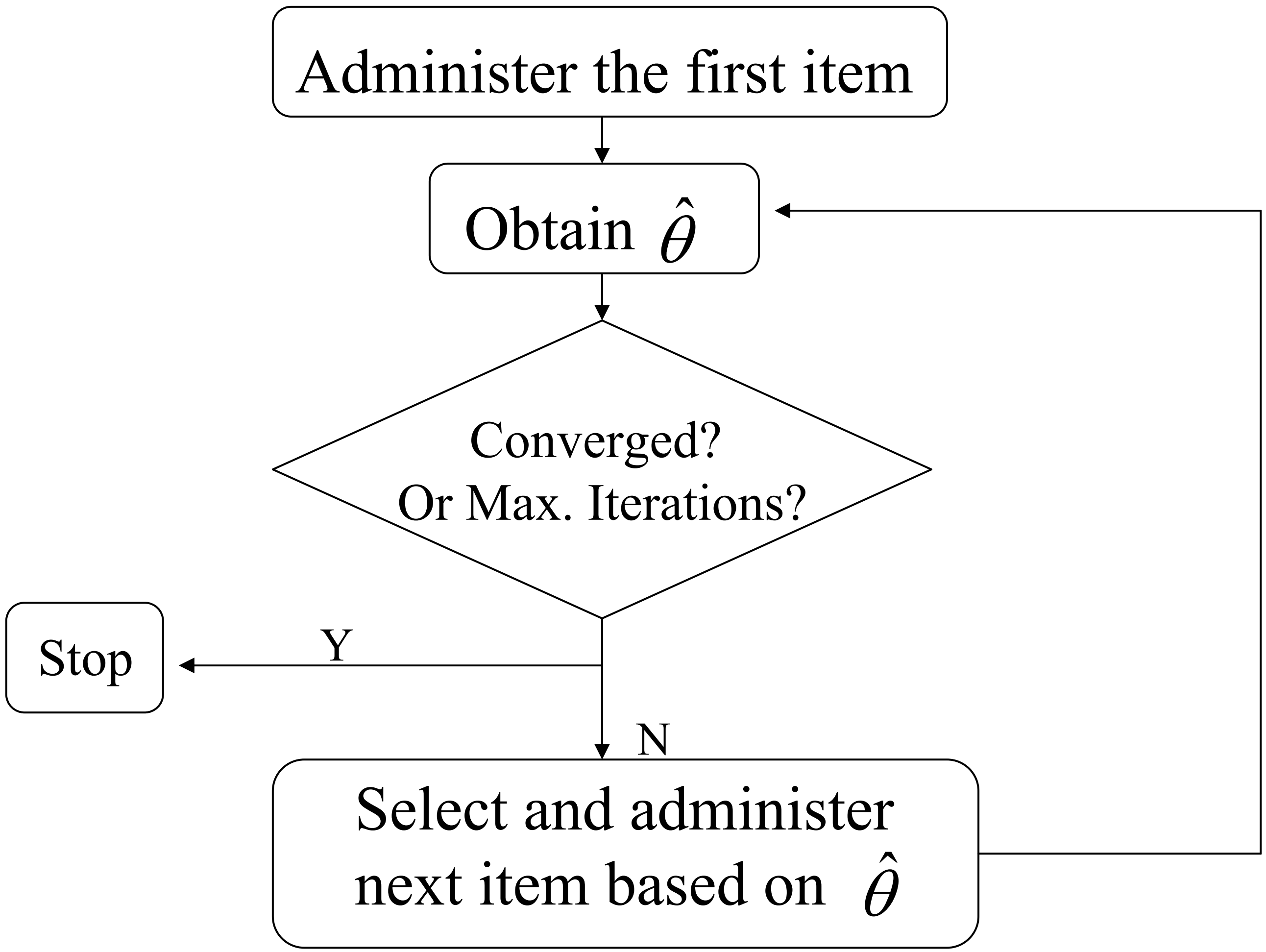
Computer Adaptive Testing for Cognitively Diagnostic Assessment

Hua-Hua Chang
Meghan Wills
Ying Cheng

Item Selection in Computerized Adaptive Testing (CAT)

- A test tailored to each examinee's latent trait level



Cognitive Diagnosis

Provide examinees with more information than just a single score.

- How? By considering the different *attributes* measured by the test.
- An attribute is a "task, subtask, cognitive process, or skill" assessed by the test, such as addition or reading comprehension.

The Item-Attribute Relationship

Which items measure which attributes is represented by the Q-matrix:

$$\begin{array}{c c} & \begin{array}{cccc} i1 & i2 & i3 & i4 \end{array} \\ \begin{array}{c} A1 \\ A2 \\ A3 \end{array} & \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

What is reported to examinees?

Traditional Testing:

$$\hat{\theta}$$

A single score

Cognitive Diagnosis:

$$\underline{\hat{\alpha}} = [\alpha_1, \alpha_2, \ldots, \alpha_K]$$

A <u>set</u> of scores:

One for each attribute.

(***K*** is the total # of attributes.)

Why is this beneficial?

Feedback from an exam can be more individualized to a student's specific strengths and weaknesses.

