Journal of Computerized Adaptive Testing

Volume 2 Number 4

December 2014

Cognitive Diagnostic Models and Computerized Adaptive Testing: Two New Item-Selection Methods That Incorporate Response Times

Matthew D. Finkelman, Wonsuk Kim, Alexander Weissman, and Robert J. Cook

DOI 10.7333/1412-0204059

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing

www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing
Volume 2, Number 4, December 2014
DOI 10.7333/1412-0204059
ISSN 2165-6592

Cognitive Diagnostic Models and Computerized Adaptive Testing: Two New Item-Selection Methods That Incorporate Response Times

Matthew D. Finkelman
Tufts Clinical and Translational Science Institute

Wonsuk Kim
Measured Progress

Alexander Weissman
Law School Admission Council

Robert J. Cook
American Board of Internal Medicine

A recent paper proposed an item-selection approach for computerized adaptive testing (CAT) in which the psychometric information per time unit is maximized. The current research extended this methodology to adaptive tests combined with use of a cognitive diagnostic model (CDM). Two new item-selection methods are introduced for the combination of CDMs and CAT: posterior-weighted Kullback-Leibler information per-time-unit, and mutual information per-time-unit. These methods were compared with standard procedures in which the amount of time required to complete an item is not considered. Simulation conditions with and without attribute-balancing constraints indicated that, on average, the new methods required more items but took less time than the standard procedures, while achieving comparable classification accuracy.

Keywords: *computerized adaptive testing, cognitive diagnostic models, Kullback-Leibler information, mutual information, diagnostic testing.*

In the past several decades, much research has sought to enhance the diagnostic power of assessments through the simultaneous measurement of multiple examinee attributes (de la Torre, Song, & Hong, 2011; Haberman & Sinharay, 2010; Hamilton, Nussbaum, Kupermintz, Kerkhoven, & Snow, 1995; Yao & Boughton, 2007). A popular psychometric framework to fa-

cilitate this goal is provided by cognitive diagnostic models (CDMs; DiBello, Roussos, & Stout, 2007; Leighton & Gierl, 2007; Rupp, Templin, & Henson, 2010). In CDMs, each examinee's latent state is described by a *K*-vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$, where the individual elements of the vector refer to the different attributes being studied. In most formulations of CDMs, including that of the current research, each element of the vector takes on a value of 0 or 1, with 0 indicating "non-mastery" and 1 indicating "mastery" in the context of an educational assessment. Examinees who complete the test typically receive feedback on each attribute, rather than a single summative score (Cheng, 2009).

A number of recent papers have studied the combination of CDMs and computerized adaptive testing (CAT; Cheng, 2009; Hsu, Wang, & Chen, 2013; Huebner, 2010; Liu, Ying, & Zhang, 2013; Mao & Xin, 2013; McGlohen & Chang, 2008; Tatsuoka & Ferguson, 2003; Wang, 2013; Wang, Chang, & Huebner, 2011; Xu, Chang, & Douglas, 2003). In CAT, the items presented to a given examinee are dependent on his/her responses to previous items. CAT thus allows for an assessment that is tailored to the examinee taking it, with items targeted to provide information about the given examinee's unknown latent trait(s). Numerous studies have found that CAT achieves substantially greater measurement efficiency than traditional paper-and-pencil tests (Gibbons et al., 2008; Moreno, Wetzel, McBride, & Weiss, 1984; Smits, Zitman, Cuijpers, den Hollander-Gijsman, & Carlier, 2012; Weiss, 1982).

A fundamental element of a CAT is the procedure by which it chooses the next item for an examinee (i.e., the item-selection method). For applications of CAT using unidimensional item response theory (IRT), the classic item-selection approaches maximize Fisher information (Lord, 1980; van der Linden & Pashley, 2000) or Kullback-Leibler information (Chang & Ying, 1996). The latter approach is more useful for CDMs because unlike Fisher information, Kullback-Leibler information is defined when the latent variable to be measured is in discrete classes (Cheng, 2009). A number of item-selection methods, based on Kullback-Leibler information or related indexes, have been proposed in the context of CDMs. These methods include maximum Kullback-Leibler information (Xu et al., 2003), maximum hybrid Kullback-Leibler information (Cheng, 2009), maximum posterior-weighted Kullback-Leibler information (PWKL; Cheng, 2009), minimum expected Shannon entropy (Tatsuoka, 2002; Xu et al., 2003), and maximum mutual information (MI; Wang, 2013).

One important property shared by all of the preceding methods is that they adhere to the traditional notion of efficiency in CAT: they seek to select items that provide maximum information. As described in Fan, Wang, Chang, and Douglas (2012), however, efficiency can also be described in terms of information per time unit. That is, in the formulation of Fan et al., the examinee's expected response time on an item is considered, in addition to the information that the item provides in estimating the examinee's trait level. Items with less than maximal information might be presented if they are expected to be answered quickly. Fan et al. showed that by incorporating anticipated response times into the item-selection process, information can accrue more rapidly in terms of the amount of time spent on the test. Thus, the efficiency of measurement can be improved when the time taken to complete the assessment is considered to be an important outcome.

Although the simulation results of Fan et al. (2012) demonstrated the utility of their approach, the study focused on applications involving unidimensional IRT models, so the extension of their methodology to CDMs is an open problem. The objective of the current study was to fill this gap by developing item-selection methods that incorporate response times for tests that use CDMs. In particular, two new item-selection methods are introduced and compared to existing

procedures in simulation. One of the new methods is a variant on PWKL item selection; the other method is a variant on MI item selection. Both variants account for a given examinee's expected response time to an item, in addition to the item's discriminatory power about the examinee's latent state $\boldsymbol{\alpha}$. PWKL was chosen for the study because this method performed well in simulations by Cheng (2009) and was also used successfully by Hsu et al. (2013). MI was chosen both on theoretical and practical grounds: it is a fundamental quantity in information theory (Cover & Thomas, 1991) and it performed well in simulations by Wang (2013). The ideas of the current paper could also be applied to Kullback-Leibler information (Xu, Chang, & Douglas, 2003) and hybrid Kullback-Leibler information (Cheng, 2009).

