Balancing Information with Imbalanced Item Banks in Multidimensional CAT

Michael J. Culbertson

Paper presented at the annual meeting of the
National Council on Measurement in Education,
Philadelphia, PA, April 4–6, 2014.

Abstract

In multidimensional Computerized Adaptive Testing (CAT), if item banks provide more information on one dimension than others, most item selection algorithms will over-emphasize the information-rich dimension. This study evaluates two proposed selection criteria based on mutual information that are intended to equalize ability estimate precision across dimensions.

1 Introduction

Computerized Adaptive Testing (CAT) provides a technological framework that is very well suited for diagnostic assessment: CAT can be administered quickly and easily in almost any location, the test specifications (e.g. content covered) can be adjusted on the fly, and the results are available to students and instructors immediately for rapid decision-making. Although unidimensional adaptive testing (originally known as "tailored" testing) has been studied since the 1970s (e.g. Lord, 1971, 1977; Weiss, 1982), CAT with the multidimensional continuous or discrete knowledge models suitable for effective diagnosis has received limited attention only recently (e.g. McGlohen & Chang, 2008; Segall, 1996). Given the potential for multdimensional models to improve the diagnostic utility of educational assessments, it will be key to understand how these models perform under CAT conditions for them to find most widespread and beneficial use in authentic educational settings.

A key principle of CAT is that test items are not equally informative about person parameters at different points on the latent scale. Rather, a (dichotomous) item is most informative when an examinee has about a 50% probability of answering it correctly. Since examinees have different ability levels, the same set of items will not be equally informative for all examinees. A CAT attempts to select the most informative items for a particular examinee from a large pool of available items, based on the evolving estimate of the examinee's ability. This optimal selection of the most informative items yields greater testing efficiencies, which become particularly crucial in diagnostic testing where the goal is to estimate many different sub-domain abilities without overburdening the examinee with an exceptionally long test.

One of the primary design considerations in CAT beyond those found in fixed-form assessment is the algorithm for selecting optimally informative items. For continuous multidimensional IRT models, algorithms have been developed based on both Fisher information and Kullback-Leibler (KL) divergence (e.g. Mulder & Linden, 2010; Mulder & van der Linden, 2009); for discrete models, indices based on KL divergence and Shannon entropy have been proposed (e.g. McGlohen & Chang, 2008). One convenient feature of the KL-based indices is their applicability to both continuous and discrete models (or to hybrid models). Moreover, Wang and Chang (2011) provided results suggesting superior performance in the continuous case of an index based on mutual information. Mutual information measures the reduction in uncertainty of one random variable due to knowledge of another (Cover & Thomas, 2006), and thus lends itself well conceptually to CAT item selection: The best item is the one that provides the greatest reduction in uncertainty about the latent variable, i.e. the item with the greatest mutual information with the latent variable. The mutual information between an item and the posterior distribution of the latent variable is equivalent to the KL divergence between subsequent posterior distributions after the item is observed (Mulder & Linden, 2010), which provides another convenient interpretation for the item selection metric—choosing the item that provides the greatest expected change in the posterior distribution after it is observed.

The mutual information between item x_k and the latent variable's posterior distribution $\boldsymbol{\theta}|\boldsymbol{x}_{k-1}$ is given by:

$$\begin{aligned} I_M(\boldsymbol{\theta}, x_k|\boldsymbol{x}_{k-1}) &= \sum_{x_k} \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, x_k|\boldsymbol{x}_{k-1}) \log \frac{f(\boldsymbol{\theta}, x_k|\boldsymbol{x}_{k-1})}{f(\boldsymbol{\theta}|\boldsymbol{x}_{k-1}) f(x_k|\boldsymbol{x}_{k-1})} d\boldsymbol{\theta} \\ &= \sum_{x_k} \int_{\boldsymbol{\theta}} f(x_k|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{x}_{k-1}) \log \frac{f(x_k|\boldsymbol{\theta})}{f(x_k|\boldsymbol{x}_{k-1})} d\boldsymbol{\theta} \\ &= \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{x}_{k-1}} \left[\mathrm{E}_{x_k|\boldsymbol{\theta}} \left[\log \frac{f(x_k|\boldsymbol{\theta})}{f(x_k|\boldsymbol{x}_{k-1})} \right] \right], \end{aligned}$$

where $f(x_k|\boldsymbol{x}_{k-1})$ is the posterior predictive density:

$$f(x_k|\boldsymbol{x}_{k-1}) = \int_{\boldsymbol{\theta}} f(x_k|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{x}_{k-1}) d\boldsymbol{\theta} = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{x}_{k-1}} \left[f(x_k|\boldsymbol{\theta}) \right].$$

Now, in a (high-dimensional) diagnostic context with continuous variables, computing the integration for mutual information quickly becomes quite onerous. However, these integrals can be framed as an expectation over the posterior distribution for the latent variable, which renders the computation amenable for Markov Chain Monte Carlo (MCMC) techniques (cf. Patz & Junker, 1999). Despite being suggested (Almond & Mislevy, 1999), using MCMC in CAT has not been studied extensively.

