Combining Computer Adaptive Testing Technology with Cognitively Diagnostic Assessment

by

Meghan K. McGlohen

Hua H. Chang

Joshua T. Wills

The University of Texas at Austin

I. Objectives

The goal of computerized adaptive testing (CAT) is to tailor a test to each individual examinee. CAT maintains many advantages in an assessment situation because it allows the test to hone in on the examinees' ability levels in an interactive manner. However, while traditional tests can accomplish assessment goals, such as a ranked comparison of examinees or grade assignments based on certain criteria, they do not provide individualized information to teachers or test-takers regarding specific content in the domain of interest (Chipman, Nichols, and Brennan, 1995). A new approach to educational research has begun to effloresce in order to address this issue. This research area, dealing with the application of cognitive diagnosis in the assessment process, aims to provide helpful information to parents, teachers, and students, which can be used to direct additional instruction and study to the areas needed most by the individual student.

Current approaches interested in cognitive diagnosis focus solely on the estimation of the knowledge state, or attribute vector. This study proposes the combination of the estimation of individual ability levels ($\hat{\theta}_j$) along with an emphasis on the diagnostic feedback provided by individual attribute vectors ($\hat{\underline{\alpha}}_j$), thus linking the current standard in testing technology with a new area of research aimed at helping teachers and students benefit from the testing process.

By combining the advantages of computerized adaptive testing with the helpful feedback provided by cognitively diagnostic assessment, this study proposes a method for customized diagnostic testing. The technique utilizes the shadow-testing algorithm to simultaneously optimize the estimation of both the ability level $\hat{\theta}_j$ and the attribute vector $\hat{\underline{\alpha}}_j$.

By combining the traditional IRT-based CAT model with a cognitive diagnosis model, examinees will be able to obtain both an ability level estimator as well as feedback concerning their states on the measured attributes instead of just one or the other.

II. Literature Review

Cognitively Diagnostic Assessment

Several models have been proposed to provide cognitively diagnostic information in the assessment process. Two of the founding models for cognitive diagnosis are Fisher's Linear Logistic Trait Model (1973) and Tatsuoka's and Tatsuoka's Rule Space methodology (1982). A plethora of additional models have used the concepts of these methods as foundations for new approaches or applications (for details, see Hartz, Roussos and Stout, 2002). Specifically, the Unified model, developed by DiBello, Stout and Roussos (1995) is based on the rule space model, and in turn, the Fusion model (Hartz *et al.*, 2002) is based on the Unified Model. The item response function for the Fusion model is illustrated below (in Equation 1), as described by Hartz, *et al.* (2002).

$$P\left(X_{ij}=1|\underline{\alpha}_j,\theta_j\right)=\pi_i^*\prod_{k=1}^{K}r_{ik}^{*\left(-\alpha_{jk}\right)\times q_{ik}}P_{c_i}\left(\theta_j\right) \tag{1}$$

where

$P_{c_i}\left(\theta_j\right)$ = The Rasch model with difficulty parameter c_i.

π_i^* = The probability of correctly applying all item *i* required attributes given that the individual processes all of the required attributes for the item *i*.

$r_{ik}^{*} = \frac{P(Y_{ijk} = 1 | \alpha_{jk} = 0)}{P(Y_{ijk} = 1 | \alpha_{jk} = 1)}$, considered an attribute-based item discrimination parameter for item *i* and attribute *k*.

Y_{jk} = 1 when examinee *j* correctly applies attribute *k* to item *i,* and 0 otherwise.

α_{jk} = 1 when examinee *j* has mastered attribute *k*, and 0 otherwise.

Also, the attribute vector for individual *j* is denoted as $\underline{\alpha}_j$, and θ_j is the residual ability parameter, which deals with content measured by the test that is not included in the Q-matrix. Though the notation is the same, this is not the single ability score θ_j of the Three Parameter Logistic model.

A software program called Arpeggio (Stout *et al.*, 2002) is available as a means of analyzing examinee response data to provide diagnostic feedback in the form of individual attribute vectors. The next helpful step in the diagnostic process would be to implement computer adaptive testing technologies to optimize the item selection process with respect to the estimation process.

Shadow Testing by means of Linear Programming

Shadow testing, a mode for test assembly which utilizes linear programming (LP) to incorporate constraints into the assembly process, was proposed by van der Linden and Reese (1998). It is an iterative process in which an ideal "shadow" test is formed before the administration of each item in an examination. Each shadow test must contain items already included in the test and must be optimal at the given estimate level while complying with all of the specified constraints (van der Linden and Chang, 2003). The best item on the shadow test is then selected as the next item to be administered. Then,

the response from this item is used in the process of formulating the next shadow test. The application of this approach results in two major advantages. First, the items actually administered in the adaptive test will follow the constraints because each of the shadow tests meets these specifications. Second, the adaptive test will converge optimally to the true value of the estimator because the shadow tests are assembled to be optimal for the current estimate level, and in turn, each selected item is the optimal one from that shadow test (van der Linden and Chang, 2003).

Once a shadow test is constructed, the best item with respect to the attribute vector estimate is selected to be the next item administered to the examinee. Two strategies, Shannon Entropy and Kullback-Leibler Information, are employed for the selection process as described in Xu, Chang and Douglas (2003).

Shannon Entropy was developed in 1948 as a measure of uncertainty from a probability standpoint. It is a nonnegative concave function of the probability distribution. In this context, Shannon Entropy is minimized; that is to say, it is desirable to have minimal uncertainty. Shannon Entropy is described by

$$Sh(\underline{\pi}) = \sum_{i=1}^{K} \pi_i \log\left(\frac{1}{\pi_i}\right) \qquad (2)$$

where π_i is the probability that the random variable of interest, call it *Y*, takes on a particular value y_i, and $\underline{\pi}$ is the probability vector containing the π_i's for all possible values of y_i (Xu *et al.*, 2003). In the context of diagnostic assessment, where we are interested in estimating attribute vectors, the function for Shannon Entropy becomes Equation 3, as described by Xu *et al.* (2003).

$$Sh\left(\underline{\pi_n}, X_j\right) = \sum_{x=0}^{1} E_n\left(\underline{\pi_n} \middle| X_j = x\right) \; P\left[X_j = x \middle| \underline{\pi_{n-1}}\right]$$

$$= \sum_{x=0}^{1} \{ \; E_n\left(\underline{\pi_n} \middle| X_j = x\right) \left(\sum_{c=1}^{2^M} P_j^x\left(\underline{\alpha_c}\right)\left[1 - P_j\left(\underline{\alpha_c}\right)\right]^{1-x} \pi_{n-1}\left(\underline{\alpha_c}\right) \right) \; \} \tag{3}$$

Kullback-Leibler (K-L) Information was introduced in 1951 as a distance measure between probability distributions. More recently, K-L Information can be used as a measure of global information for the purpose of item selection in IRT (Chang and Ying, 1996) and as an index in the item selection process in diagnostic assessment (Xu *et al.*, 2003).

$$K(f,g) = \int \log\left(\frac{f(x)}{g(x)}\right) f(x)\mu(dx) \tag{4}$$

For the cognitive diagnosis context, K-L Information is used as an item selection criterion. The integral become a sum because the variables are discrete; then the sum is taken across all possible attribute patterns. Thus the function becomes

$$K_j\left(\underline{\hat{\alpha}}\right) = \sum_{c=1}^{2^M} \left[\sum_{x=0}^{1} \log\left(\frac{P(X_j = x|\underline{\hat{\alpha}})}{P(X_j = x|\underline{\alpha_c})} \right) P\left((X_j = x|\underline{\hat{\alpha}})\right) \right] \tag{5}$$

where $\underline{\hat{\alpha}}$ is the currant estimate for the attribute vector and $\underline{\alpha_c}$ is the possible candidate attribute vector generated by the j^{th} item (Xu *et al.*, 2003). This yields an information index relating our current attribute vector estimate with the possible attribute vector estimate resulting from the administration of the next item (*j*) for every possible remaining item. The item with the largest value of $K_j\left(\underline{\hat{\alpha}}\right)$ is then selected as the next item.