Cognitive Diagnostic Models

As mentioned above, CDMs define an examinee's latent state via a vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_K)$ identifying the attributes mastered by the examinee $(\{k: \alpha_k = 1\})$ and the attributes not mastered by the examinee $(\{k: \alpha_k = 0\})$. Another important quantity in CDMs is the Q-matrix (Tatsuoka, 1985), which shows the items that measure each attribute. Letting I denote the number of items under study, the Q-matrix has I rows and K columns. The $[i, k]$ entry of the Q-matrix, q_{ik}, is then defined as 1 if item i measures attribute k, and 0 if it does not. For instance, if the first four rows of a Q-matrix are

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix}, \tag{1}$$

then item 1 measures the first and third attributes, item 2 measures only the fourth attribute, item 3 measures the first, second, and fifth attributes, and item 4 measures the second and third attributes. The Q-matrix can be determined *a priori* by content experts (Cheng, 2009; Hsu et al., 2013), although it can also be identified, validated, or refined via empirical evidence (Chiu, 2013; de la Torre, 2008; Liu, Xu, & Ying, 2012).

A variety of different CDMs have been developed and studied (DiBello et al., 2007). Some examples are the deterministic input, noisy "and" gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001), the deterministic input, noisy "or" gate (DINO) model (Templin & Henson, 2006), the noisy input, deterministic "and" gate (NIDA) model (Junker & Sijtsma, 2001; Maris, 1999), and the reparameterized unified model (RUM; Roussos et al., 2007). The theoretical underpinnings of a model from a cognitive psychology perspective (Rupp & Templin, 2008) should be considered in any application, as well as the fit of the model to the data (de la Torre & Douglas, 2004).

Although each CAT item-selection method examined here is applicable to all of the CDMs mentioned above, the simulation study presented below used the RUM. A fundamental aspect of the RUM is that it considers the probability of an individual attribute being applied correctly to an item, given the examinee's mastery level of the attribute. In particular, let Y_{ik} be an indicator variable representing whether a given examinee successfully applies attribute k to item i; to simplify, notation for the indexing of examinees is suppressed. The probability that attribute k is successfully applied to item i, given that attribute k has been mastered by the examinee, is then

defined as $\pi_{ik} = P(Y_{ik} = 1 | \alpha_k = 1)$. Next, denote by π_i^* the probability that all attributes measured by item *i* are successfully applied to the item—hence, the item is answered correctly—given that all such attributes have been mastered. Assuming local independence between the attributes, and recalling that attribute *k* is measured by item *i* if and only if $q_{ik} = 1$,

$$\pi_i^* = \prod_{k=1}^{K} \pi_{ik}^{q_{ik}} \tag{2}$$

(Roussos et al., 2007).

Because the value in Equation 2 corresponds to an examinee who has mastered all attributes relevant to item *i*, the chance of a correct response to this item can be no greater than π_i^*. In other words, π_i^* represents the upper bound for the chance of a correct response to item *i*. This upper bound is obtained for examinees who have mastered all attributes that are measured by the item. The chance of a correct response can be lower than π_i^* if one of the relevant attributes has not been mastered. Let $r_{ik} = P(Y_{ik} = 1 | \alpha_k = 0)$ represent the probability that attribute *k* is successfully applied to item *i*, given that attribute *k* has not been mastered by the examinee. It is assumed that $r_{ik} \leq \pi_{ik}$, so that the probability of successful application is not higher by a non-master of the attribute than by a master. Let $r_{ik}^* = r_{ik} / \pi_{ik}$; note that $r_{ik}^* \leq 1$ and a value closer to 0 indicates greater discrimination between masters and non-masters of the attribute. Under the "reduced" form of the RUM, the probability of a correct response to item *i*, given an underlying examinee attribute pattern of $\boldsymbol{\alpha}$, is

$$P_i(\boldsymbol{\alpha}) = \pi_i^* \prod_{k=1}^{K} r_{ik}^{*\,(1-\alpha_k)q_{ik}} \tag{3}$$

(Roussos et al., 2007). Here the product term serves to lower the probability of a correct response for examinees who have not mastered some relevant attributes. The reduced RUM is a special case that simplifies the RUM model by excluding a term related to "supplemental" abilities that are outside of the Q-matrix. See Roussos et al. for technical details; see Kim (2011) for a recent application of the reduced RUM.

Methods for CAT Item Selection with CDMs

It is assumed that item parameters have been obtained for a specific model (i.e., a CDM) and are treated as known. It is further assumed that a prior distribution $p_0(\boldsymbol{\alpha})$ has been placed on the 2^K possible attribute patterns, such that $\sum_{\boldsymbol{\alpha}} p_0(\boldsymbol{\alpha}) = 1$ and all $p_0(\boldsymbol{\alpha}) \geq 0$. This prior distribution is updated after each item, so that after *m* items have been administered, the posterior probability of a given attribute pattern $\boldsymbol{\alpha}'$ is equal to

$$p_m(\boldsymbol{\alpha}') = \frac{p_0(\boldsymbol{\alpha}') \prod_{i=1}^{m} P_i(\boldsymbol{\alpha}')^{u_i} \left[1 - P_i(\boldsymbol{\alpha}')\right]^{1-u_i}}{\sum_{\boldsymbol{\alpha}} p_0(\boldsymbol{\alpha}) \prod_{i=1}^{m} P_i(\boldsymbol{\alpha})^{u_i} \left[1 - P_i(\boldsymbol{\alpha})\right]^{1-u_i}}. \tag{4}$$

Here, u_i is the examinee's response to item i (0 = incorrect, 1 = correct), and $P_i(\boldsymbol{\alpha}')^{u_i}\left[1-P_i(\boldsymbol{\alpha}')\right]^{1-u_i}$ is the likelihood of $\boldsymbol{\alpha}'$ with respect to the response u_i (i.e., the modeled probability that an examinee with attribute pattern $\boldsymbol{\alpha}'$ would produce the response u_i to item *i*). The product assumes local independence, whereas the denominator ensures that the set of posterior probabilities sums to 1.

After all posterior probabilities have been computed, the maximum *a posteriori* (MAP) estimate of $\boldsymbol{\alpha}$ can be obtained. The MAP estimate after m items is defined as $\hat{\boldsymbol{\alpha}}_m = \arg\max_a \{p_m(\boldsymbol{\alpha})\}$, which makes it the attribute pattern with the largest posterior probability of being the examinee's true state.