Julia R.

$\hat{\theta} = 75$

$$\underline{\hat{\alpha}} = [0000111]$$

Halle B.

$\hat{\theta} = 75$

$$\underline{\hat{\alpha}} = [0101100]$$

Models for CD

- Conjunctive latent class models
  - DINA model (e.g., Junker & Sijtsma, 2001)
  - NIDA model (Maris, 1999)
  - Fusion model (Hartz, 2002)

Why Use These Models?

1. Students are assessed with respect to the attributes

2. Items are assessed with respect to their capacity to measure the attributes

3. The parameters are estimable

How can we include both?

Traditional Testing

(Single Score, $\hat{\theta}$)

Cognitive Diagnosis

(Set of Scores, $\underline{\hat{\alpha}}$)

CAT administration

This project looks at three possible approaches to item selection.

Several Approaches

Select items based on:

a) Traditional θ estimates only

b) Cognitively diagnostic $\underline{\alpha}$ estimates only

c) Both θ and $\underline{\alpha}$ estimates simultaneously

Item Selection in Traditional CAT Maximum Fisher Information Method

- Three-parameter logistic model (Birnbaum, 1968)

$$P(Y=1\,|\,\theta)=c+(1-c)\frac{1}{1+e^{-a(\theta-b)}}$$

- Maximum Fisher Information Method

$$I(\hat{\theta})=\frac{(1-c)a^2e^{a(\hat{\theta}-b)}}{[1+e^{a(\hat{\theta}-b)}]^2[(1-c)+c(1+e^{a(\hat{\theta}-b)})]}$$

$$Var(\hat{\theta}_n)=1/I(\theta)\ \ as\ n\rightarrow\infty$$

DINA Model
(Deterministic Input; Noisy "And" Gate)

- An examinee who masters all attributes required by an item should be able to answer it correctly.

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \longrightarrow$$

$\eta_{ij} = 1$ Person *i* masters all attributes required by item *j*

$\eta_{ij} = 0$ Otherwise

- Slipping (s) and Guessing (g)

- Item response function

$$P\left(Y_{ij} = 1 \mid \alpha\right) = \left(1 - s_j\right)^{\eta_{ij}} g_j^{1-\eta_{ij}}$$

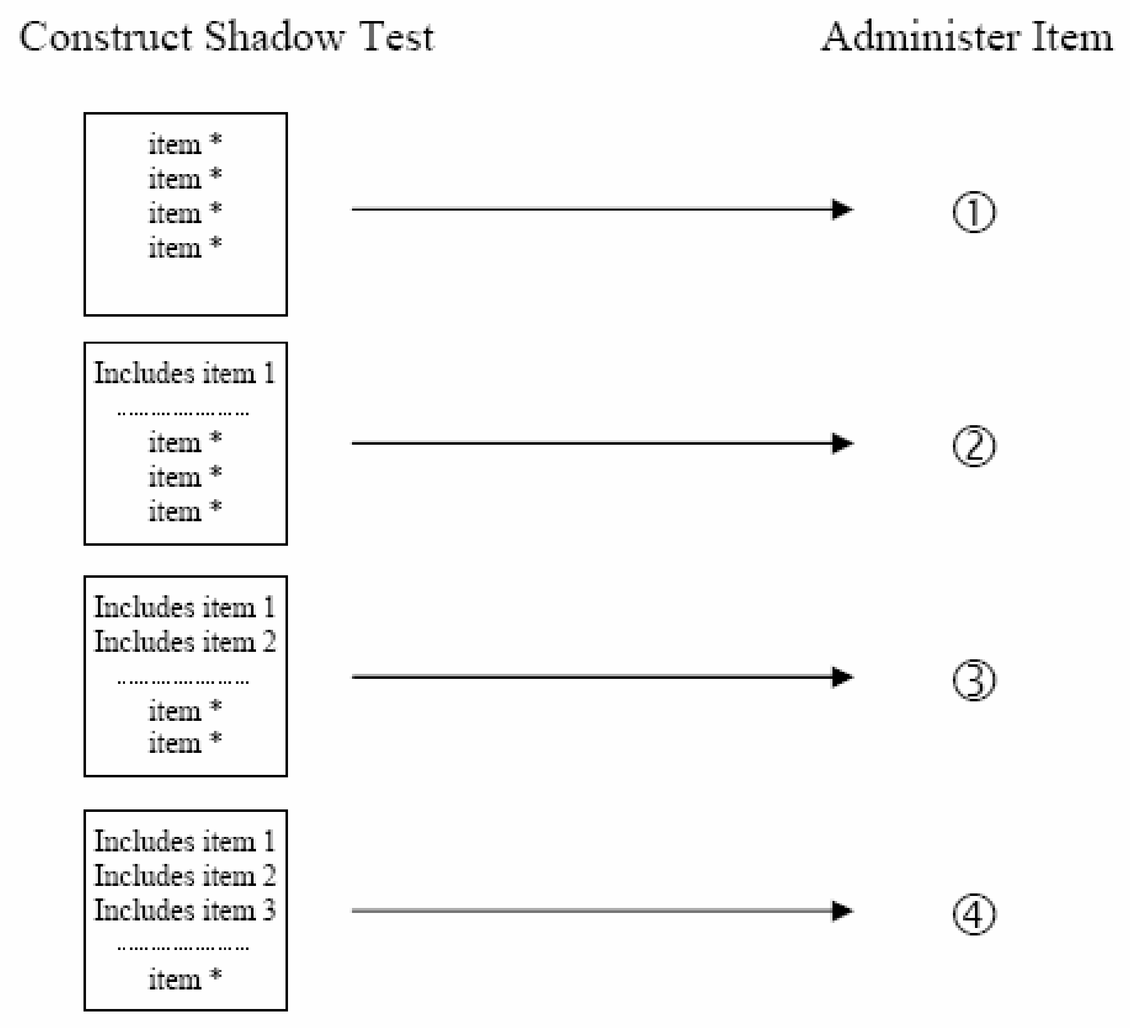
Select items based on $\underline{\alpha}$ estimates only