In this study, maximizing Fisher Information with respect to the current estimate of the single score, $\hat{\theta}_j$, is used as the optimization function of the LP in the Shadow Test.

Then the minimization of Shannon Entropy and the maximization of K-L Information with respect to the attribute mastery vector, $\underline{\hat{\alpha}}_j$, are each used in different sub-conditions to select the best item from the Shadow Test. Thus, the item selected to be administered next will be a good item with respect to both theta and attribute mastery estimation an will obey any specified constraints in the LP. Van der Linden and Reese (1998) outline several non-statistical constraints of interest, including content balance, item time availability, and test length. This study will use a similar approach to applying these non-statistical constraints.

III. Method

The item parameters are pre-calibrated based on a simple random sample of 2000 examinees of three administrations of a state-mandated large-scale assessment using Bilog MG and Arpeggio 1.2 to obtain the 3PL and fusion model-based item parameters respectively. This large-scale assessment is comprised of a math portion and a reading portion. The item parameters from the three administrations are tripled to increase the size of the item bank. This results in items banks with 396 items for the math portion of the test and 324 items for the reading portion. The examinees' estimated parameters (both single score thetas and attribute mastery patterns) are used as true scores in the CAT simulation. Out of the 6000 examinees, 3000 are randomly sampled with replacement to be employed in the CAT simulation.

The study's design will include three conditions for comparison: one which selects items based on the $\hat{\theta}$ estimate only, one which selects items based on the $\underline{\hat{\alpha}}$ estimate only, and one which selects items based on both estimates. The first condition

will use the conventional method of focusing solely on $\hat{\theta}$ during item selection, and then will estimate the $\underline{\hat{\alpha}}$ vectors at the end using MLE estimation. The second condition, focusing solely on $\underline{\hat{\alpha}}$, will mimic the approach outlined in Xu *et al*. (2003) to select items and will calculate the $\hat{\theta}$ values afterwards based on the response patterns, also using MLE estimation.

The third condition first involves the construction of a Shadow Test that is optimized according to the ability level $\hat{\theta}$ (as outlined by van der Linden and Reese, 1998) before the administration of each item. Then the best item for measuring the attribute vector $\underline{\hat{\alpha}}$ is selected from the shadow test based on the current $\underline{\hat{\alpha}}$ using Shannon Entropy or Kullback-Leibler Information as outlined in Xu *et al*. (2003).

For conditions 1 and 3, the 5-4-3-2-1 item exposure control method proposed by McBride and Martin (1983) is implemented. Results of the three conditions are compared with regard to the accuracy of both the attribute classification rate and the ability levels by comparing the estimated values with the true values.

IV. Results

The evaluation of the information obtained from the various conditions involves three areas of examination. First, it is important that the various methods accurately estimated the values of the single score, theta. Second, the methods should also accurately estimate the attribute mastery patterns. An acceptable method would estimate both the theta values and the attribute mastery patterns well. Third, the item exposure rates of the various methods are examines because item exposure control is an important issue in test security in computerized adaptive testing. The following sections deal with

each of these issues. Optimally, examinees' single scores and attribute vectors should be accurately estimated while maintaining minimal item exposure rates for test security.

Theta Estimation

First, the theta estimates of the different conditions are of particular interest. The theta estimates are compared with the true theta values to determine how well each of the methods succeeds in accurately estimating the single score theta. The comparison between the true theta and its corresponding estimate is accomplished by examining the values of the correlation coefficient, the root mean square error, and the bias statistics.

Approaches where the probability of obtaining a correct response is based on the 3PL model are grouped together. Likewise, approaches where the probabilities for obtaining a correct response are based on the fusion model are grouped together. It would be impractical to compare results across these differences because how the response patterns are generated is a fundamental aspect of the CAT simulation and it would be ineffective to discern the differences do to model choice from the differences due to the various methods and conditions.

Correlation coefficients for all of the item selection methods within each condition are presented in Table 1 for the response probabilities based on the 3PL model and in Table 2 for the response probabilities based on the fusion model. Values of bias are presented in Table 3 for probabilities based on the 3PL model and in Table 4 for those based on the fusion model. The root mean square error values are presented in Table 5 for probabilities based on the 3PL model and in Table 6 for probabilities based on the fusion model.

Table 1: *Correlations of the true theta values and the estimated theta values for 3PL-based probabilities.*

	θ-Based Item Selection		θ- & α-Based Item Selection	
	Fisher	K-L	Shannon	K-L
Math				
Blueprint Q-matrix	0.965	0.966	0.969	0.971
Intuitive Q-matrix	0.967	0.967	0.975	0.975
Reading				
Blueprint Q-matrix	0.950	0.954	0.960	0.956
Intuitive Q-matrix	0.952	0.950	0.951	0.956

Table 2: *Correlations of the true theta values and the estimated theta values for fusion-based probabilities.*

	α-Based Item Selection		θ- & α-Based Item Selection	
	Shannon	K-L	Shannon	K-L
Math				
Blueprint Q-matrix	0.768	0.782	0.786	0.790
Intuitive Q-matrix	0.795	0.749	0.817	0.812
Reading				
Blueprint Q-matrix	0.718	0.763	0.752	0.762
Intuitive Q-matrix	0.203	0.222	0.227	0.230

Table 3: Bias statistics of the estimated theta values for 3PL-based probabilities.

	θ-Based Item Selection		θ- & α-Based Item Selection	
	Fisher	K-L	Shannon	K-L
Math				
Blueprint Q-matrix	-0.031	-0.036	-0.022	-0.023
Intuitive Q-matrix	-0.019	-0.008	-0.017	-0.011
Reading				
Blueprint Q-matrix	-0.062	-0.057	-0.047	-0.050
Intuitive Q-matrix	-0.049	-0.060	-0.055	-0.054

Table 4: *Bias statistics of the estimated theta values for fusion-based probabilities.*

	α-Based Item Selection		θ- & α-Based Item Selection	
	Shannon	K-L	Shannon	K-L
Math				
Blueprint Q-matrix	0.103	0.164	0.093	0.073
Intuitive Q-matrix	0.077	0.069	0.078	0.097
Reading				
Blueprint Q-matrix	0.160	0.144	0.061	0.081
Intuitive Q-matrix	0.674	0.113	-0.190	-0.426

Table 5: *Root Mean Square Error of the estimated theta values for 3PL-based probabilities.*

	θ-Based Item Selection		θ- & α-Based Item Selection	
	Fisher	K-L	Shannon	K-L
Math				
Blueprint Q-matrix	0.296	0.298	0.282	0.265
Intuitive Q-matrix	0.295	0.293	0.262	0.258
Reading				
Blueprint Q-matrix	0.335	0.324	0.302	0.318
Intuitive Q-matrix	0.324	0.334	0.329	0.317

Table 6: *Root Mean Square Error of the estimated theta values for fusion-based probabilities.*

	α-Based Item Selection		θ- & α-Based Item Selection	
	Shannon	K-L	Shannon	K-L
Math				
Blueprint Q-matrix	0.692	0.685	0.674	0.666
Intuitive Q-matrix	0.689	0.748	0.658	0.667
Reading				
Blueprint Q-matrix	0.748	0.681	0.696	0.687
Intuitive Q-matrix	1.351	1.174	1.364	1.345

In Table 1, all the correlations are above 0.9, and the values are similar across conditions 1 and 3. Correlations are typically lower for the reading test than the math test. Similarly, the bias values in Table 3 are all small and similar across the two conditions, but are smaller for the math test than the reading test. The root mean square

errors also tend to be greater for the reading test. Notice that in general, the methods perform more poorly on the reading test than the math test. Overall the math test seems to have more accurate estimates than the reading test, which indicates that the reading test is not as good at measuring a single overall reading score than the math test is at measuring a single overall math score.