As the CAT progresses, a number of similar distributions must each be computed (namely, the successive posterior distributions for the latent ability), and it is possible that the burn-in period necessary for each successive distribution may decrease as the posterior distribution focuses more tightly on the final ability estimate.

Unidimensional CAT often requires non-statistical constraints to balance domain content. CAT based on multidimensional models could possibly provide a more measurement-informed alternative to non-statistical constraints, since the item selection algorithm can automatically balance measurement precision of different sub-domains instead of balancing raw numbers of items; however, this built-in content balancing is threatened if imbalanced item information leads to more items selected from information-rich sub-domains. This can happen because the mutual information criterion (like many other proposed criteria) allows for "compensatory" improvement in the posterior ability distribution—since information is only considered overall, even if the posterior variance for one dimension is very small, the criterion will continue to select items from that dimension if they uniformly continue to reduce the posterior variance of the precise dimension more than items for the imprecise dimension would reduce its (relatively larger) posterior variance. An altered criterion that considers only the variance of the least-precise dimension may provide more uniform measurement precision across the model.

The purpose of this paper is to examine performance of the mutual information item selection criterion with multidimensional models when the item bank provides different levels of information for different dimensions.

2 Methods

Simulations were conducted with two continuous latent dimensions. The covariance of the latent variables was parameterized as $\boldsymbol{\theta} = (I_p - \Gamma)^{-1}\boldsymbol{\xi}$, where $p = 2$ is the number of dimensions, I_p is the p-by-p identity matrix, $\boldsymbol{\xi} \sim N_p(0, I_p)$ is an uncorrelated standard multivariate normal vector, and Γ is a p-by-p matrix of path coefficients. In this case, only one entry of Γ was non-zero, yielding $\theta_1 = \xi_1$ and $\theta_2 = \gamma\xi_1 + \xi_2$, and γ is the covariance between θ_1 and θ_2. In a graphical modeling context, θ_1 would be called the "parent" of θ_2 (the "child"). Each item loaded onto to only one of the dimensions, and the items followed a 2-parameter logistic (2PL) model:

$$P(X_j = 1|\theta) = \frac{\exp[a_j\theta_k - b_j]}{1 + \exp[a_j\theta_k - b_j]},$$

where θ_k is the component of $\boldsymbol{\theta}$ corresponding to the dimensions for the given item. Note that in this parameterization of the 2PL model, a and b are scaling and location parameters for θ (as opposed to the usual $\exp[a(\theta - b)]$).

The simulation study had four factors (Table 1): dimension correlation, information distribution (balanced, imbalanced), test length, and item selection criterion.

Table 1: Adaptive Item Selection Simulation Design

Factor	N	Levels
Sub-Domain Relationships	4	Independent, Weakly coupled, Strongly coupled
Information Distribution	3	Balanced, Parent-Favored, Child-Favored
Test Length	3	6, 12, 20
Item Selection Criterion	5	Unrestricted, Item-Restricted, Dimension-Restricted (point-estimate), Dimension-Restricted (expected), Random

Three different correlations were implemented:

- independent sub-domains ($\gamma = 0$),
- weakly coupled sub-domains ($\gamma = 0.45$, correlation $= 0.4$), and
- strongly coupled sub-domains ($\gamma = 1.0$, correlation $= 0.7$).

The item bank was composed of 50 items from each dimension. Test length was 6, 12, or 20.

Item parameters came from three conditions: balanced, parent-favored, and child-favored. In the balanced information condition, item discrimination parameters were sampled from the same distribution $a \sim \sigma_k^{-1}\text{Uniform}(0.6, 1.4)$ for all sub-domains, where σ_k is standard deviation for sub-domain k:

$$\sigma_k^2 = \sum_v r_{kv}^2, \quad R = (I_p - \Gamma)^{-1}.$$

In the parent-favored conditions, item discriminations parameters for parent sub-domains were sampled from a high-information distribution $a \sim \sigma_k^{-1}\text{Uniform}(1.2, 1.6)$, and item discrimination for the remaining items were sampled from a low-information distribution $a \sim \sigma_k^{-1}\text{Uniform}(0.4, 0.8)$. In the child-favored conditions, the distributions for item-discrimination parameters were reversed. Item difficulty parameters were equally spaced on $b \in [-1.5, 1.5]$.