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

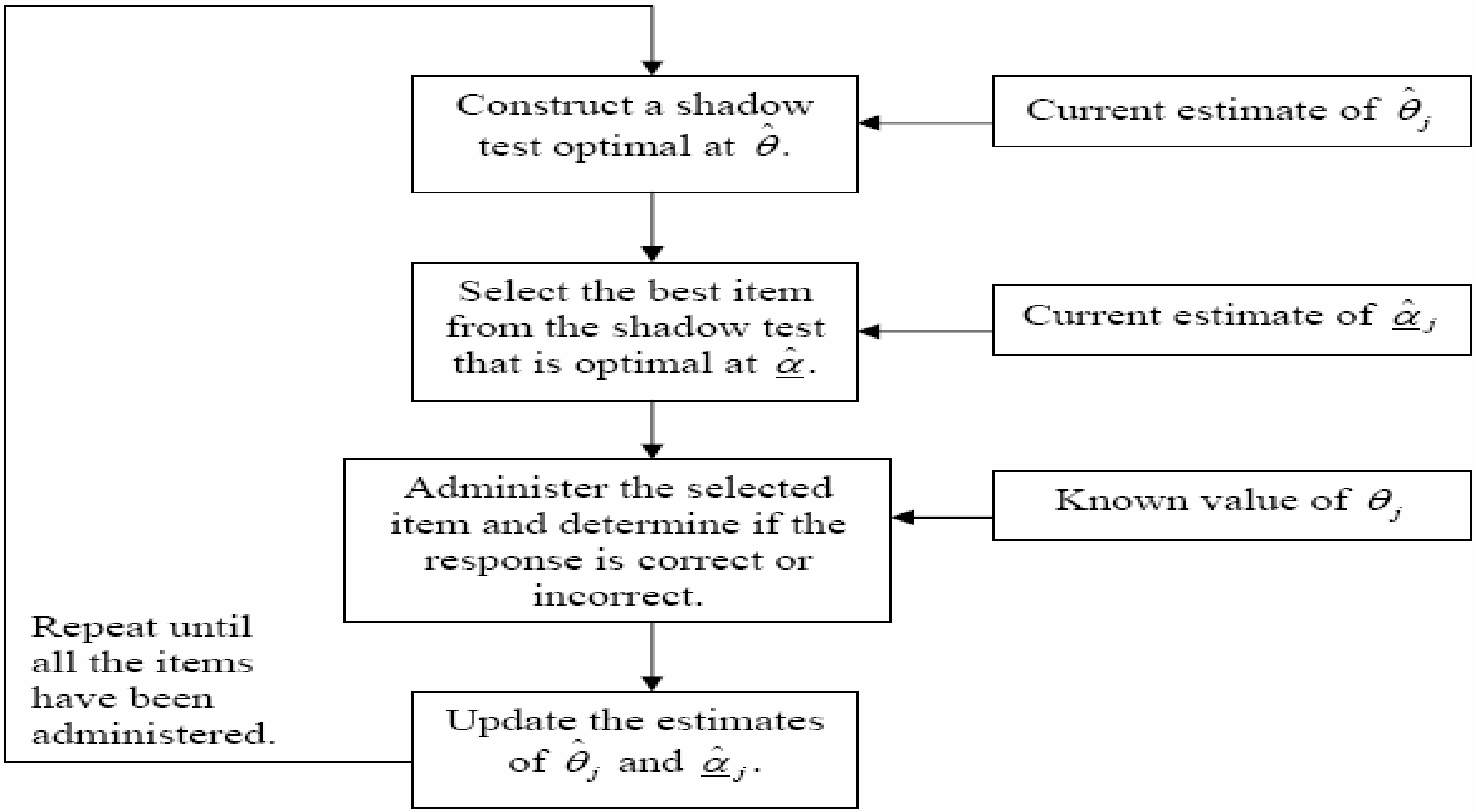
$$K_i\left(\underline{\hat{\alpha}}\right) = \sum_{c=1}^{2^M}\left\{\sum_{x=0}^{1}\log\left(\frac{P(X_i = x \mid \underline{\hat{\alpha}})}{P(X_i = x \mid \underline{\alpha}_c)}\right)P\left(X_i = x \mid \underline{\hat{\alpha}}\right)\right\}$$