In Table 5, the root mean square error is lower in condition 1 for the math test, but is lower for condition 3 in the reading test. Overall conditions 1 and 3 seem to perform comparably at accurately measuring the single score theta for the examinees when the 3PL model is used for calculating response probabilities. Within condition 1, item selection based on maximizing Fisher Information and K-L Information tend to perform equally well, as do minimizing Shannon Entropy and maximizing K-L Information within condition 3.

Now the results of the methods using the fusion model to calculated response probabilities are examined. As expected, conditions where the probabilities of obtaining a correct response are based on the fusion model do not estimate the values of the single score theta very well. This is intuitive because the value of theta is not present anywhere in the probability function for the item responses. Hence, the values of the correlation coefficients are lower and the root mean square error and bias values are greater than desired, but what is more interesting for this study is a comparison between the different methods within this fusion-based model approach. Higher correlations, lower bias estimates and lower root mean square error values in Tables 2, 4 and 6 respectively illustrate that condition 3 performs better than condition 2 at estimating single theta scores. This means that an item selection method that takes theta estimates and attribute

mastery patterns into account yields better theta estimates than the item selection method that only takes attribute mastery patterns into account. To evaluate which method(s) are best overall however, the accuracy of the attribute mastery classifications must also be considered.

Attribute Mastery Estimation

Optimally, an assessment approach will accurately estimate the attribute mastery of each attribute as well as the entire attribute pattern for the examinees. To evaluate the attribute mastery estimation, the correct classification rates of each measured attribute and the entire attribute pattern are presented in the following tables. Tables 7 presents these correct classification rates, or "hit rates," of each method using the 3PL to determine the response pattern probabilities for the math test. Table 8 holds the same information for the reading test. Tables 9 and 10 present the attributes' correct classification rates for the fusion model-based probabilities for the math test and reading test, respectively. A list of the attributes measured by each Q-matrix for each subject portion is presented in Appendix A.

Table 7: *The math test's attribute mastery hit rates using 3PL-based probabilities.*

		Condition 1		Condition 3	
Blueprint Q-matrix:	Attribute	Fisher	K-L	Shannon	K-L
	1	0.797	0.792	0.789	0.795
	2	0.712	0.696	0.715	0.703
	3	0.686	0.681	0.678	0.710
	4	0.710	0.725	0.718	0.703
	5	0.783	0.796	0.780	0.794
	6	0.833	0.827	0.833	0.816
	7	0.808	0.810	0.805	0.816
	8	0.814	0.817	0.808	0.818
	9	0.557	0.564	0.585	0.556
	10	0.807	0.796	0.825	0.823
	11	0.746	0.748	0.763	0.770
	Mean 1-11	0.750	0.750	0.754	0.755
	Whole Pattern	0.169	0.162	0.169	0.176
Intuitive Q-matrix:	Attribute	Fisher	K-L	Shannon	K-L
	1	0.835	0.838	0.863	0.838
	2	0.813	0.823	0.804	0.746
	3	0.826	0.824	0.834	0.833
	4	0.765	0.765	0.706	0.720
	5	0.795	0.806	0.772	0.779
	6	0.801	0.764	0.797	0.804
	7	0.741	0.749	0.711	0.720
	8	0.804	0.809	0.778	0.757
	9	0.800	0.813	0.848	0.828
	10	0.622	0.598	0.672	0.659
	11	0.847	0.860	0.790	0.776
	12	0.825	0.822	0.767	0.758
	13	0.685	0.664	0.646	0.622
	Mean 1- 13	0.782	0.780	0.768	0.757
	Whole Pattern	0.220	0.211	0.206	0.193

Table 8: *The reading test's attribute mastery hit rates using 3PL-based probabilities*.

		Condition 1		Condition 3	
Blueprint Q-matrix:	Attribute	Fisher	K-L	Shannon	K-L
	1	0.830	0.830	0.787	0.781
	2	0.847	0.850	0.868	0.863
	3	0.795	0.779	0.797	0.802
	4	0.863	0.861	0.856	0.864
	5	0.815	0.824	0.829	0.823
	6	0.827	0.834	0.843	0.861
	Mean 1-6	0.829	0.830	0.830	0.832
	Whole Pattern	0.586	0.590	0.583	0.580
Intuitive Q-matrix:	Attribute	Fisher	K-L	Shannon	K-L
	1	0.758	0.758	0.756	0.754
	2	0.814	0.812	0.799	0.808
	3	0.751	0.760	0.753	0.723
	4	0.753	0.744	0.767	0.754
	5	0.787	0.787	0.784	0.792
	6	0.766	0.767	0.762	0.769
	7	0.792	0.788	0.804	0.803
	Mean 1-7	0.774	0.774	0.775	0.772
	Whole Pattern	0.468	0.465	0.465	0.443

Table 9: *The math test's attribute mastery hit rates using fusion-based probabilities*.

		Condition 2		Condition 3	
Blueprint Q-matrix:	Attribute	Shannon	K-L	Shannon	K-L
	1	0.233	0.205	0.797	0.801
	2	0.849	0.871	0.800	0.801
	3	0.847	0.222	0.688	0.703
	4	0.887	0.847	0.781	0.797
	5	0.274	0.853	0.882	0.878
	6	0.356	0.877	0.890	0.876
	7	0.894	0.904	0.879	0.898
	8	0.816	0.907	0.883	0.885
	9	0.882	0.368	0.643	0.554
	10	0.939	0.997	0.937	0.939
	11	0.907	0.955	0.813	0.835
	Mean 1-11	0.717	0.728	0.817	0.815
	Whole Pattern	0.040	0.007	0.160	0.170
Intuitive Q-matrix:	Attribute	Shannon	K-L	Shannon	K-L
	1	0.905	0.242	0.882	0.855
	2	0.802	0.159	0.786	0.736
	3	0.925	0.876	0.896	0.876
	4	0.645	0.222	0.772	0.760
	5	0.873	0.916	0.871	0.872
	6	0.941	0.896	0.933	0.926
	7	0.837	0.261	0.726	0.746
	8	0.440	0.339	0.819	0.797
	9	0.910	0.978	0.904	0.913
	10	0.868	0.307	0.756	0.710
	11	0.734	0.260	0.846	0.832
	12	0.900	0.279	0.851	0.856
	13	0.858	0.265	0.705	0.660
	Mean 1- 13	0.818	0.461	0.827	0.811
	Whole Pattern	0.090	0.029	0.157	0.141

Table 10: *The reading test's attribute mastery hit rates using fusion-based probabilities*.