Five item selection criteria were compared:

- the overall mutual information criterion, $I_M(\boldsymbol{\theta}, x_k|\boldsymbol{x}_{k-1})$, (unrestricted);
- the overall mutual information criterion, computed only for items that measure the sub-domain θ_* with the largest posterior variance (item-restricted);
- mutual information between items and the posterior distribution only of the least-precise sub-domain using the current expected a posterior estimate for the remaining sub-dimensions ($\boldsymbol{\theta}_\dagger$), $I_M(\theta_*, x_k|\boldsymbol{x}_{k-1}, \hat{\boldsymbol{\theta}}_\dagger)$, (dimension-restricted, point-estimate);

- the expected mutual information between items and the posterior distribution of the least-precise sub-domain, $\mathrm{E}_{\boldsymbol{\theta}_\dagger|\boldsymbol{x}_{k-1}}\left[I_M(\theta_*, x_k|\boldsymbol{x}_{k-1}, \boldsymbol{\theta}_\dagger)\right]$ (dimension-restricted, expected); and
- random item selection.

The mutual information integration was computed via MCMC for the joint posterior knowledge distribution (see Section 3). For the item- and dimension-restricted conditions, the focal sub-domain for the first item was selected randomly.

For each condition, 1000 examinees were simulated. Disturbances ξ for each sub-domain were sampled from independent standard normal distributions. Expected a posteriori (EAP) person parameters were obtained via MCMC. Based on brief preliminary studies, interim person parameter estimates were based on 20,000 draws from the Markov chain, and final person parameter estimates were based on 40,000 draws. Since the difference between the prior distribution (standard normal) and the posterior distribution after the first item, as well as the the difference between successive posterior distributions, was not very large, no draws were discarded as burn-in. To reduce the effect of autocorrelation, the chain was thinned to keep only 1 in 5 draws (i.e., for 20,000 draws for an interim person parameter estimate, the Markov chain was actually advanced 100,000 iterations). These parameters were chosen to be highly conservative (i.e., an overly large number of draws) in order to reduce artifacts from MCMC estimation, which was not related to the central purpose of this study, and the number of draws could likely be considerably reduced with additional research to optimize MCMC parameters for use in CAT.

Simulation results were evaluated by three criteria:

- person parameter recovery, both overall (Euclidean distance) and by sub-domain (root mean squared error);
- item pool usage, both overall and by sub-domain; and
- number of items tested per sub-domain.

Particular attention was also be given to the relative precision of person parameter estimates across sub-domains. Computation for this study was conducted using the Open Science Grid* (OSG; Pordes et al., 2007; Sfiligoi et al., 2009).

3 Computation of Mutual Information

As mentioned above, since the mutual information between an item and the posterior distribution for person parameters can be written as a posterior expectation,

*The Open Science Grid is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

the computation of mutual information is straightforward via MCMC for the posterior person parameters. For dichotomous items, let $P_{\boldsymbol{\theta}} = P(X_k = 1|\boldsymbol{\theta})$ and $\mathrm{E}P_{\boldsymbol{\theta}} = \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{x}_{k-1}}[P_{\boldsymbol{\theta}}]$. Then, the mutual information can be written as:

$$\begin{aligned} I_M(\boldsymbol{\theta}, x_k|\boldsymbol{x}_{k-1}) &= \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{x}_{k-1}}\left[\mathrm{E}_{x_k|\boldsymbol{\theta}}\left[\log\frac{f(x_k|\boldsymbol{\theta})}{f(x_k|\boldsymbol{x}_{k-1})}\right]\right] \\ &= \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{x}_{k-1}}\left[P_{\boldsymbol{\theta}}\log P_{\boldsymbol{\theta}} + (1-P_{\boldsymbol{\theta}})\log(1-P_{\boldsymbol{\theta}})\right] - \\ &\quad \mathrm{E}P_{\boldsymbol{\theta}}\log\mathrm{E}P_{\boldsymbol{\theta}} - (1-\mathrm{E}P_{\boldsymbol{\theta}})\log\left(1-\mathrm{E}P_{\boldsymbol{\theta}}\right). \end{aligned}$$

For N draws, $\boldsymbol{\theta}^t, t = 1, \ldots, N$, from the Markov chain for the posterior distributions $\boldsymbol{\theta}|\boldsymbol{x}_{k-1}$, compute the sample averages:

$$\begin{aligned} \bar{p}_1 &= N^{-1}\sum_t^N P_{\boldsymbol{\theta}^t} \\ \bar{p}_2 &= N^{-1}\sum_t^N P_{\boldsymbol{\theta}^t}\log P_{\boldsymbol{\theta}^t} + (1-P_{\boldsymbol{\theta}^t})\log(1-P_{\boldsymbol{\theta}^t}). \end{aligned}$$