The posterior-weighted Kullback-Leibler (PWKL) criterion. The PWKL criterion is based on the notion of Kullback-Leibler information, which has been described elsewhere (Chang & Ying, 1996; Cover & Thomas, 1991), including in the context of CDMs (Henson & Douglas, 2005). A brief summary is given here and formal details are provided in the Appendix. Briefly, let $K_i(\boldsymbol{\alpha}',\boldsymbol{\alpha}'')$ denote the Kullback-Leibler information of item *i* for attribute patterns $\boldsymbol{\alpha}'$ and $\boldsymbol{\alpha}''$. A larger value of $K_i(\boldsymbol{\alpha}',\boldsymbol{\alpha}'')$ indicates that under $\boldsymbol{\alpha}'$, item *i* provides more information to discern between $\boldsymbol{\alpha}'$ and $\boldsymbol{\alpha}''$.

PWKL sums multiple $K_i(\boldsymbol{\alpha}',\boldsymbol{\alpha}'')$ values per item and then compares the results across items, thereby creating a criterion for CAT item selection. Because $\hat{\boldsymbol{\alpha}}_m$ is the most likely attribute pattern for the given examinee (according to the posterior distribution after *m* items), the focus is on the Kullback-Leibler values that involve this pattern. In particular, the Kullback-Leibler information between $\hat{\boldsymbol{\alpha}}_m$ and every other attribute pattern $\boldsymbol{\alpha}''$ is computed (setting $\boldsymbol{\alpha}' = \hat{\boldsymbol{\alpha}}_m$). The results are then summed across $\boldsymbol{\alpha}''$, weighting the Kullback-Leibler values by the posterior probability of $\boldsymbol{\alpha}''$ in order to give greater influence to attribute patterns that are more likely to be the true state. Mathematically, the PWKL value for item *i* after the administration of *m* items is equal to

$$PWKL_i^m = \sum_{\boldsymbol{\alpha}''} K_i(\hat{\boldsymbol{\alpha}}_m,\boldsymbol{\alpha}'')\, p_m(\boldsymbol{\alpha}'') \tag{5}$$

(Cheng, 2009). The item with the largest $PWKL_i^m$ value, among the set of eligible items in the bank, is then selected to be item $m+1$ of the test (Cheng, 2009).

The mutual information (MI) criterion. Because MI has been thoroughly treated elsewhere (Cover & Thomas, 1991; Wang & Chang, 2011; Weissman, 2007), including in the context of CDMs (Wang, 2013), this section provides only a brief account of MI, with formal details in the Appendix. MI has been described as measuring "the reduction in the uncertainty of one random variable due to the knowledge of the other" (Cover & Thomas, 1991, p. 18). In the context of CDMs, interest lies in reducing the uncertainty of $\boldsymbol{\alpha}$ because of knowledge of an examinee's response to an item.

To use MI as an item-selection criterion with CDMs, this information index must be computed for every candidate item after each stage of the test. Denote U_i as a random variable indicating the examinee's response to item *i* (for dichotomous items, $U_i = 0$ or $U_i = 1$). As previously, let u_i represent the examinee's actual response to item *i*, that is, the realized value of U_i. Let

$P_m(U_i=1)$ and $P_m(U_i=0)$ denote the marginal probabilities of a correct answer and incorrect answer, respectively, to item *i* after *m* items have been administered. For a given attribute pattern $\boldsymbol{\alpha}'''$, let $P_m(\boldsymbol{\alpha}''', U_i=1)$ denote the joint probability of $\boldsymbol{\alpha}'''$ and a correct response to item *i* after *m* items, and $P_m(\boldsymbol{\alpha}''', U_i=0)$ denote the joint probability of $\boldsymbol{\alpha}'''$ and an incorrect response to item *i* after *m* items. The MI between $\boldsymbol{\alpha}$ and the response to item *i*, following the administration of *m* items, is then

$$I_m(\boldsymbol{\alpha};U_i)=\sum_{\boldsymbol{\alpha}'''}\sum_{u_i=0}^{1}P_m(\boldsymbol{\alpha}''',U_i=u_i)\log\frac{P_m(\boldsymbol{\alpha}''',U_i=u_i)}{p_m(\boldsymbol{\alpha}''')P_m(U_i=u_i)}, \tag{6}$$

where $p_m(\boldsymbol{\alpha}''')$ is again the posterior probability of $\boldsymbol{\alpha}'''$ after *m* items. Larger values of $I_m(\boldsymbol{\alpha};U_i)$ indicate greater information; hence, the item with the largest $I_m(\boldsymbol{\alpha};U_i)$ value, among the set of eligible items in the bank, is selected to be item $m+1$ of the test.

Methods incorporating response times. The majority of response-time models, including the one used in this research, relate an examinee's observable response times to a latent variable representing the examinee's speed. Let T_i denote a random variable representing the amount of time spent by an examinee on item *i* (for simplicity, the notation for examinee indexing is suppressed). Let t_i denote the realized value of T_i. Finally, let τ denote the examinee's latent speed (examinees with higher values of τ tend to have faster responses). Under the model of van der Linden (2006), which was used in the current study, the density of T_i for an examinee of speed τ is given by

$$f(t_i;\tau,A_i,\beta_i)=\frac{A_i}{t_i\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[A_i(\ln t_i-(\beta_i-\tau))\right]^2\right\}. \tag{7}$$

Here β_i is the "time intensity" parameter of item *i* (a larger value of β_i indicates that the item tends to require more time). A_i is a parameter that gives the item's discriminatory power for distinguishing different values of τ (greater discriminatory power is obtained with a higher value of A_i). Note that the latter parameter is usually represented by α_i; A_i is used here to avoid confusion with the examinee's attribute pattern, which is denoted by $\boldsymbol{\alpha}$. Equation 7 assumes that T_i is lognormal (i.e., the logarithm of T_i follows the normal distribution). The mean and standard deviation of this normal distribution are $\beta_i-\tau$ and $1/A_i$, respectively.

During the test, an examinee's latent speed can be estimated, based on responses to previous items. In particular, after *m* items have been administered, the maximum likelihood estimate (MLE) of τ is equal to

$$\hat{\tau}_m=\frac{\sum_{i=1}^{m}A_i^2(\beta_i-\ln t_i)}{\sum_{i=1}^{m}A_i^2} \tag{8}$$

(van der Linden, 2011a). $\hat{\tau}_m$ can then be used to estimate the amount of time that an examinee will spend on any item *i*. Specifically, the expected amount of time spent on item *i*, assuming a

latent speed of $\hat{\tau}_m$, is given by

$$\mathrm{E}(T_i \mid \hat{\tau}_m) = \exp\left(\beta_i - \hat{\tau}_m + \frac{1}{2A_i^2} \right) \tag{9}$$

(Fan et al., 2012).