Select $\underline{\hat{\alpha}}$ such that K is maximum

Both θ and $\underline{\alpha}$ estimates simultaneously (Wills, Chang,& Wills)



Both θ and $\underline{\alpha}$ estimates simultaneously

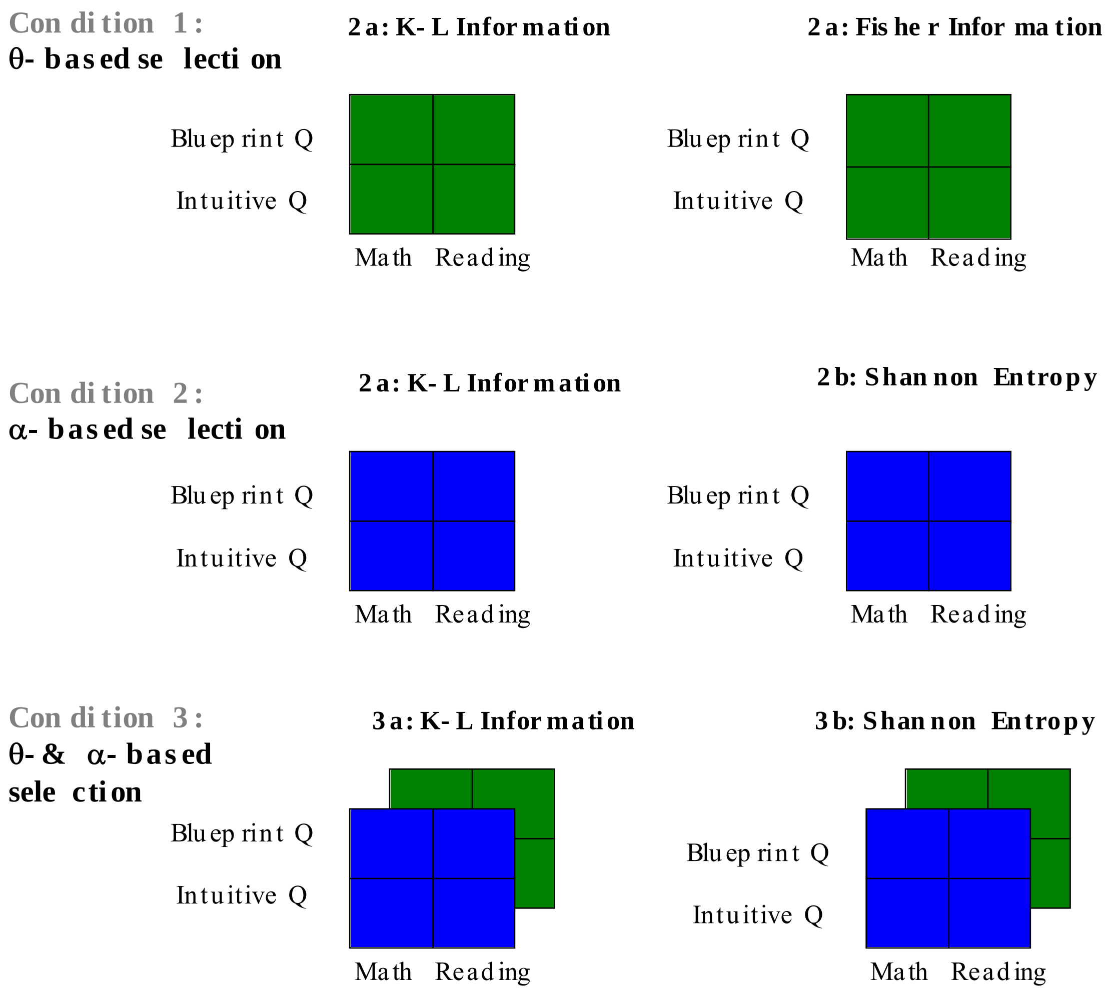


Micro-level Research Design

Test Subject

Q-Matrix		Math	Reading
	Blueprint		
	Intuitive		

Macro-level Research Design



Evaluation is based on:

- Theta Accuracy
- Attribute Mastery Pattern Accuracy
- Item Exposure

Theta Accuracy

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

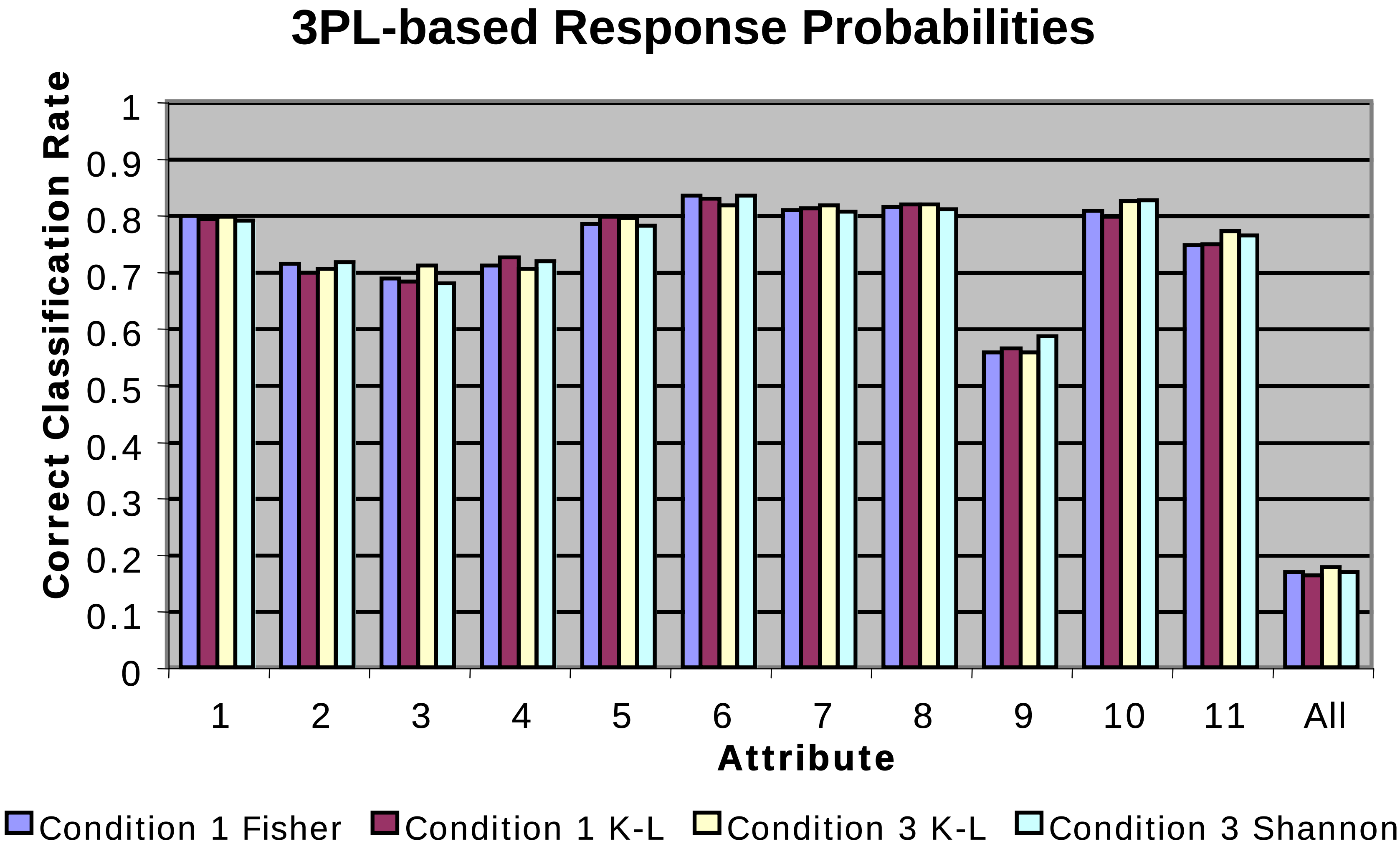
	Condition 1		Condition 3	
	Fisher	K-L	Shannon	K-L
Math				
Blueprint Q-matrix	0.955	0.957	0.940	0.946
Intuitive Q-matrix	0.959	0.943	0.942	0.937
Reading				
Blueprint Q-matrix	0.922	0.933	0.941	0.912
Intuitive Q-matrix	0.916	0.916	0.926	0.934