Then, the approximate mutual information is simply:

$$I_M(\boldsymbol{\theta}, x_k|\boldsymbol{x}_{k-1}) = \bar{p}_2 - \bar{p}_1\log\bar{p}_1 - (1-\bar{p}_1)\log(1-\bar{p}_1).$$

The mutual information between items and the posterior distribution only of the least-precise sub-domain, θ_*, using the current expected a posterior estimate for the remaining sub-dimensions, $\hat{\boldsymbol{\theta}}_\dagger$, cannot be expressed as a posterior expectation, but require integration only over a single dimension. It can be computed as:

$$\begin{aligned} I_M(\theta_*, x_k|\boldsymbol{x}_{k-1}, \hat{\boldsymbol{\theta}}_\dagger) &= \sum_{x_k}\int_{\theta_*} f(\theta_*, x_k|\boldsymbol{x}_{k-1}, \hat{\boldsymbol{\theta}}_\dagger)\log\frac{f(\theta_*, x_k|\boldsymbol{x}_{k-1}, \hat{\boldsymbol{\theta}}_\dagger)}{f(\theta_*|\boldsymbol{x}_{k-1}, \hat{\boldsymbol{\theta}}_\dagger)f(x_k|\boldsymbol{x}_{k-1}, \hat{\boldsymbol{\theta}}_\dagger)}d\theta_* \\ &= \sum_{x_k}\int_{\theta_*} f(x_k|\theta_*, \hat{\boldsymbol{\theta}}_\dagger)f(\theta_*|\boldsymbol{x}_{k-1}, \hat{\boldsymbol{\theta}}_\dagger)\log\frac{f(x_k|\theta_*, \hat{\boldsymbol{\theta}}_\dagger)}{f(x_k|\boldsymbol{x}_{k-1}, \hat{\boldsymbol{\theta}}_\dagger)}d\theta_* \\ &\propto \sum_{x_k}\int_{\theta_*} f(x_k|\theta_*, \hat{\boldsymbol{\theta}}_\dagger)f(\boldsymbol{x}_{k-1}|\theta_*, \hat{\boldsymbol{\theta}}_\dagger)f(\theta_*)\log\frac{f(x_k|\theta_*, \hat{\boldsymbol{\theta}}_\dagger)}{f(x_k|\boldsymbol{x}_{k-1}, \hat{\boldsymbol{\theta}}_\dagger)}d\theta_*, \end{aligned}$$

where the proportionality constant does not depend on θ_* or x_k. The informed posterior predictive density $f(x_k|\boldsymbol{x}_{k-1}, \hat{\boldsymbol{\theta}}_\dagger)$ can be computed as:

$$\begin{aligned} f(x_k|\boldsymbol{x}_{k-1}, \hat{\boldsymbol{\theta}}_\dagger) &= \frac{f(x_k, \boldsymbol{x}_{k-1}|\hat{\boldsymbol{\theta}}_\dagger)}{f(\boldsymbol{x}_{k-1}|\hat{\boldsymbol{\theta}}_\dagger)} \\ &= \frac{\int_{\theta_*} f(x_k, \boldsymbol{x}_{k-1}|\theta_*, \hat{\boldsymbol{\theta}}_\dagger)f(\theta_*)d\theta_*}{\int_{\theta_*} f(\boldsymbol{x}_{k-1}|\theta_*, \hat{\boldsymbol{\theta}}_\dagger)f(\theta_*)d\theta_*}. \end{aligned}$$

Note that because the denominator above does not depend on x_k, it constitutes a constant offset for all items under consideration for selection, and thus does not need to be computed for item selection purposes. The integral can be computed via Gauss-Hermite quadrature with weights w_v and abscissae τ_v as:

$$f(x_k|\boldsymbol{x}_{k-1}, \hat{\boldsymbol{\theta}}_\dagger) \propto \sum_v w_v f(x_k, \boldsymbol{x}_{k-1}|\tau_v, \hat{\boldsymbol{\theta}}_\dagger).$$

Finally, the expected mutual information between items and the posterior distribution of the least-precise sub-domain is calculated similarly:

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\theta}_\dagger|\boldsymbol{x}_{k-1}}\left[I_M(\theta_*, x_k|\boldsymbol{x}_{k-1}, \boldsymbol{\theta}_\dagger)\right] &= \int_{\boldsymbol{\theta}_\dagger} f(\boldsymbol{\theta}_\dagger|\boldsymbol{x}_{k-1}) I_M(\theta_*, x_k|\boldsymbol{x}_{k-1}, \boldsymbol{\theta}_\dagger) d\boldsymbol{\theta}_\dagger \\
&= \sum_{x_k} \int_{\boldsymbol{\theta}} f(x_k|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{x}_{k-1}) \log \frac{f(x_k|\boldsymbol{\theta})}{f(x_k|\boldsymbol{x}_{k-1}, \boldsymbol{\theta}_\dagger)} d\boldsymbol{\theta} \\
&= \mathrm{E}_{\boldsymbol{\theta}|\boldsymbol{x}_{k-1}}\left[\mathrm{E}_{x_k|\boldsymbol{\theta}}\left[\log \frac{f(x_k|\boldsymbol{\theta})}{f(x_k|\boldsymbol{x}_{k-1}, \boldsymbol{\theta}_\dagger)}\right]\right].
\end{aligned}$$

For N draws, $\boldsymbol{\theta}^t, t = 1, \ldots, N$, from the Markov chain for the posterior distributions $\boldsymbol{\theta}|\boldsymbol{x}_{k-1}$, the expected mutual information-based selection index is thus:

$$\mathrm{E}\tilde{I}(\theta_*, x_k|\boldsymbol{x}_{k-1}) = N^{-1} \sum_t \sum_{x_k} f(x_k|\boldsymbol{\theta}^t) \log \frac{f(x_k|\boldsymbol{\theta}^t)}{\sum_v w_v f(x_k, \boldsymbol{x}_{k-1}|\tau_v, \boldsymbol{\theta}_\dagger^t)}.$$

4 Results

When the two dimensions are independent and information is balanced, the mutual information criteria all select an equal number of items from each sub-domain, and the person parameter estimates for each sub-domain have approximately the same error (Table 2). When one sub-domain has more information than the other, however, the full MI criterion selects many more items from the information-rich sub-domain, yielding much more precise estimates. The item-restricted and dimension-restricted MI criteria, on the other hand, select more items from the information-poor sub-domain, and thus achieve a better balance of estimation precision across sub-domains. The improvement in precision in the weaker sub-domain comes at the cost of precision in the stronger sub-domain, of course; and while the restricted criteria approximately equalize precision of sub-domain estimates, the overall multidimensional error (as measured by the Euclidean distance between the final person estimate and the true person parameter vectors) is larger for the restricted criteria.

For weakly correlated sub-domains, even when information for the linear combination of person disturbances ($\boldsymbol{\theta} = (I_p - \Gamma)^{-1}\boldsymbol{\xi}$) is balanced, the parent sub-domain

Table 2: Person Parameter Estimation and Number of Items by Sub-domain

				Parent		Child	
Relationship	Information	Criterion	Euclidean	RMSE	#	RMSE	#
Independent	Balanced	Full	0.73	0.56	6	0.60	6
		Item	0.72	0.56	6	0.59	6
		Point	0.74	0.60	6	0.59	6
	Parent	Full	0.80	0.44	9	0.82	3
		Item	0.85	0.65	3	0.71	9
		Point	0.86	0.66	3	0.70	9
Weak	Balanced	Full	0.75	0.56	6	0.64	6
		Item	0.76	0.60	5	0.62	7
		Point	0.73	0.59	5	0.57	7
	Parent	Full	0.87	0.45	9.6	0.90	2.4
		Item	0.90	0.70	2	0.74	10
		Point	0.93	0.72	2	0.77	10
	Child	Full	0.83	0.79	3	0.55	9
		Item	0.84	0.68	8	0.67	4
		Point	0.86	0.67	8	0.70	4
Strong	Balanced	Full	0.79	0.51	6	0.76	6
		Item	0.88	0.69	1	0.75	11
		Point	0.87	0.67	1	0.73	11
	Parent	Full	0.88	0.39	11	0.95	1
		Item	0.96	0.74	1	0.80	11
		Point	0.98	0.72	1	0.85	11
	Child	Full	0.87	0.72	1.3	0.72	10.7
		Item	0.93	0.74	1	0.77	11
		Point	0.94	0.78	0	0.78	12

Note: Results are shown for 12-item tests. Other conditions performed similarly (see section 6). Since the "parent" and "child" labels are reversible in the independent condition, only one set of imbalanced information results is shown.

obtains additional information from items for the child sub-domain. Thus, even though the full MI criterion selects the same number of items from each sub-domain, the difference in estimation precision between the parent and child sub-domains is slightly inflated; and, the dimension-restricted criteria select slightly more items for the child sub-domain to compensate. For strongly correlated sub-domains, this trend is enhanced, and the dimension-restricted criteria select items almost exclusively from the child sub-domain to compensate. Due to the sub-domain correlations under the conditions of this study, this still yields a substantial amount of information for the parent sub-domain. In fact, even when item discrimination parameters favor the child sub-domain, the restricted criteria still continue to select items from the child sub-domain, since the strong correlation between the two dimensions yields sufficient information about the weaker parent sub-domain with items from the stronger child sub-domain.