Response-time models have been used for a number of purposes, including detecting collusion between examinees (van der Linden, 2009a), setting time limits on tests (van der Linden, 2011b), and assembling tests with constraints on speededness (van der Linden, 2011a). Fan et al. (2012) used response-time models to create a CAT item-selection method that enhances time efficiency. For a unidimensional IRT model, let θ represent an examinee's latent level of ability, $\hat{\theta}_m$ its MLE after *m* items, and $I_i(\hat{\theta}_m)$ the Fisher information of item *i* at $\hat{\theta}_m$. Instead of basing item selection solely on the latter quantity, which is a traditional approach for unidimensional CAT (Lord, 1980; van der Linden & Pashley, 2000), Fan et al. maximized information per time unit, i.e., they proposed selecting the item that maximizes

$$\frac{I_i(\hat{\theta}_m)}{\mathrm{E}(T_i \mid \hat{\tau}_m)}, \tag{10}$$

among the set of eligible items in the bank. This criterion takes into account not only the amount of information provided by a given item, but also its expected time. Hence, if two items ("Item A" and "Item B") have equal Fisher information, but Item A's expected time is less than Item B's, the former is preferred to the latter. For the lognormal model of van der Linden (2006), the denominator of Equation 10 is given by Equation 9; it is assumed that all response-time parameters have been estimated and are treated as known. Such estimation can be accomplished via computer-based pretesting, for which the response time is logged for each examinee and item (van der Linden, 2011b). The unit of time used in the response-time model (e.g., seconds, minutes) does not affect the item chosen.

To extend this approach to CDMs, $PWKL_i^m$ or $I_m(\boldsymbol{\alpha};U_i)$ is substituted into Equation 10 in place of $I_i(\hat{\theta}_m)$. The PWKL per-time-unit criterion is thus

$$\frac{PWKL_i^m}{\mathrm{E}(T_i \mid \hat{\tau}_m)} = \frac{\sum_{\boldsymbol{\alpha}''} K_i(\hat{\boldsymbol{\alpha}}_m, \boldsymbol{\alpha}'')\, p_m(\boldsymbol{\alpha}'')}{\mathrm{E}(T_i \mid \hat{\tau}_m)}; \tag{11}$$

and the MI per-time-unit criterion is

$$\frac{I_m(\boldsymbol{\alpha};U_i)}{\mathrm{E}(T_i \mid \hat{\tau}_m)} = \frac{\sum_{\boldsymbol{\alpha}'''} \sum_{u_i=0}^{1} P_m(\boldsymbol{\alpha}''', U_i = u_i) \log \dfrac{P_m(\boldsymbol{\alpha}''', U_i = u_i)}{p_m(\boldsymbol{\alpha}''') P_m(U_i = u_i)}}{\mathrm{E}(T_i \mid \hat{\tau}_m)}. \tag{12}$$

Again, when using the lognormal model of van der Linden (2006), the denominator of Equations 11 and 12 are given by Equation 9. The two new item-selection methods introduced in this paper are designed to maximize Equation 11 or 12 among the set of eligible items in the bank.

Interestingly, when using Equation 11 or 12 as an item-selection criterion, precise estimation of $\hat{\tau}_m$ becomes unnecessary, because of the separability of the person and item parameters in the

conditional expectation in Equation 9 that appears in the denominator of both Equations 11 and 12. That is, Equation 9 can be factored as

$$E(T_i \mid \hat{\tau}_m) = \exp(-\hat{\tau}_m)\left[\exp\left(\beta_i + \frac{1}{2A_i^2}\right)\right] \tag{13}$$

(van der Linden, 2011a); hence for any item *i*, the estimated person parameter $(\hat{\tau}_m)$ is separable from the item parameters $(A_i$ and $\beta_i)$. For a given examinee, $\exp(-\hat{\tau}_m)$ remains constant when evaluating the PWKL or MI per-time-unit criterion for all eligible items, so the item selected by maximizing either criterion does not depend on the specific value of $\hat{\tau}_m$. In fact, any value of τ would result in the selection of the same item as that chosen when $\hat{\tau}_m$ is used. In this paper, $\hat{\tau}_m$ was used in Equations 11 and 12 to promote consistency with the formulation of Fan et al. (2012); in practice, an arbitrary value of τ (such as $\tau = 0)$ can be used equivalently.

Simulation Design

A simulation study was conducted to compare the four CAT item-selection methods outlined above: PWKL, MI, PWKL per-time-unit, and MI per-time-unit. The investigation used a synthetic item bank that had previously been used for simulation by Finkelman, Kim, and Roussos (2009) and Finkelman, Kim, Roussos, and Verschoor (2010). This item bank contained 300 items following the reduced RUM. A total of five attributes were under study; the Q-matrix was composed so that the mean number of attributes measured per item was two (specifically, 80 items measured one attribute, 140 items measured two attributes, and 80 items measured three attributes). The particular attribute(s) measured by a given item were chosen at random, with each attribute having an equal chance of representation. For the item parameters of the reduced RUM, π_i^* values were randomly sampled from the uniform distribution with a lower bound of 0.75 and an upper bound of 0.95, and r_{ik}^* values were sampled from the uniform distribution with a lower bound of 0.40 and an upper bound of 0.85. These bounds were selected so that results would be comparable to the results found empirically by Jang (2005, 2006) and Roussos, Hartz, and Stout (2003).

Because there were five attributes under study, there were $2^5 = 32$ possible attribute patterns (i.e., 32 possible manifestations of the vector $\boldsymbol{\alpha}$). Each attribute pattern was assumed to have a prior probability of $p_0(\boldsymbol{\alpha}) = 1/32$. One hundred simulated examinees ("simulees") with each attribute pattern were generated, for a total of 3,200 simulees. A discrete uniform prior had also been employed by Cheng (2009) to represent the case in which the prior distribution is uninformative.

For parameters related to the response-time model, A_i was sampled from the uniform distribution with a lower bound of 2 and an upper bound of 4, and β_i was sampled from the normal distribution with a mean of 0 and a variance of 0.25. The speed parameter τ was sampled from the standard normal distribution. All of these parameters were chosen to match a simulation condition of Fan et al. (2012).