		Condition 2		Condition 3	
Blueprint Q-matrix:	Attribute	Shannon	K-L	Shannon	K-L
	1	0.899	0.833	0.799	0.803
	2	0.898	0.923	0.884	0.899
	3	0.890	0.822	0.855	0.844
	4	0.858	0.868	0.889	0.906
	5	0.911	0.807	0.869	0.867
	6	0.882	0.928	0.929	0.922
	Mean 1-6	0.890	0.863	0.871	0.874
	Whole Pattern	0.677	0.686	0.637	0.640
Intuitive Q-matrix:	Attribute	Shannon	K-L	Shannon	K-L
	1	0.929	0.967	0.924	0.941
	2	0.882	0.881	0.883	0.887
	3	0.888	0.898	0.896	0.889
	4	0.908	0.866	0.880	0.859
	5	0.900	0.867	0.867	0.868
	6	0.919	0.891	0.901	0.896
	7	0.893	0.862	0.893	0.885
	Mean 1-7	0.903	0.890	0.892	0.889
	Whole Pattern	0.711	0.722	0.699	0.708

This information may be more easily compared across the various approaches through graphical representation. The correct classification rates of the attribute mastery estimates are illustrated graphically in the following figures.

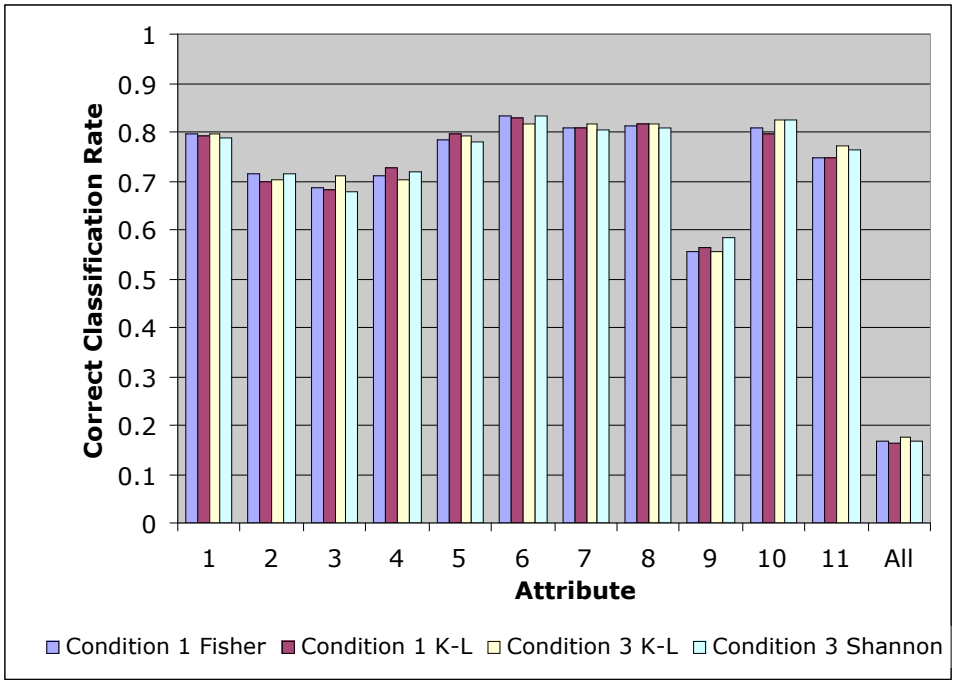

Figure 1: *Correct Attribute Mastery Classification for the Math Blueprint Q-matrix using 3PL-based Probabilities.*

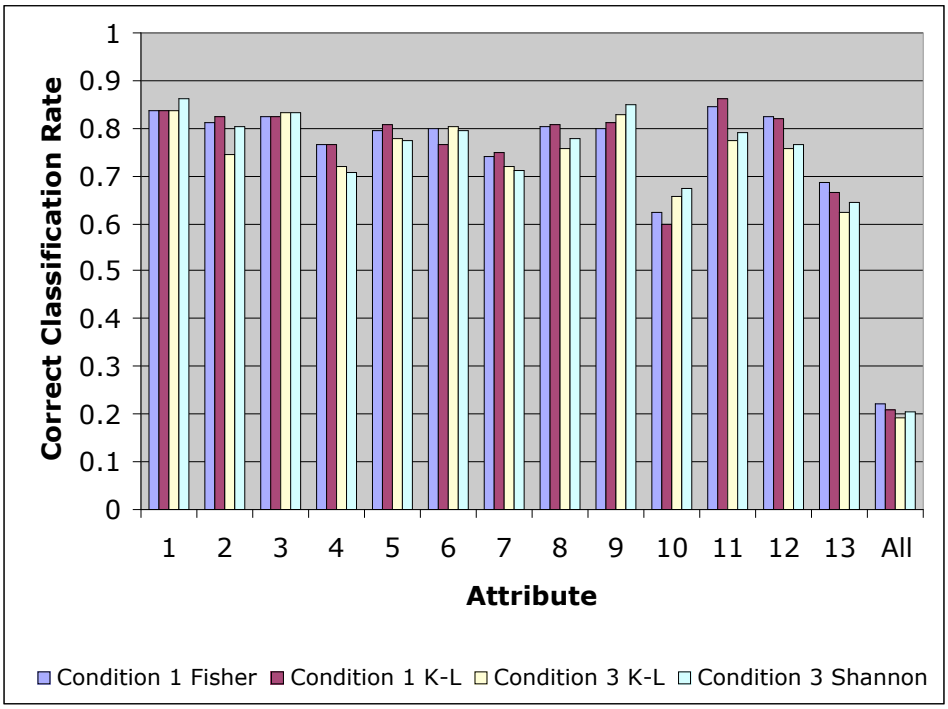

Figure 2: *Correct Attribute Mastery Classification for the Math Intuitive Q-matrix using 3PL-based Probabilities.*

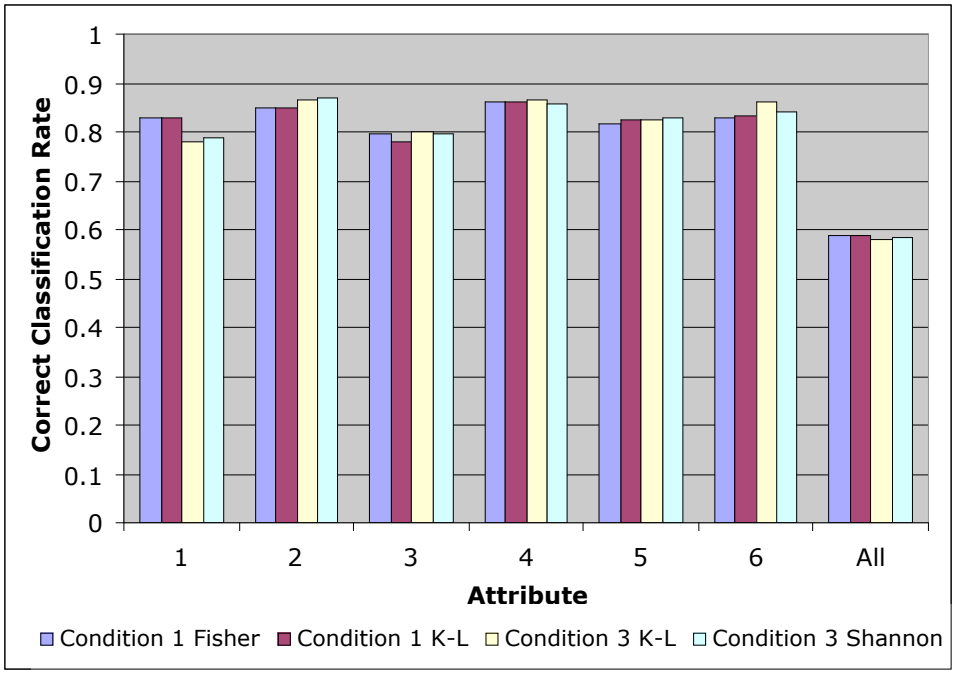

Figure 3: *Correct Attribute Mastery Classification for the Reading Blueprint Q-matrix using 3PL-based Probabilities.*