Attribute Mastery Estimates
with the math Blueprint Q-Matrix

Table 3:
*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

Attribute Mastery Estimates
with the math Blueprint Q-Matrix



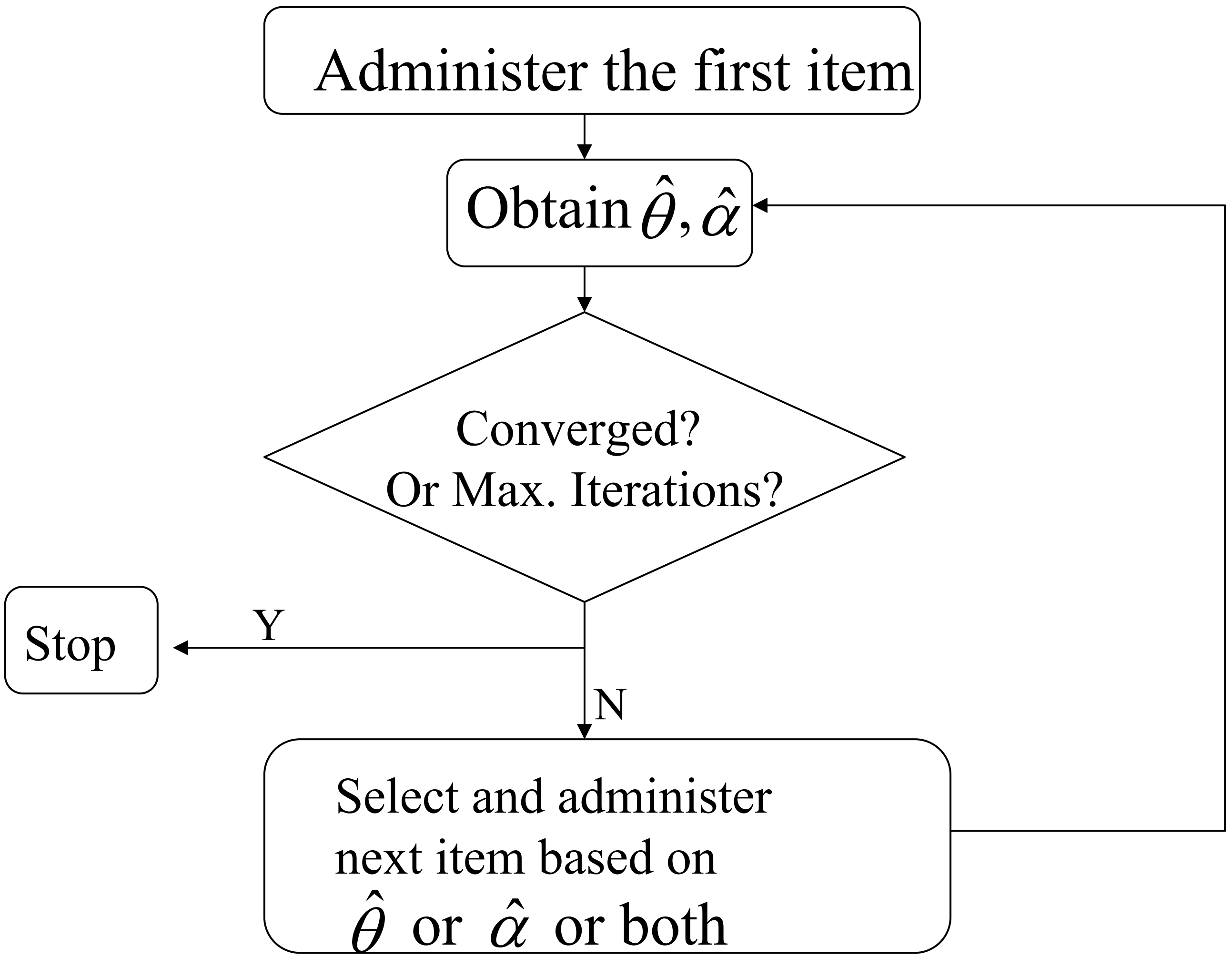
Overall Results

Item selection Method	Performance
based on θ	good overall
based on $\underline{\alpha}$	poorer than the other 2
based on θ and $\underline{\alpha}$	good overall

The Dual Information Method for Item Selection in Cognitive Diagnostic CAT

Cheng & Chang

Cognitive Diagnostic CAT (CD-CAT)



Current CD-CAT Methods

- Adaptive based on θ
- Adaptive based on α (Xu et. al., 2003)
- Adaptive based on both θ and α (McGlohen, 2004)
  - Step 1: Select a set of items that are optimal on θ
  - Step 2: Select from the above set the item that is optimal on α

This Study

- Proposes a new item selection method on the basis of synchronized information on both θ and α

- Is it adequate to adapt solely on the basis of either θ and α ?

  The two pieces of information "are substantially (but far from perfectly) correlated at the item and at the test-taker level." (Budescu, Karelitz & Douglas, 2002)

The Dual Information Method (I)

- Kullback-Leibler (KL) Information on θ (Chang & Ying, 1996)
  - The KL info. of the *j*th item to distinguish θ from $\hat{\theta}_m$

$$KL_j(\theta \| \hat{\theta}_m) = P_j(\hat{\theta}_m)\log[\frac{P_j(\hat{\theta}_m)}{P_j(\theta)}] + [1 - P_j(\hat{\theta}_m)]\log[\frac{1 - P_j(\hat{\theta}_m)}{1 - P_j(\theta)}]$$

  - Global information for item selection

$$KL_j(\hat{\theta}_m) = \int_{\hat{\theta}_m - \delta_m}^{\hat{\theta}_m + \delta_m} KL_j(\theta \| \hat{\theta}_m) d_\theta$$

The Dual Information Method (II)

- Kullback-Leibler (KL) Information on α (Xu et. al., 2003)