For the item pools in this study, the mutual information-based selection criteria tend to over-select a small subset (less than half) of the most-informative items, leaving over half of the items never selected. The number of items selected for each sub-domain (Table 2) reflect the item pool usage trends among the different selection criteria: The restricted criteria under-utilize items from the information-rich sub-domain and over-utilize items from the information-poor sub-domain. While the restricted criteria reduce the number of never-used items on the sub-domain they favor, these items become over-exposed, and the number of never-used items on the unfavored sub-domain increases, increasing disparities in item exposure rates.

5 Discussion

This study demonstrates that the restricted mutual information item selection criteria are effective at equalizing measurement precision in CAT when an item pool provides more-informative items for some dimensions than others. Equalizing the measurement precision across dimensions does come at a cost, though: Global measurement precision may decrease (as in this study), and item-pool usage degrades as item from more-informative sub-domains are under used and items from less-informative sub-domains are over exposed. If the purpose of a CAT is diagnostic and not summative, the former cost may be inconsequential, as the individual sub-domain estimates are of most interest. The latter cost may be mitigated by incorporating exposure control mechanisms. Alternatively, test developers may simply need to focus on producing more items for less-informative sub-domains to compensate for their overexposure. Moreover, when sub-domain relationships are strong, items from sub-domains at the bottom of the network (few descendants, many ancestors) tend to be heavily emphasized by the restricted MI criteria due to their relative poverty of information and the fact that they provide strong information for their ancestors, further reducing the need for selecting items from sub-domains high in the network (few ancestors, many

descendants). This may indicate a need for content balancing to ensure coverage of the unique content of these sub-domains.

Finally, in terms of sub-domain person parameter estimation and item usage, the item-restricted and dimension-restricted performed very similarly. On one hand, this may suggest an indifference between the techniques for operational usage. However, the restricted criteria differ markedly in their computational performance: The item-restricted criterion requires summation over the number of draws from the Markov chain, while the point-estimate dimension-restricted criterion sums only over a fixed number of Gaussian quadrature points. The expected dimension-restricted criterion is computationally most intensive, requiring summation over the product of draws from the Markov chain and the quadrature points. While the dimension-restricted criterion needs to be computed for more items than the item-restricted criterion, if the number of Markov chain draws is large, the point-estimate dimension-restricted criterion still takes considerably less time to compute. When the number of dimensions is relatively small (as in the simplified network in this study), this difference can be considerable, though as the number of dimensions increases, the time required to compute the interim ability estimates begins to dominate the time taken to compute the item selection criteria, and the difference between the item-restricted and point-estimate dimension-restricted becomes less important to the total CAT computation time. In a related study with 12 dimensions, using the same conservative MCMC parameters (i.e., an overly large number of draws), item selection using the item-restricted and point-estimate dimension-restricted criteria took approximately 3-4 seconds per item on average. While this may be rather slow for scenarios in which many tests must be managed by a single processor, in a distributed testing environment in which each examinee is sitting at a separate computer with its own processor, this computation time can be easily accommodated while examinees are reading and deliberating about each test item, and the time could likely be further reduced with optimized MCMC parameters.

References

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement*, *23*, 223–237. doi: 10.1177/01466219922031347

Cover, T. M., & Thomas, J. A. (2006). Elements of information theory. In (chap. 2). Wiley. doi: 10.1002/047174882X

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika*, *36*, 227–242. doi: 10.1007/BF02297844

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, *1*, 95–100. doi: 10.1177/014662167700100115

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

Mulder, J., & Linden, W. J. V. D. (2010). Multidimensional adaptive testing with Kullback-Leibler information item selection. In W. J. van der Linden & C. A. Glas (Eds.), *Elements of adaptive testing* (pp. 77–101). New York, NY: Springer New York. doi: 10.1007/978-0-387-85461-8

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika*, *74*, 273–296. doi: 10.1007/s11336-008-9097-5

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178. doi: 10.3102/10769986024002146

Pordes, R., Petravick, D., Kramer, B., Olson, D., Livny, M., Roy, A., … Quick, R. (2007). The Open Science Grid. *Journal of Physics: Conference Series*, *78*, 012057. doi: 10.1088/1742-6596/78/1/012057

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354. doi: 10.1007/BF02294343

Sfiligoi, I., Bradley, D. C., Holzman, B., Mhashilkar, P., Padhi, S., & Wurthwein, F. (2009). The pilot way to grid resources using glideinWMS. *2009 WRI World Congress on Computer Science and Information Engineering*, 428–432. doi: 10.1109/CSIE.2009.950

Wang, C., & Chang, H.-H. (2011). Item selection in multidimensional computerized adaptive testing—gaining information from different angles. *Psychometrika*, *76*, 363–384. doi: 10.1007/s11336-011-9215-7

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *4*, 273–285. doi: 10.1177/014662168200600408

6 Complete Results

This section provides the complete results for all conditions of the item selection study.