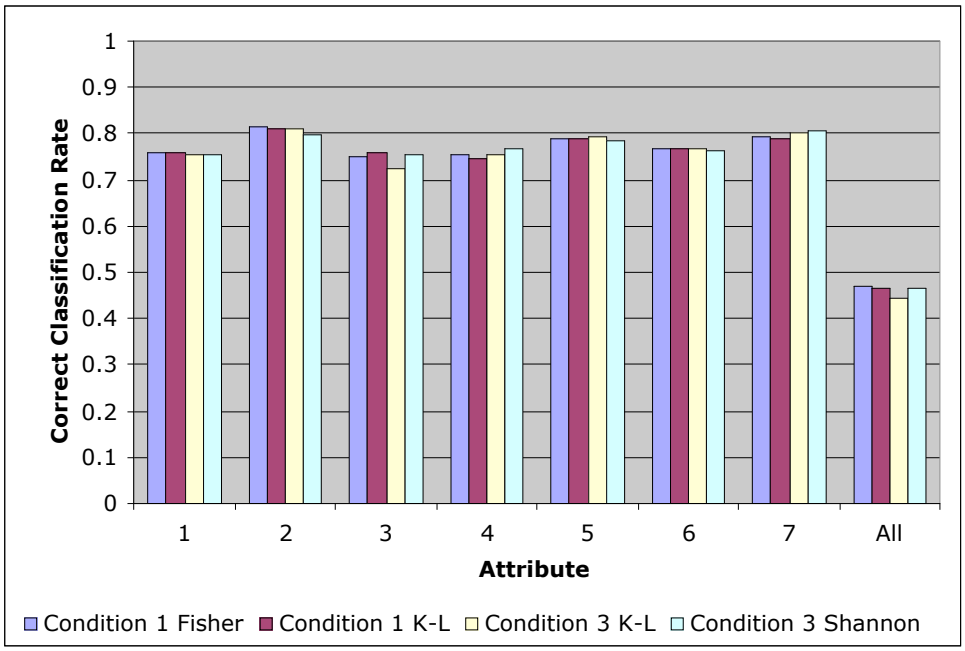

Figure 4: *Correct Attribute Mastery Classification for the Reading Intuitive Q-matrix using 3PL-based Probabilities.*

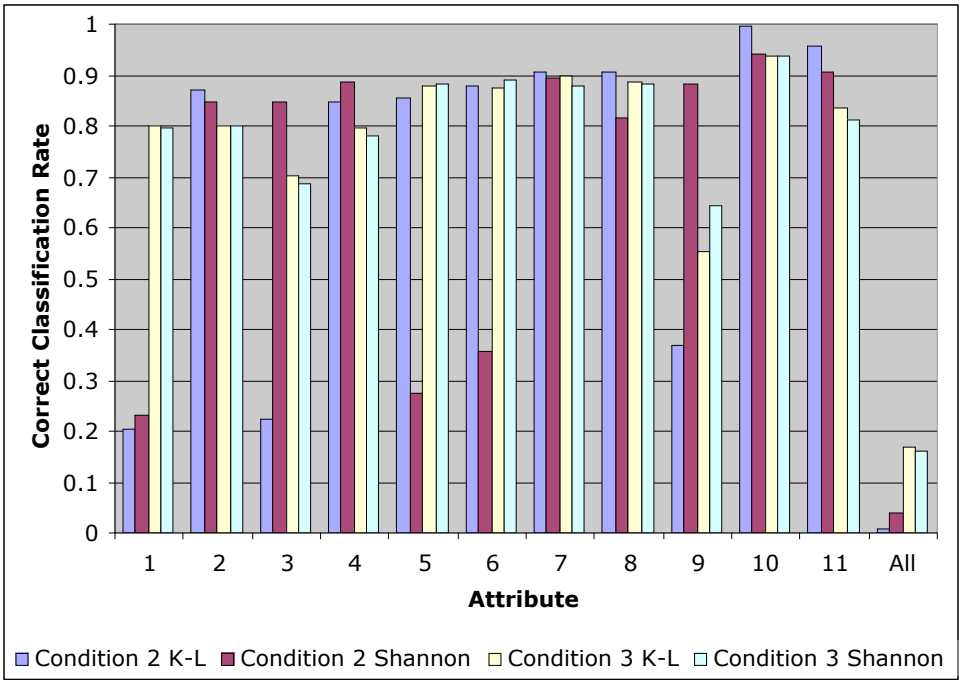

Figure 5: *Correct Attribute Mastery Classification for the Math Blueprint Q-matrix using fusion-based Probabilities*.

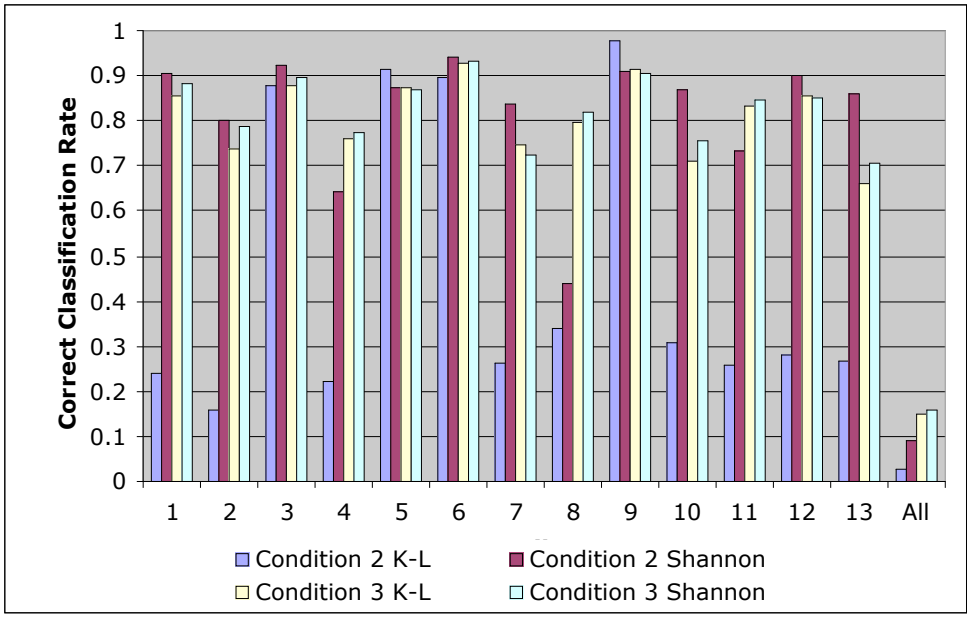

Figure 6: *Correct Attribute Mastery Classification for the Math Intuitive Q-matrix using fusion-based Probabilities*.