All item responses and response times were generated at the start of the study; it was then determined post-hoc which items would be selected by each item-selection method if administration were conducted adaptively. For the PWKL and PWKL per-time-unit methods, the MAP es-

timate $\hat{\boldsymbol{\alpha}}_m$ might not be unique (i.e., multiple $\boldsymbol{\alpha}$ vectors might be "tied" for being the MAP). The chance that the MAP lacks uniqueness generally depends on both the stage of the test and the prior distribution for $\boldsymbol{\alpha}$. If there is a discrete uniform prior distribution on $\boldsymbol{\alpha}$, then all attribute patterns are necessarily "tied" for being the MAP before the first item is administered. Under this prior distribution, it is also common for ties to occur in the initial stages of testing. Later in the test, the posterior probabilities of the different attribute patterns typically become sufficiently diffuse that ties are unlikely. In this study, when ties occurred, the calculations of Equations 5 and 11 were done for each $\hat{\boldsymbol{\alpha}}_m$ value (i.e., each value tied for being the MAP), and the results were averaged. The MI and MI per-time-unit criteria do not require a definition of $\hat{\boldsymbol{\alpha}}_m$ in the item-selection process and therefore did not need this method of averaging.

Two levels of attribute balancing constraints on item selection were studied. In the first level, item selection was unconstrained: the PWKL, MI, PWKL per-time-unit, and MI per-time-unit criteria were administered in their forms as mentioned above. In the second level, item selection was constrained by an attribute balancing rule after the administration of the first item, which itself was unconstrained. Specifically, the attribute balancing rule required that the selected item measure the attribute that had been measured the fewest times previously for the simulee being examined. If there was a tie for least representation among two or more attributes, the selected item was constrained to measure at least one of the attributes that was part of the tie. The item with the largest value of a given method (PWKL, MI, PWKL per-time-unit, or MI per-time-unit), among the set of eligible items satisfying the constraint, was then defined as the item selected by that method.

To determine when to halt test administration and make a final estimate of $\boldsymbol{\alpha}$, a variable-length procedure proposed by Hsu et al. (2013) was adopted. Specifically, the test was ended as soon as the following two conditions were both satisfied: (1) one of the candidate attribute patterns (i.e., the MAP estimate $\hat{\boldsymbol{\alpha}}_m$) had a posterior probability of 0.80 or greater, and (2) none of the other attribute patterns had a posterior probability greater than 0.10. The values 0.80 and 0.10 were mentioned by Hsu et al. as appropriate for low-stakes (diagnostic) tests. After testing was stopped, the MAP estimate was used as the final estimate of $\boldsymbol{\alpha}$.

The following outcomes were recorded for each item-selection method under study:

1. The proportion of simulees for whom attribute *k* was classified correctly, $k = 1,...,5$.
2. The proportion of simulees for whom the entire estimated attribute pattern was correct (i.e., for whom all five attributes were classified correctly).
3. The mean number of items administered.
4. The mean amount of time required to complete the test.

Results

Table 1 displays the results for the unconstrained condition, in which attribute balancing was not enforced. The table shows that the classification accuracy rates were high: the proportion of correct classifications was at least 0.94 for every combination of attribute and item-selection method. The accuracy rates were also similar across the item-selection methods: for each attribute, the proportion of correct classifications never differed by more than 0.01 across the different methods. Furthermore, the proportion of simulees for whom the entire attribute pattern was classified correctly was either 0.82 or 0.83 for every item-selection method. The latter result is consistent with the requirement that the posterior probability of an attribute pattern be at least 0.80

before termination can occur.

The last two columns of Table 1 indicate that for the unconstrained condition, the choice between the standard methods (PWKL and MI) and their respective per-time-unit versions was consequential in terms of respondent burden. Use of PWKL per-time-unit, rather than PWKL, resulted in a 19% increase in the mean number of items administered (from 24.3 to 28.8 items), but a 34% decrease in the mean amount of time required to complete the test (from 47.6 to 31.2 time units). The use of MI per-time-unit, rather than MI, resulted in a 21% increase in the mean number of items administered (from 24.1 to 29.1 items), but a 30% decrease in the mean amount of time required to complete the test (from 44.9 to 31.4 time units). These results were consistent with the respective goals of the different methods: PWKL and MI seek to select the most informative items, thus achieving a low average number of items administered, whereas their per-time-unit versions seek to administer the most time-efficient items, reducing the mean amount of time spent on the test. Note that the choice of units used to measure time (e.g., half-minutes, minutes) does not affect the percentage decrease in mean testing time achieved by the per-time-unit methods.

Table 1. Results for Unconstrained Item Selection (*N* = 3,200 Simulees)

	Proportion of Correct Classifications by Attribute							
Method	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Entire Attribute Pattern	Mean No. of Items	Mean Time Required
PWKL	0.96	0.94	0.96	0.96	0.96	0.83	24.3	47.6
PWKL Per-Time-Unit	0.96	0.94	0.96	0.96	0.97	0.83	28.8	31.2
MI	0.96	0.95	0.95	0.96	0.96	0.82	24.1	44.9
MI Per-Time-Unit	0.96	0.95	0.95	0.96	0.96	0.82	29.1	31.4

One final comparison to be made from Table 1 is the comparison between the PWKL-based and MI-based procedures (i.e., PWKL versus MI, and PWKL per-time-unit versus MI per-time-unit). The table shows that for the unconstrained condition, the difference between PWKL and MI was relatively small in terms of respondent burden. Use of MI in place of PWKL resulted in a 1% decrease in the mean number of items administered (from 24.3 to 24.1 items) and a 6% decrease in the mean time required (from 47.6 to 44.9 time units). Use of MI per-time-unit resulted in a 1% increase in the mean number of items administered (from 28.8 to 29.1 items) and a 1% increase in the mean time required (from 31.2 to 31.4 time units).

Table 2 displays the results for the constrained condition, in which attribute balancing was required. As in the previous condition, classification accuracy rates were high (at least 0.94 for every combination of attribute and item-selection method) and similar across item-selection methods (the difference between methods was no more than 0.02 for any attribute). The proportion of simulees for whom the entire attribute pattern was classified correctly was again either 0.82 or 0.83 for all methods.