Table 3: Person Parameter Estimation and Number of Items by Sub-domain: 6-item tests

				Parent		Child	
Relationship	Information	Criterion	Euclidean	RMSE	#	RMSE	#
Independent	Balanced	Random	1.03	0.84	2.9	0.81	3.1
		Full	0.89	0.72	3.0	0.71	3.0
		Item	0.90	0.71	3.0	0.72	3.0
		Point	0.87	0.72	3.0	0.69	3.0
		Expected	0.84	0.64	3.0	0.70	3.0
	Parent	Random	1.05	0.73	3.0	0.94	3.0
		Full	0.99	0.50	5.9	1.03	0.1
		Item	0.96	0.72	2.0	0.81	4.0
		Point	0.98	0.75	2.0	0.81	4.0
		Expected	0.94	0.71	2.0	0.79	4.0
Weak	Balanced	Random	1.00	0.79	3.0	0.82	3.0
		Full	0.92	0.70	3.0	0.77	3.0
		Item	0.91	0.75	2.6	0.71	3.4
		Point	0.90	0.70	2.6	0.75	3.4
		Expected	0.92	0.70	2.6	0.76	3.4
	Parent	Random	1.02	0.71	3.0	0.90	3.0
		Full	0.96	0.52	6.0	0.99	0.0
		Item	1.06	0.84	1.0	0.85	5.0
		Point	1.05	0.81	1.0	0.86	5.0
		Expected	1.05	0.83	1.0	0.85	5.0
	Child	Random	1.02	0.87	3.0	0.78	3.0
		Full	0.99	0.93	0.0	0.64	6.0
		Item	0.98	0.79	4.0	0.77	2.0
		Point	0.95	0.75	4.0	0.79	2.0
		Expected	0.95	0.76	4.0	0.77	2.0
Strong	Balanced	Random	1.04	0.74	3.0	0.91	3.0
		Full	0.92	0.63	3.1	0.85	2.9
		Item	0.98	0.75	1.0	0.83	5.0
		Point	0.97	0.73	1.0	0.83	5.0
		Expected	0.97	0.74	1.0	0.81	5.0
	Parent	Random	1.03	0.72	3.0	0.93	3.0
		Full	0.96	0.54	6.0	0.97	0.0
		Item	1.06	0.80	1.0	0.90	5.0
		Point	1.05	0.80	1.0	0.88	5.0
		Expected	1.05	0.79	1.0	0.92	5.0
	Child	Random	1.06	0.82	3.0	0.88	3.0
		Full	0.97	0.80	0.0	0.76	6.0
		Item	0.98	0.78	1.0	0.82	5.0
		Point	0.96	0.76	0.0	0.79	6.0
		Expected	0.97	0.79	0.0	0.80	6.0

Note: Results are shown for 6-item tests. Since the "parent" and "child" labels are reversible in the independent condition, only one set of imbalanced information results is shown.

Table 4: Person Parameter Estimation and Number of Items by Sub-domain: 12-item tests