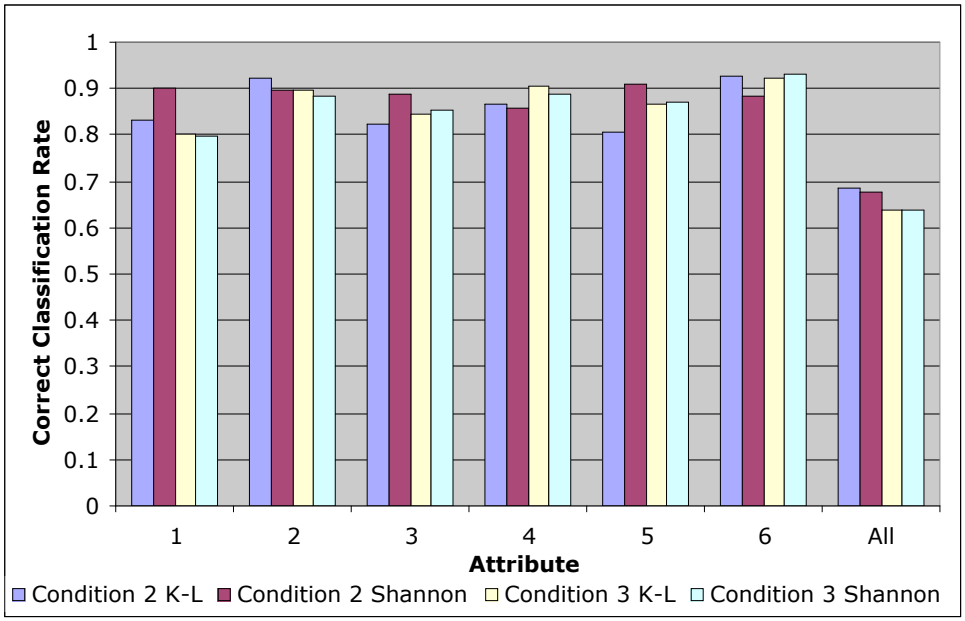

Figure 7: *Correct Attribute Mastery Classification for the Reading Blueprint Q-matrix using fusion-based Probabilities*.

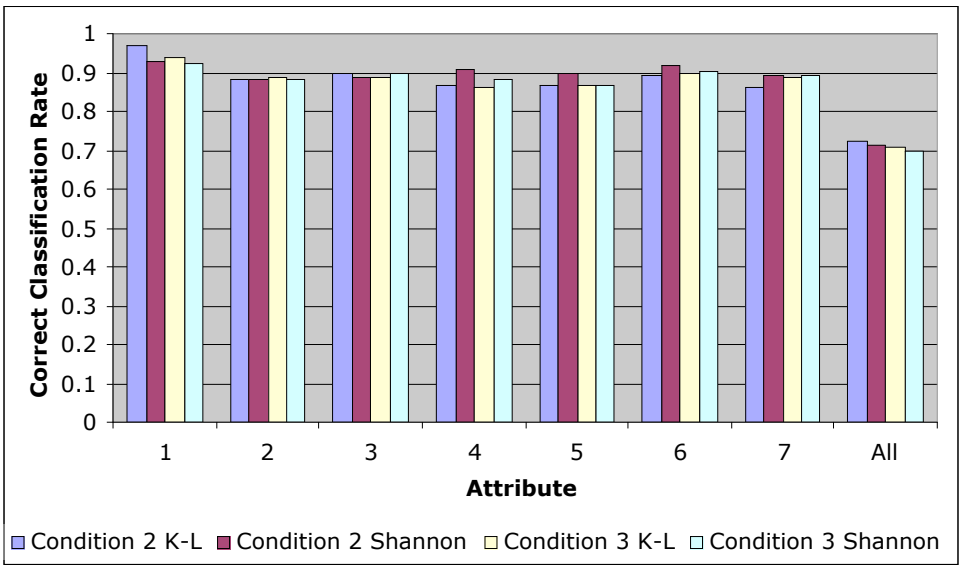

Figure 8: *Correct Attribute Mastery Classification for the Reading Intuitive Q-matrix using fusion-based Probabilities*.

With regard to the approaches based on the 3PL model for determining item response probabilities, condition 1 and condition 3 for math and reading portion of the

test on the individual attribute level as well as for the entire attribute pattern. An examination of the above figures indicates that within condition 1, the use of Fisher Information and K-L Information in selecting items optimally for the current theta estimate both perform equally well.

Results are more irregular for the methods based on the fusion model for item response probabilities. Notice that for the math test, condition 2 and condition 3 perform similarly, but a holistic examination shows that condition 3 more consistently classifies the examinees as masters or non-masters of the attributes, while the methods in condition 2 show more fluctuation. Within condition 2, using Shannon Entropy seems to produce more accurate attribute mastery classifications than using K-L Information. For the reading portion, the second condition yields slightly higher correct classifications for many of the attributes and for the overall mastery patterns, and overall the differences between the two conditions correct classification rates are quite small. It is surprising that condition 2 did not perform much better than condition 3. Condition 2 only selected items based on the current attribute mastery pattern estimate, $\underline{\alpha}_j$, and condition 3 takes both $\underline{\alpha}_j$ and θ_j into account, so it is logical that condition 2 would perform quite a bit better than condition 3 with regard to correct attribute mastery estimation, but the results were comparable. Thus, a test administrator would not have to sacrifice much attribute mastery classification precision in order to obtain higher precision in theta estimates. Within condition 3, utilizing K-L Information and Shannon Entropy seem to perform equally well in correctly estimating attribute mastery.

Item Exposure

In computerized adaptive testing, it is desirable to keep item exposure to a minimum to assure test security. But some items are better at measuring the underlying construct than others, so minimizing exposure control and maximizing measurement precision have a give-and-take relationship. Due to the importance of test security, the items exposure rated for the various methods implemented in this study will not be examined. The item exposure rates for the CAT-simulation based on the 3PL model are presented in Table 11 for the math portion of the test and Table 12 for the reading portion. Likewise, the item exposure rates for the CAT-simulation based on the fusion model are presented in Table 13 for the math portion of the test and Table 14 for the reading portion.

Table 11: *Item exposure for the math test using 3PL-based probabilities.*

	Condition 1				Condition 3			
Blueprint Q-matrix:	Fisher		K-L		Shannon		K-L	
Exposure Rate	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.
Not Exposed	114	0.29	153	0.39	137	0.35	147	0.37
0.0 to 0.099	169	0.43	136	0.34	134	0.34	139	0.35
0.1 to 0.199	39	0.10	36	0.09	42	0.11	32	0.08
0.2 to 0.299	29	0.07	28	0.07	46	0.12	23	0.06
0.3 to 0.399	13	0.03	8	0.02	15	0.04	20	0.05
0.4 to 0.499	10	0.03	9	0.02	4	0.01	17	0.04
0.5 to 0.599	10	0.03	12	0.03	5	0.01	6	0.02
0.6 to 0.699	10	0.03	8	0.02	9	0.02	10	0.03
0.7 to 0.799	2	0.01	6	0.02	4	0.01	2	0.01
0.8 to 0.899	0	0.00	0	0.00	0	0.00	0	0.00
0.9 to 1.000	0	0.00	0	0.00	0	0.00	0	0.00
Intuitive Q-matrix:	Fisher		K-L		Shannon		K-L	
Exposure Rate	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.
Not Exposed	152	0.38	170	0.43	126	0.32	122	0.31
0.0 to 0.099	132	0.33	114	0.29	149	0.38	154	0.39
0.1 to 0.199	31	0.08	33	0.08	39	0.10	40	0.10
0.2 to 0.299	32	0.08	29	0.07	22	0.06	26	0.07
0.3 to 0.399	21	0.05	19	0.05	36	0.09	24	0.06
0.4 to 0.499	11	0.03	12	0.03	16	0.04	18	0.05
0.5 to 0.599	5	0.01	7	0.02	6	0.02	7	0.02
0.6 to 0.699	10	0.03	7	0.02	2	0.01	3	0.01
0.7 to 0.799	2	0.01	5	0.01	0	0.00	2	0.01
0.8 to 0.899	0	0.00	0	0.00	0	0.00	0	0.00
0.9 to 1.000	0	0.00	0	0.00	0	0.00	0	0.00