A comparison of the constrained condition to the unconstrained condition shows that the former resulted in greater respondent burden than the latter. In particular, both the mean number of items and the mean time required were higher in the constrained condition. These results were anticipated; because the constrained condition did not allow the methods to select items based purely on their discriminatory power, it was expected that more items and time would be needed

to reach the same level of classification accuracy. Another way to understand the comparison between the constrained and unconstrained conditions is to note that with a relatively uniform representation of attributes in the item bank, the unconstrained item-selection methods implicitly tended to measure each attribute approximately (but not exactly) the same number of times (results not shown). Enforcing explicit constraints on attribute balance influenced item selection and resulted in reduced efficiency.

Table 2. Results for Constrained Item Selection (*N* = 3,200 Simulees)

	Proportion of Correct Classifications by Attribute							
Method	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Entire Attribute Pattern	Mean No. of Items	Mean Time Required
PWKL	0.96	0.95	0.96	0.97	0.96	0.83	32.1	62.0
PWKL Per-Time-Unit	0.96	0.95	0.96	0.97	0.96	0.83	36.5	42.6
MI	0.96	0.94	0.96	0.96	0.97	0.82	30.8	59.5
MI Per-Time-Unit	0.96	0.96	0.96	0.96	0.96	0.83	36.8	43.2

Turning to the comparison between the standard item-selection methods and their per-time-unit versions, results of the constrained condition were similar to those of the unconstrained condition. Using PWKL per-time-unit, rather than PWKL, resulted in a 14% increase in the mean number of items administered (from 32.1 to 36.5 items), but a 31% decrease in the mean amount of time required to complete the test (from 62.0 to 42.6 time units). Using MI per-time-unit, instead of only MI, resulted in a 19% increase in the mean number of items administered (from 30.8 to 36.8 items), but a 27% decrease in the mean amount of time required to complete the test (from 59.5 to 43.2 time units).

Finally, a comparison of the PWKL-based and MI-based item-selection methods to each other again showed slight differences. Use of MI in place of PWKL produced 4% decreases in both the mean number of items administered (from 32.1 to 30.8 items) and the mean time required (from 62.0 to 59.5 time units). Use of MI per-time-unit produced 1% increases in both the mean number of items administered (from 36.5 to 36.8 items) and the mean time required (from 42.6 to 43.2 time units), compared to PWKL per-time-unit.

Discussion

The objective of this research was to extend the information-per-time-unit item-selection approach of Fan et al. (2012) to the CDM setting. Results suggested that this approach has potential to substantially improve the time-efficiency of a diagnostic assessment conducted via CAT. In particular, the per-time-unit versions of PWKL and MI lessened the average testing time by at least 30% in the unconstrained condition and at least 27% in the constrained condition. These reductions are similar to those found in Table 1 of Fan et al. for the unidimensional IRT case. Ultimately, the decision of whether to use PWKL per-time-unit or MI per-time-unit, rather than PWKL or MI, depends in large part on the goal of the practitioner. If the goal is to minimize the average number of items administered, then the standard methods (PWKL and MI) might be preferred. However, if the goal is to minimize the average testing time, the per-time-unit versions might be favored. The latter methods are expected to be useful in certain testing applications, and

therefore constitute an important addition to the existing set of CAT item-selection procedures. Moreover, although the per-time-unit methods require response-time information about each item in the bank, such information is becoming more available as computerized testing becomes more common (van der Linden, 2006).

The simulation design presented here also allowed for a comparison of PWKL-based and MI-based item-selection procedures. PWKL and MI tended to produce similar results to each other in both their traditional forms and their per-time-unit versions. PWKL resulted in slightly greater respondent burden than MI, and PWKL per-time-unit resulted in slightly lower respondent burden than MI per-time-unit. The question of whether these findings represent a systematic trend or an artifact of the data should be investigated in future research.

One important limitation of the study is that it relied exclusively on simulated data. Designing realistic simulation studies that combine CAT, CDMs, and response-time data is not without its challenges. This study employed the reduced RUM as the model for CDMs and van der Linden's (2006) model for response times. These two models were applied independently in the creation of simulated data. In practice, an association could exist between an examinee's answer (correct or incorrect) and his/her response time. Because the results observed in this study were intuitive (PWKL per-time-unit and MI-per-time unit tended to present more items, but take less time, than their traditional counterparts), it might be expected that these results could also be found in different settings. Nevertheless, additional research should be conducted to better understand the relation between examinees' answers and response times when a CDM is used, and how this relation affects the relative efficiencies of the different CAT item-selection methods. Simulations in which the item parameters of the CDM are correlated with the item parameters of the response-time model should be conducted as well.

Further studies could also be performed in which exposure control methodology is used to limit the percentage of times that an item is administered; for example, the procedure of Sympson and Hetter (1985) is readily applicable to the methods described here. As discussed by Cheng (2009), however, exposure control is generally less critical for CDMs than some other psychometric models, because CDMs are often applied in low-stakes settings. Nevertheless, the combination of exposure control and CDMs has been considered (e.g., Hsu et al., 2013), and the study of exposure control alongside the CDM-based per-time-unit item-selection methods is an important topic. An additional area of future research is to examine the types of items that are selected by the per-time-unit methods in empirical applications. It is possible that these methods will tend to choose items that are qualitatively different from CAT methods that do not take time-efficiency into account. For example, the PWKL per-time-unit and MI per-time-unit approaches might result in the selection of items that require fewer steps to solve, because such items would also be expected to require less time. If studies reveal that the per-time-unit methods tend to avoid certain desired item types, constraints could be added to ensure that enough items of those types are included in each administration. Such constraints, as well as modifications of the item-selection procedures to incorporate exposure control, would be expected to reduce the efficiency of the CAT. Also, before the new CAT item-selection methods are employed operationally, the response-time model being used should be validated in the context of cognitive assessment.

The Fan et al. (2012) study is not the only previous attempt to incorporate response times in CAT item selection. Van der Linden (2009b), van der Linden, Scrams, and Schnipke (1999), and van der Linden and Xiong (2013) also combined response-time information with CAT. However, the work of van der Linden and colleagues focused on control of differential speededness, rather

than the selection of items to maximize information per time unit. Although the goal of the current paper has been to extend the research of Fan et al. on the latter topic to CDMs, controlling differential speededness alongside CDMs is also possible and is a potential topic for future work.