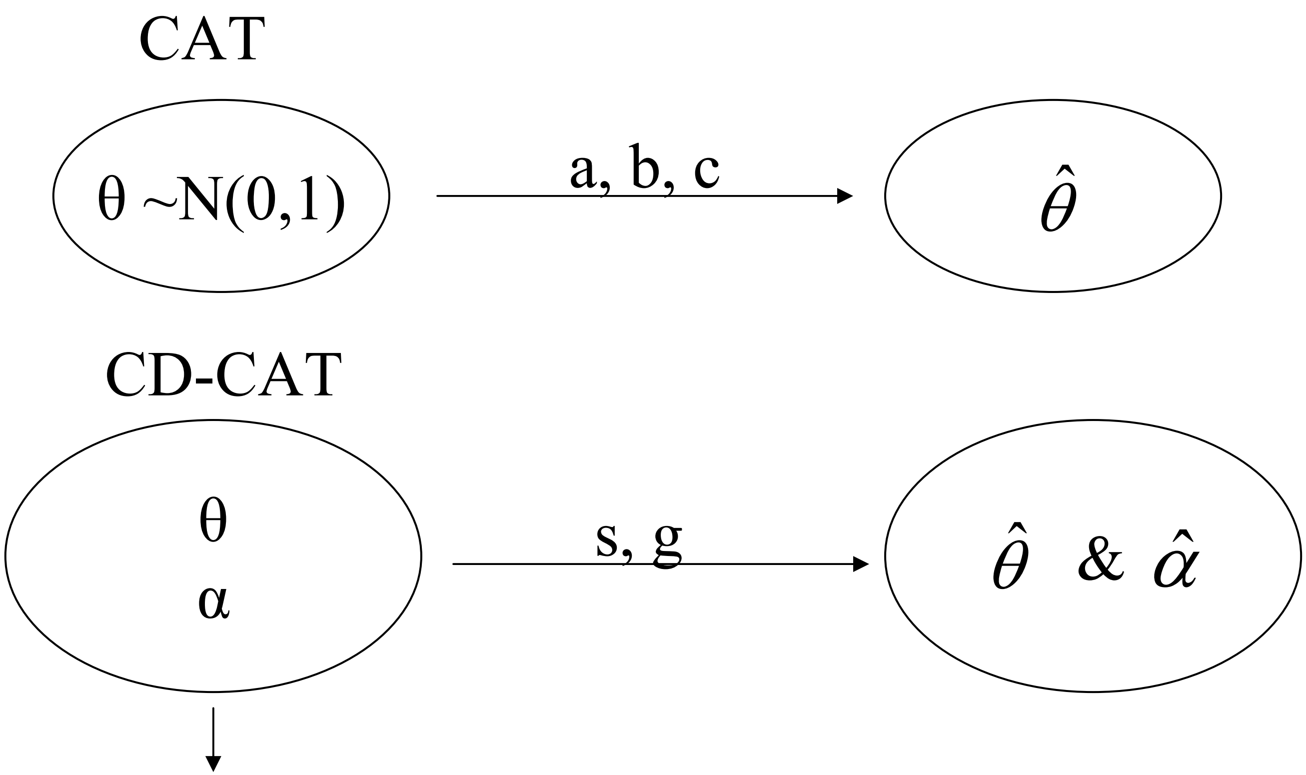
$$KL_j(\hat{\alpha}_m) = \sum_{c=1}^{2^K} [\sum_{x=0}^{1} \log(\frac{P(X_j = x \mid \hat{\alpha}_m)}{P(X_j = x \mid \alpha_c)}) P(X_j = x \mid \hat{\alpha}_m)]$$

- Dual information: a weighted sum

$$KL_j(\hat{\theta}_m, \hat{\alpha}_m) = w \cdot KL_j(\hat{\alpha}_m) + (1 - w) \cdot KL_j(\hat{\theta}_m)$$

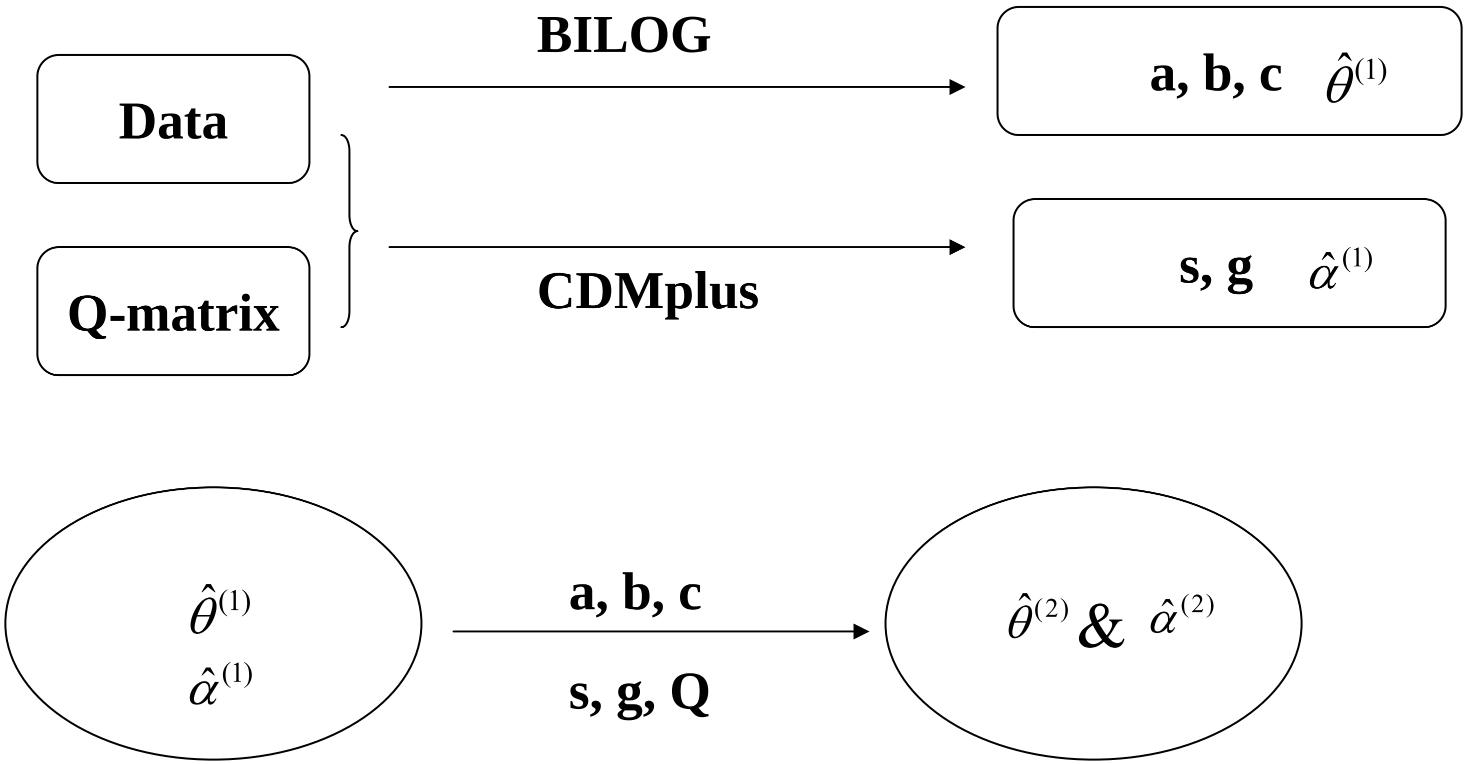
- Item selection: maximum dual information