Table 12: *Item exposure for the reading test using 3PL-based probabilities.*

	Condition 1				Condition 3			
Blueprint Q-matrix:	Fisher		K-L		Shannon		K-L	
Exposure Rate	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.
Not Exposed	100	0.25	116	0.29	121	0.31	123	0.31
0.0 to 0.099	115	0.29	100	0.25	115	0.29	103	0.26
0.1 to 0.199	32	0.08	37	0.09	22	0.06	34	0.09
0.2 to 0.299	28	0.07	21	0.05	12	0.03	12	0.03
0.3 to 0.399	13	0.03	17	0.04	13	0.03	17	0.04
0.4 to 0.499	14	0.04	9	0.02	18	0.05	5	0.01
0.5 to 0.599	11	0.03	8	0.02	6	0.02	9	0.02
0.6 to 0.699	8	0.02	9	0.02	6	0.02	13	0.03
0.7 to 0.799	0	0.00	4	0.01	5	0.01	5	0.01
0.8 to 0.899	3	0.01	2	0.01	6	0.02	3	0.01
0.9 to 1.000	0	0.00	1	0.00	0	0.00	0	0.00
Intuitive Q-matrix:	Fisher		K-L		Shannon		K-L	
Exposure Rate	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.
Not Exposed	100	0.25	111	0.28	103	0.26	109	0.28
0.0 to 0.099	115	0.29	106	0.27	115	0.29	114	0.29
0.1 to 0.199	32	0.08	36	0.09	32	0.08	30	0.08
0.2 to 0.299	26	0.07	22	0.06	16	0.04	13	0.03
0.3 to 0.399	8	0.02	13	0.03	14	0.04	19	0.05
0.4 to 0.499	20	0.05	11	0.03	22	0.06	13	0.03
0.5 to 0.599	10	0.03	10	0.03	15	0.04	13	0.03
0.6 to 0.699	8	0.02	6	0.02	1	0.00	10	0.03
0.7 to 0.799	3	0.01	7	0.02	6	0.02	3	0.01
0.8 to 0.899	0	0.00	2	0.01	0	0.00	0	0.00
0.9 to 1.000	0	0.00	0	0.00	0	0.00	0	0.00

Table 13: *Item exposure for the math test using fusion-based probabilities.*

	Condition 2				Condition 3			
Blueprint Q-matrix:	Shannon		K-L		Shannon		K-L	
Exposure Rate	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.
Not Exposed	156	0.39	311	0.79	160	0.40	181	0.46
0.0 to 0.099	134	0.34	22	0.06	116	0.29	108	0.27
0.1 to 0.199	42	0.11	15	0.04	51	0.13	34	0.09
0.2 to 0.299	11	0.03	8	0.02	29	0.07	19	0.05
0.3 to 0.399	13	0.03	0	0.00	8	0.02	17	0.04
0.4 to 0.499	11	0.03	0	0.00	12	0.03	7	0.02
0.5 to 0.599	13	0.03	2	0.01	3	0.01	17	0.04
0.6 to 0.699	9	0.02	3	0.01	9	0.02	3	0.01
0.7 to 0.799	5	0.01	8	0.02	7	0.02	8	0.02
0.8 to 0.899	1	0.00	6	0.02	1	0.00	2	0.01
0.9 to 1.000	1	0.00	21	0.05	0	0.00	0	0.00
Intuitive Q-matrix:	Shannon		K-L		Shannon		K-L	
Exposure Rate	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.
Not Exposed	127	0.32	205	0.52	139	0.35	148	0.37
0.0 to 0.099	145	0.37	115	0.29	143	0.36	130	0.33
0.1 to 0.199	50	0.13	21	0.05	43	0.11	44	0.11
0.2 to 0.299	28	0.07	13	0.03	19	0.05	18	0.05
0.3 to 0.399	21	0.05	2	0.01	19	0.05	28	0.07
0.4 to 0.499	17	0.04	0	0.00	19	0.05	9	0.02
0.5 to 0.599	5	0.01	5	0.01	5	0.01	7	0.02
0.6 to 0.699	1	0.00	9	0.02	5	0.01	6	0.02
0.7 to 0.799	1	0.00	16	0.04	4	0.01	5	0.01
0.8 to 0.899	1	0.00	2	0.01	0	0.00	1	0.00
0.9 to 1.000	0	0.00	8	0.02	0	0.00	0	0.00

Table 14: *Item exposure for the reading test using fusion-based probabilities.*

	Condition 2				Condition 3			
Blueprint Q-matrix:	Shannon		K-L		Shannon		K-L	
Exposure Rate	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.
Not Exposed	93	0.23	200	0.51	131	0.33	134	0.34
0.0 to 0.099	131	0.33	71	0.18	95	0.24	101	0.26
0.1 to 0.199	51	0.13	7	0.02	35	0.09	25	0.06
0.2 to 0.299	7	0.02	4	0.01	10	0.03	20	0.05
0.3 to 0.399	2	0.01	1	0.00	11	0.03	3	0.01
0.4 to 0.499	7	0.02	1	0.00	13	0.03	8	0.02
0.5 to 0.599	7	0.02	2	0.01	10	0.03	2	0.01
0.6 to 0.699	13	0.03	2	0.01	3	0.01	9	0.02
0.7 to 0.799	9	0.02	6	0.02	9	0.02	17	0.04
0.8 to 0.899	3	0.01	7	0.02	7	0.02	2	0.01
0.9 to 1.000	1	0.00	23	0.06	0	0.00	3	0.01
Intuitive Q-matrix:	Shannon		K-L		Shannon		K-L	
Exposure Rate	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.	Freq.	Prop.
Not Exposed	87	0.22	194	0.49	102	0.26	114	0.29
0.0 to 0.099	135	0.34	81	0.20	112	0.28	117	0.30
0.1 to 0.199	48	0.12	7	0.02	44	0.11	31	0.08
0.2 to 0.299	12	0.03	2	0.01	9	0.02	10	0.03
0.3 to 0.399	6	0.02	0	0.00	16	0.04	7	0.02
0.4 to 0.499	2	0.01	0	0.00	18	0.05	7	0.02
0.5 to 0.599	10	0.03	0	0.00	15	0.04	17	0.04
0.6 to 0.699	9	0.02	2	0.01	2	0.01	13	0.03
0.7 to 0.799	13	0.03	11	0.03	6	0.02	7	0.02
0.8 to 0.899	2	0.01	2	0.01	0	0.00	1	0.00
0.9 to 1.000	0	0.00	25	0.06	0	0.00	0	0.00

An example of desirable item exposure would have all of the items exposed to less than twenty percent of the examinees and would have no items that were not administered at all. The 5-4-3-2-1 exposure control method did not perform as ideally as this, but the exposure tendencies can be compared across the various methods. For instance, the method in condition 2 based on maximizing K-L Information has between two and six percent of the items exposed to al least ninety percent of the examinees. This

is an unacceptably high exposure rate. Condition 3 tends to have better exposure rates than condition 2, and tends to have comparable exposure rates to condition 1. The different methods within condition 1 and within condition 3 also tend to be comparable with regard to exposure control. If these item exposure rates are greater than desired for a particular assessment, alternative exposure control techniques might be an interesting area of future research.