As mentioned above, the items chosen by the per-time-unit methods do not depend on the value of τ used in the computation of expected time, so calculating $\hat{\tau}_m$ is unnecessary for the item-selection process. This observation raises the question of why response-time modeling is needed for per-time-unit item selection—specifically, whether the simple mean of an item's elicited response times during pretesting could be used in lieu of $\mathrm{E}(T_i \mid \hat{\tau}_m)$ in the denominator of Equations 11 and 12. In fact, Lau and Wang (2000) discussed the idea of simply dividing item information by average response time within the context of computerized classification testing. Although the response-time model used in the current research requires an assumption that the speed parameter is a constant (including among items measuring different attributes, in the CDM context), the response-time model can be used even when examinees receive different items from one another during pretesting. This benefit can be weighed against its relative complexity, compared with the approach of simply dividing information by the average response time observed during pretesting. A formal comparison of the two approaches could be illuminating.

It is notable that the estimation of $\boldsymbol{\alpha}$ in the simulation study was based solely on the given examinee's response pattern. As mentioned by van der Linden (2006) and Fan et al. (2012), estimation of the trait(s) of interest (here, $\boldsymbol{\alpha}$) could potentially be enhanced by incorporating information about the examinee's speed. As in the current study, Fan et al. did not pursue joint estimation of τ and the ability parameter, which for Fan et al. was the unidimensional IRT parameter θ. The authors stated that accounting for speed in the estimation of θ could be effective, but should be done cautiously; this statement also holds for the estimation of $\boldsymbol{\alpha}$.

Although the research presented here represents a step toward maximizing information per time unit in CDMs, further studies are needed. In addition to addressing the topics mentioned above, future work should compare the standard item-selection methods with their per-time-unit counterparts under different conditions. Simulations should be conducted using a CDM other than the reduced RUM, and the item bank, stopping rule, constraints, and number of attributes measured should be varied. The response-time parameters—and potentially the response-time model itself—should be varied, as well. Finally, all CAT item-selection methods should be analyzed under the assumption of an incorrectly specified Bayesian prior distribution in order to assess their robustness. The investigation of each of these topics would constitute an advance toward obtaining time-efficient diagnostic information via CAT.

References

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

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

Chiu, C.-Y. (2013). Statistical refinement of the Q-matrix in cognitive diagnosis. *Applied Psychological Measurement, 37,* 598-618. CrossRef

Cover, T.M., & Thomas, J.A. (1991). *Elements of information theory.* New York, NY: John Wiley. CrossRef

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. CrossRef

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

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement, 35,* 296-316. CrossRef

DiBello, L., Roussos, L.A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C.V. Rao, & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26, Psychometrics)* (pp. 979-1027). Amsterdam: Elsevier. CrossRef

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. CrossRef

Finkelman, M., Kim, W., & Roussos, L.A. (2009). Automated test assembly for cognitive diagnosis models using a genetic algorithm. *Journal of Educational Measurement, 46,* 273-292. CrossRef

Finkelman, M.D., Kim, W., Roussos, L., & Verschoor, A. (2010). A binary programming approach to automated test assembly for cognitive diagnosis models. *Applied Psychological Measurement, 34,* 310-326. CrossRef

Gibbons, R.D., Weiss, D.J., Kupfer, D.J., Frank, E., Fagiolini, A., Grochocinski, V.J., et al. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59,* 361-368. CrossRef

Haberman, S.J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75,* 209-227. CrossRef

Haertel, E.H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 301-321. CrossRef

Hamilton, L.S., Nussbaum, E.M., Kupermintz, H., Kerkhoven, J.I.M., & Snow, R.E. (1995). Enhancing the validity and usefulness of large-scale educational assessments: II. NELS:88 science achievement. *American Educational Research Journal, 32,* 555-581. CrossRef

Henson, R.A., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement, 29,* 262-277. CrossRef

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. CrossRef

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment, Research & Evaluation, 15*. Available online.

Jang, E.E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Unpublished doctoral dissertation, University of Illinois, Champaign, IL.

Jang, E.E. (2006, April). Pedagogical implications of cognitive skills diagnostic assessment for teaching and learning. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA.

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. CrossRef

Kim, Y.-H. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified Model. *Language Testing, 28,* 509-541. CrossRef

Lau, C.A., & Wang, T. (2000). A new item selection procedure for mixed item type in computerized classification testing. Paper presented at the annual meeting of the American

Educational Research Association. New Orleans, LA.
Leighton, J.P., & Gierl, M.J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge, UK: Cambridge University Press. CrossRef
Liu, J., Xu, G., & Ying, Z. (2012). Data-driven learning of Q-matrix. *Applied Psychological Measurement, 36,* 548-564. CrossRef
Liu, J., Ying, Z., & Zhang, S. (2013). A rate function approach to computerized adaptive testing for cognitive diagnosis. *Psychometrika.* Advance online publication. CrossRef
Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.
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. CrossRef
Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64,* 187-212. CrossRef
McGlohen, M., & Chang, H.-H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods, 40,* 808-821. CrossRef
Moreno, K.E., Wetzel, C.D., McBride, J.R., & Weiss, D.J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery and computerized adaptive testing subtests. *Applied Psychological Measurement, 8,* 155-163. CrossRef
Roussos, L.A., DiBello, L.V., Stout, W., Hartz, S.M., Henson, R.A., & Templin, J.L. (2007). The fusion model skills diagnosis system. In J.P. Leighton & M.J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275-318). Cambridge, UK: Cambridge University Press. CrossRef
Roussos, L.A., Hartz, S.M., & Stout, W.M. (2003, April). Real data applications of the Fusion Model skills diagnostic system. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.
Rupp, A.A., & Templin, J.L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6,* 219-262. CrossRef
Rupp, A.A., Templin, J., & Henson, R.A. (2010). *Diagnostic measurement: Theory, methods, and applications.* New York, NY: Guilford.
Smits, N., Zitman, F.G., Cuijpers, P., den Hollander-Gijsman, M.E., & Carlier, I.V.E. (2012). A proof of principle for using adaptive testing in routine Outcome Monitoring: The efficiency of the Mood and Anxiety Symptoms Questionnaire—Anhedonic Depression CAT. *BMC Medical Research Methodology, 12,* 4. CrossRef
Sympson, J.B., & Hetter, R.D. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.
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. CrossRef
Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society, Series B (Statistical Methodology), 65,* 143-157. CrossRef
Tatsuoka, K.K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10,* 55-73. CrossRef
Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11,* 287-305. CrossRef
van der Linden, W.J. (2006). A lognormal model for response times on test items. *Journal of*

Educational and Behavioral Statistics, 31, 181-204. CrossRef

van der Linden, W.J. (2009a). A bivariate lognormal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics, 34,* 378-394. CrossRef

van der Linden, W.J. (2009b). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement, 33,* 25-41. CrossRef

van der Linden, W.J. (2011a). Test design and speededness. *Journal of Educational Measurement, 48,* 44-60. CrossRef

van der Linden, W.J. (2011b). Setting time limits on tests. *Applied Psychological Measurement, 35,* 183-199. CrossRef

van der Linden, W.J., & Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1-25). Dordrecht, The Netherlands: Kluwer Academic Publishers.