				Parent		Child	
Relationship	Information	Criterion	Euclidean	RMSE	#	RMSE	#
Independent	Balanced	Random	0.84	0.66	6.0	0.68	6.0
		Full	0.73	0.56	6.0	0.60	6.0
		Item	0.72	0.56	6.0	0.59	6.0
		Point	0.74	0.60	6.0	0.59	6.0
		Expected	0.76	0.60	6.0	0.62	6.0
	Parent	Random	0.90	0.60	5.9	0.83	6.1
		Full	0.80	0.44	9.0	0.82	3.0
		Item	0.85	0.65	3.0	0.71	9.0
		Point	0.86	0.66	3.0	0.70	9.0
		Expected	0.84	0.67	3.0	0.67	9.0
Weak	Balanced	Random	0.88	0.67	6.0	0.74	6.0
		Full	0.75	0.56	5.9	0.64	6.1
		Item	0.76	0.60	5.0	0.62	7.0
		Point	0.73	0.59	5.0	0.57	7.0
		Expected	0.74	0.59	5.0	0.61	7.0
	Parent	Random	0.92	0.60	6.0	0.86	6.0
		Full	0.87	0.45	9.6	0.90	2.4
		Item	0.90	0.70	2.0	0.74	10.0
		Point	0.93	0.72	2.0	0.77	10.0
		Expected	0.90	0.72	2.0	0.73	10.0
	Child	Random	0.95	0.84	6.0	0.68	6.0
		Full	0.83	0.79	3.0	0.55	9.0
		Item	0.84	0.68	8.0	0.67	4.0
		Point	0.86	0.67	8.0	0.70	4.0
		Expected	0.83	0.67	8.0	0.67	4.0
Strong	Balanced	Random	0.94	0.63	5.9	0.86	6.1
		Full	0.79	0.51	6.1	0.76	5.9
		Item	0.88	0.69	1.0	0.75	11.0
		Point	0.87	0.67	1.0	0.73	11.0
		Expected	0.91	0.71	1.0	0.77	11.0
	Parent	Random	0.95	0.58	6.0	0.94	6.0
		Full	0.88	0.39	11.1	0.95	0.9
		Item	0.96	0.74	1.0	0.80	11.0
		Point	0.98	0.72	1.0	0.85	11.0
		Expected	0.98	0.75	1.0	0.83	11.0
	Child	Random	0.95	0.73	6.0	0.81	6.0
		Full	0.87	0.72	1.3	0.72	10.7
		Item	0.93	0.74	1.0	0.77	11.0
		Point	0.94	0.78	0.0	0.78	12.0
		Expected	0.91	0.76	0.0	0.75	12.0

Note: Results are shown for tests with 12-item tests. Since the "parent" and "child" labels are reversible in the independent condition, only one set of imbalanced information results is shown.

Table 5: Person Parameter Estimation and Number of Items by Sub-domain: 20-item tests

				Parent		Child	
Relationship	Information	Criterion	Euclidean	RMSE	#	RMSE	#
Independent	Balanced	Random	0.74	0.58	10.0	0.60	10.0
		Full	0.59	0.47	10.2	0.48	9.8
		Item	0.61	0.50	10.0	0.48	10.0
		Point	0.62	0.50	10.0	0.50	10.0
		Expected	0.61	0.49	10.0	0.48	10.0
	Parent	Random	0.79	0.50	10.1	0.74	9.9
		Full	0.71	0.38	13.1	0.73	6.9
		Item	0.74	0.59	4.2	0.60	15.8
		Point	0.72	0.57	4.2	0.60	15.8
		Expected	0.73	0.57	4.2	0.59	15.8
Weak	Balanced	Random	0.75	0.58	10.1	0.63	9.9
		Full	0.64	0.49	9.6	0.53	10.4
		Item	0.65	0.52	8.2	0.52	11.8
		Point	0.64	0.51	8.2	0.53	11.8
		Expected	0.64	0.51	8.1	0.52	11.9
	Parent	Random	0.80	0.49	10.0	0.79	10.0
		Full	0.76	0.38	13.9	0.79	6.1
		Item	0.81	0.64	3.1	0.67	16.9
		Point	0.82	0.64	3.1	0.67	16.9
		Expected	0.78	0.62	3.1	0.64	16.9
	Child	Random	0.83	0.75	9.8	0.59	10.2
		Full	0.72	0.69	6.7	0.47	13.3
		Item	0.75	0.61	13.8	0.60	6.2
		Point	0.75	0.61	13.8	0.60	6.2
		Expected	0.73	0.61	13.8	0.56	6.2
Strong	Balanced	Random	0.84	0.55	10.0	0.80	10.0
		Full	0.71	0.45	10.1	0.68	9.9
		Item	0.87	0.71	1.0	0.74	19.0
		Point	0.89	0.71	1.0	0.78	19.0
		Expected	0.86	0.68	1.0	0.72	19.0
	Parent	Random	0.83	0.47	9.9	0.83	10.1
		Full	0.80	0.34	15.5	0.88	4.5
		Item	0.93	0.73	1.0	0.78	19.0
		Point	0.95	0.69	1.0	0.84	19.0
		Expected	0.94	0.72	1.0	0.80	19.0
	Child	Random	0.86	0.65	9.9	0.74	10.1
		Full	0.80	0.66	4.4	0.68	15.6
		Item	0.86	0.71	1.0	0.71	19.0
		Point	0.88	0.72	0.1	0.75	19.9
		Expected	0.90	0.77	0.2	0.77	19.8

Note: Results are shown for 20-item tests. Since the "parent" and "child" labels are reversible in the independent condition, only one set of imbalanced information results is shown.