Overall Performance

Evaluation of the different techniques encompasses an evaluation of theta estimation accuracy, attribute mastery estimation accuracy, and item exposure control. Overall results should thus be considered across all of the areas. With regard to estimating theta, conditions 1 and 3 produce comparable results, and condition 3 performs better than condition 2. Surprisingly, conditions 1 and 3 also produce comparable results with regard to the attribute mastery estimates. Condition 3 outperforms condition 2 with regard to theta estimation, attribute mastery pattern estimation, and item exposure control. But between condition 1 and condition 3, there is no clear-cut winner. Both methods perform well and similarly with regard to theta estimation, attribute mastery estimation, and item exposure control.

Notice that the results for the intuitive Q-matrix for the reading test are quite poor with respect to all three of the discussed criteria, and certainly performed more poorly than the corresponding results based on the Blueprint Q-matrix. This emphasizes the importance of Q-matrix construction. It would be inappropriate to underestimate the significance of Q-matrix development. Therefore, great care should go into this

important initial step in the cognitive diagnosis process. Additional precautions could be taken, such as asking a content expert to review a Q-matrix before the cognitively diagnostic analysis.

In sum, selecting items based on current theta estimates or based on both theta and attribute mastery estimates by means of shadow test are both good methods with regard to single score estimation, attribute mastery estimation and item exposure control. Determining which of these two to implement then depends on other issues the test administrator may be facing. If other constraints are necessary in the testing process, then the shadow test approach that selects items based on both single score and attribute mastery estimation is better because it can easily and efficiently incorporate the additional requirements. Such constrains include content balancing, item type constraints, testlet constraints, among others. Van der Linden and Reese (1998) and van der Linden (2000) expound on incorporating such constraints.

On the other hand, if a test administrator prefers a more simple approach to item selection and does not have access to special software like CPLEX, nor the need for additional constraints, then an item selection method based on the current single score estimates would suffice. Furthermore the different approaches within these item selection techniques (i.e. Fisher Information versus K-L Information and Shannon Entropy versus K-L Information) seem to have little difference in the results of this study, so a test administrator may have their choice between these.

V. Educational Implications

This study compares the accuracy of three possible item selection methods in a computerized adaptive testing situation focused on estimating diagnostic attribute information in addition to the conventional single score estimation. This simulation study is important in an educational context because it explores the accuracy of these methods with regard to these to assessment approaches. Test administrators can also use simulation studies like this one to determine how well the various attributes of interest are being measured. For instance, in Figures 1 and 2, the mastery status for attributes 9 and 10, respectively, are not estimated as accurately as the other attributes. Notice that both of these attributes deal with the students' ability to estimate a reasonable solution to the item. Test administrators could use this information to evaluate poorly measured attributes or to examine items measuring the attribute to try to better assess what they originally had in mind for this attribute. In addition, the attribute-based item difficulty parameter, π^*, and attribute-based item discrimination parameters, r^*, for the poorly estimated attribute(s) in this evaluation process. The results of this study also stress the importance of the construction of the Q-matrix in cognitively diagnostic assessment.

Cognitive diagnosis is important in educational assessment because it provides helpful feedback to students about specific elements of the measured content domain. It is rapidly becoming a requirement of effective, educationally beneficial test development (*No Child Left Behind Act, 2001)*. The challenge then becomes how to adapt the methods developed within the CAT framework to enable this new approach. This study proposes the application of the Shadow Test procedure to achieve the best of both worlds. While this study was conducted using the Fusion model's framework for cognitive diagnosis,

the procedure can be generalized to any diagnostic model which estimates the attribute states of the examinees, such as the Noisy Inputs Deterministic 'And' gate (NIDA) model (see Maris, 1999), the Generalized Latent Trait Model (GLTM) (Embretson, 1984), or the Rule Space method (Tatsuoka and Tatsuoka, 1982).

APPENDIX A
List of Attributes Measured by Each Test

Blueprint Q-matrix for the math test:

1. Demonstrate an understanding of number concepts.
2. Demonstrate an understanding of mathematical relations.
3. Demonstrate an understanding of geometric properties and relationships.
4. Demonstrate an understanding of measurement concepts using metric and customary units.
5. Demonstrate an understanding of probability and statistics.
6. Use the operation of addition to solve problems.
7. Use the operation of subtraction to solve problems.
8. Use the operation of multiplication and/or division to solve problems.
9. Estimate solutions to a problem situation and/or evaluate the reasonableness of a solution to a problem situation.
10. Determine solution strategies and analyze or solve problems.
11. Express or solve problems using mathematical representation.

Intuitive Q-matrix of the math test:

1. Understanding representation
2. Counting
3. Multiplication
4. Division
5. Addition
6. Subtraction
7. Understanding geometric shapes
   (turning, flipping, draw a line of symmetry, etc.)
8. Read a chart
9. Set up an arithmetic calculation from verbal information
10. Estimation
11. Read a table of numbers
12. Using standard units of measure
13. Understanding and forming order of magnitude

Blueprint Q-matrix for the reading test:

1. Determine the meaning of words in a variety of written texts.
2. Identify supporting ideas in a variety of written texts.
3. Summarize a variety of written texts.
4. Perceive relationships and recognize outcomes in a variety of written texts.
5. Analyze information in a variety of written texts in order to make inferences and generalizations.
6. Recognize points of view, propaganda, and/or statements of fact and opinion in a variety of written texts.

Intuitive Q-matrix of the reading test:

1. Chronology
2. Causality (determining why)
3. Word Meaning
4. General Summary
5. Observing/Remembering Details
6. Knowing Fact versus Opinion
7. Speculating from Contextual Clues

References

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

Chipman, S. F., Nichols, P. D. & Brennan, R. L. (1995). Introduction. In P. D. Nichols, S. F. Chipman, and R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (p. 327-361). Hillsdale, NJ: Lawrence Erlbaum Associates.

DiBello, L., Stout, W. & Roussos, L. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, and R. L. Brennan (Eds.), *Cognitively Diagnostic Assessment* (p. 327-361). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Fisher, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Hartz, S., Roussos, L. & Stout, W. (2002). Skills Diagnosis: Theory and Practice. User Manual for Arpeggio software. Princeton, NJ: Educational Testing Service.

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

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New Horizons in Testing* (pp. 223-236). New York: Academic Press.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*, 379-423, 623-656.

Stout, W. *et al*. (2002) Arpeggio Software Program, version 1.1. Princeton, NJ: Educational Testing Service

Tatsuoka, K. K. & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7*, 215-231.

U.S. House of Representatives (2001), 'Text of No Child Left Behind Act'.

Van der Linden, W. & Chang, H. (2003). Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach. *Applied Psychological Measurement, 27*, 107-120.

Van der Linden, W. & Reese, L. (1998, September). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259-270.

Xu, X., Chang, H., & Douglas, J. (2003). A simulation study to compare CAT strategies for cognitive diagnosis. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.