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. CrossRef

van der Linden, W.J., & Xiong, X. (2013). Speededness and adaptive testing. *Journal of Educational and Behavioral Statistics, 38,* 418-438. CrossRef

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. CrossRef

Wang, C., & Chang, H.-H. (2011). Item selection in multidimensional computerized adaptive tests—gaining information from different angles. *Psychometrika, 76,* 363-384. CrossRef

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. CrossRef

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473-492. CrossRef

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement, 67,* 41-58. CrossRef

Xu, X., Chang, H.-H., & Douglas, J. (2003, April). A simulation study to compare CAT strategies for cognitive diagnosis. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.

Yao, L., & Boughton, K.A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31,* 83-105. CrossRef

Acknowledgements

The authors would like to thank two anonymous reviewers, as well as G. Gage Kingsbury, for their suggested improvements to a previous version of this paper. We are also indebted to Louis Roussos for the use of his synthetic item bank.

Author Address

Matthew D. Finkelman, 35 Kneeland St.. Boston, MA, 02111, Email: MFinkelman@tuftsmedicalcenter.org.

Appendix:
Kullback-Leibler Information and Mutual Information

In general, the Kullback-Leibler information $K(f,g)$ is the expected log likelihood ratio of two probability distributions $f(x)$ and $g(x)$, where the expectation is taken with respect to $f(x)$:

$$K(f,g) = E_f\left\{\log\frac{f(x)}{g(x)}\right\} \tag{A1}$$

(Chang & Ying, 1996; Cover & Thomas, 1991). For CDMs, the goal is to discern between different attribute patterns, and the random variable in question is the response to a given item. If the item is dichotomous, as is assumed by the reduced RUM, the Kullback-Leibler information of item *i* for attribute patterns $\boldsymbol{\alpha}'$ and $\boldsymbol{\alpha}''$ is equal to

$$K_i(\boldsymbol{\alpha}',\boldsymbol{\alpha}'') = P_i(\boldsymbol{\alpha}')\log\left[\frac{P_i(\boldsymbol{\alpha}')}{P_i(\boldsymbol{\alpha}'')}\right] + \left[1-P_i(\boldsymbol{\alpha}')\right]\log\left[\frac{1-P_i(\boldsymbol{\alpha}')}{1-P_i(\boldsymbol{\alpha}'')}\right] \tag{A2}$$

(Henson & Douglas, 2005).

The general definition of MI, for any two random variables *Y* and *Z*, involves the marginal distributions of these variables, as well as their joint distribution. Let $h(y)$ and $h(z)$ represent the marginal distributions of *Y* and *Z*, respectively, and let $h(y,z)$ represent their joint distribution. Here, *y* and *z* denote realized values of *Y* and *Z*, respectively. The mutual information $I(Y;Z)$ between these variables is the Kullback-Leibler information between the joint distribution and the product of the marginal distributions. Applying Equation A1, this can be written as

$$I(Y;Z) = \sum_y\sum_z h(y,z)\log\frac{h(y,z)}{h(y)h(z)} \tag{A3}$$

(Cover & Thomas, 1991; Wang, 2013; Wang & Chang, 2011; Weissman, 2007).

MI can be understood intuitively by noting that if *Y* and *Z* are independent, then the joint distribution is equal to the product of marginal distributions: $h(y,z) = h(y)h(z)$ for all *y* and *z*. In this case $\log\frac{h(y,z)}{h(y)h(z)} = 0$ and hence $I(Y;Z) = 0$, which is consistent with the intuitive notion that independent variables provide no information about each other. However, if *Y* and *Z* are not independent, it can be shown that $I(Y;Z) > 0$, with larger values of $I(Y;Z)$ indicating greater information (Cover & Thomas, 1991).

The calculation in Equation 6 requires a number of different quantities. The marginal distribution of $\boldsymbol{\alpha}$ after *m* items have been administered is directly provided by the set of posterior probabilities $\{p_m(\boldsymbol{\alpha})\}$; these values represent the most current information regarding the unknown parameter $\boldsymbol{\alpha}$. The marginal distribution of an examinee's response to item *i* also involves the posterior probabilities $\{p_m(\boldsymbol{\alpha})\}$, as they affect the overall probability of a correct answer to the item. Specifically, after *m* items, the marginal probability of a correct answer to item *i* is obtained by computing the probability of a correct answer given $\boldsymbol{\alpha}$, then taking a summation weighted by the posterior distribution of $\boldsymbol{\alpha}$. The marginal probability of a correct answer to item *i* is thus

$$P_m(U_i = 1) = \sum_{\boldsymbol{\alpha}} p_m(\boldsymbol{\alpha}) P_i(\boldsymbol{\alpha}), \tag{A4}$$

where for the reduced RUM, $P_i(\boldsymbol{\alpha})$ is given by Equation 3. The marginal probability of an incorrect answer to item *i* after *m* items is $P_m(U_i = 0) = 1 - P_m(U_i = 1)$. Finally, the joint probability of an attribute pattern $\boldsymbol{\alpha}$ and a particular response u_i is equal to the marginal (posterior) probability of $\boldsymbol{\alpha}$ times the conditional probability of u_i given $\boldsymbol{\alpha}$. Hence, after *m* items, the joint probability of $\boldsymbol{\alpha}$ and a correct response to item *i* is equal to $P_m(\boldsymbol{\alpha}, U_i = 1) = p_m(\boldsymbol{\alpha}) P_i(\boldsymbol{\alpha})$; the joint probability of $\boldsymbol{\alpha}$ and an incorrect response to item *i* is equal to $P_m(\boldsymbol{\alpha}, U_i = 0) = p_m(\boldsymbol{\alpha}) \left[1 - P_i(\boldsymbol{\alpha})\right]$. These quantities are combined to provide the formula for $I_m(\boldsymbol{\alpha}; U_i)$ given in Equation 6.