Simulation Design (I)



Note here the θ and α can’t be generated separately.

Simulation Design (II)



Details of Data and Simulation (I)

- Data: 2,000 x 36 response matrix (TAKS data)
- Q-matrix adopted from McGlohen's (2004) dissertation
- Calibrate 36 sets of *a*, *b*, *c*, *s*, *g* and 2,000 $\hat{\theta}^{(1)}$ and $\hat{\alpha}^{(1)}$
- The item pool tripled: 108 items
- Item length: 36
- First item is randomly selected from the bank
- Theta estimation: EAP

  Alpha estimation: MLE

Details of Data and Simulation (II)

- 11 levels of weights: w = 0, 0.1, 0.2, … 1.0

  w = 0 : Solely rely on KL(θ)

  w = 1.0: Solely rely on KL(α)

Evaluation Criteria

- Measurement precision:
  - Bias = $\sum_{j=1}^{N}(\hat{\theta}_j^{(2)} - \hat{\theta}_j^{(1)}) / N$
  - MSE = $\sum_{j=1}^{N}(\hat{\theta}_j^{(2)} - \hat{\theta}_j^{(1)})^2 / N$
  - Correlation: $\rho_{(\hat{\theta}^{(2)}, \hat{\theta}^{(1)})}$
- Examinee profile recovery rate computed for each attribute and the entire profile pattern

  [0 0 1 1 1] → [0 0 0 1 1]

Results: Measurement Precision

Weight	Bias	MSE	$\rho_{\theta,\hat{\theta}}$
0	0.110	0.468	0.714
0.1	0.091	0.474	0.708
0.2	0.094	0.480	0.705
0.3	0.049	0.467	0.703
0.4	0.077	0.465	0.710
0.5	0.066	0.480	0.700
0.6	0.064	0.463	0.710
0.7	0.091	0.483	0.702
0.8	0.087	0.478	0.704
0.9	0.071	0.470	0.705
1.0	0.075	0.475	0.705
Average	0.079	0.473	0.706

Results: Profile Recovery

Weight	Attr. 1	Attr. 2	Attr. 3	Attr. 4	Attr. 5	Attr. 6	Whole pattern
0	0.892	0.930	0.883	0.868	0.934	0.900	0.548
0.1	0.906	0.961	0.862	0.921	0.946	0.935	0.686
0.2	0.917	0.955	0.835	0.941	0.948	0.920	0.670
0.3	0.908	0.971	0.834	0.941	0.923	0.921	0.658
0.4	0.912	0.962	0.860	0.928	0.931	0.920	0.688
0.5	0.925	0.947	0.837	0.934	0.934	0.905	0.659
0.6	0.928	0.961	0.782	0.932	0.936	0.924	0.615
0.7	0.912	0.956	0.795	0.937	0.938	0.924	0.627
0.8	0.909	0.956	0.854	0.931	0.934	0.926	0.686
0.9	0.910	0.959	0.797	0.936	0.936	0.912	0.623
1.0	0.917	0.952	0.843	0.932	0.945	0.926	0.677
Average	0.912	0.955	0.835	0.927	0.937	0.919	0.649

Conclusion (I)

- The dual information method achieves comparable measurement precision and profile recovery as Meghan's (2004) method

- By manipulating the weight, it is clear that an item selection method solely rely on θ is not adequate in providing diagnostic feedback

Conclusion (II)

- The relationship is not monotone: increasing the weight on KL(α) would not necessarily lead to higher recovery rate.

- Cognitive diagnosis would benefit from the additional piece of information KL(θ) provides.