

# Computerized, Internet-Based, or Paper-Pencil Test, In Which Direction Should Testing Go?

--- A Close Look at Graduate Record Examination

Hua-Hua Chang  
University of Illinois

The 25<sup>th</sup> IRT Workshop  
October 2009  
Enschede, The Netherlands

## CAT: One of the Most Important Applications of IRT

- Computerized Testing
  - Adaptive Testing (CAT)
  - Linear Testing (CLT)
- Internet Based Testing (IBT)
  - ETS Definition
  - General Definition
- Paper & Pencil Testing
- Which mode is promising?

## Computerized Testing

- Computerized Adaptive Testing (CAT)
- Computerized Linear Testing (CLT)
  - Single Stage (non adaptive)
  - Multiple Stage (adaptive)
  - CAT or CLT?
- Continuous Testing
- Test Security
- Issues in Item Selection
- How to assemble many parallel forms?



3

## Historical Review of CAT

- Lord (1970), called *tailored* tests
  - Examinee will be measured most effectively if items are neither too difficult nor too easy.
- But he did not have a chance to oversee the development due to his injury.
- ETS rushed into implementation of Computerized GRE in early 90's
  - paper/pencil tests were eliminated.

4

## Educational Testing Service



## Frederic Lord (1912-2000)



In May 1982, when a major conference was held at ETS commemorating Lord's 70th birthday, participants came from as far away as Australia and Europe. The conference resulted in the publication of a major book in his honor, entitled *Principals of Modern Psychological Measurement*, edited by Howard Wainer and Samuel Messick. The book brings together new scholarly contributions from around the world that were stimulated wholly, or in part, by the work of Lord.

## Historical Review (Conti.)

- 2002 ETS suspended CAT-GRE and reintroduced P&P versions in China, Hong Kong, Taiwan, and Korea.
- 2005 ETS announced a plan to replace CAT-GRE by IBT-GRE in October 2006.
- 2007 Delayed Release of New GRE Test
- July 2007 ETS announced the cancellation of the IBT-GRE plan.
- ETS is going to replace CAT-GRE with a multi-stage CLT. Let's just wait....

7

## Successful CAT Applications

- CAT has become a popular mode of educational assessment
  - Examples: GMAT, ASVAB, and etc.
- More & more tests are becoming CAT-based
- Patient Reported Outcome Application.
- K-12 Cognitive Diagnostic CAT in China
- Advantages:
  - Efficiency, new item types, continuous testing

8

## Item Selection Criteria in CAT

The most important component in CAT is the item selection procedure

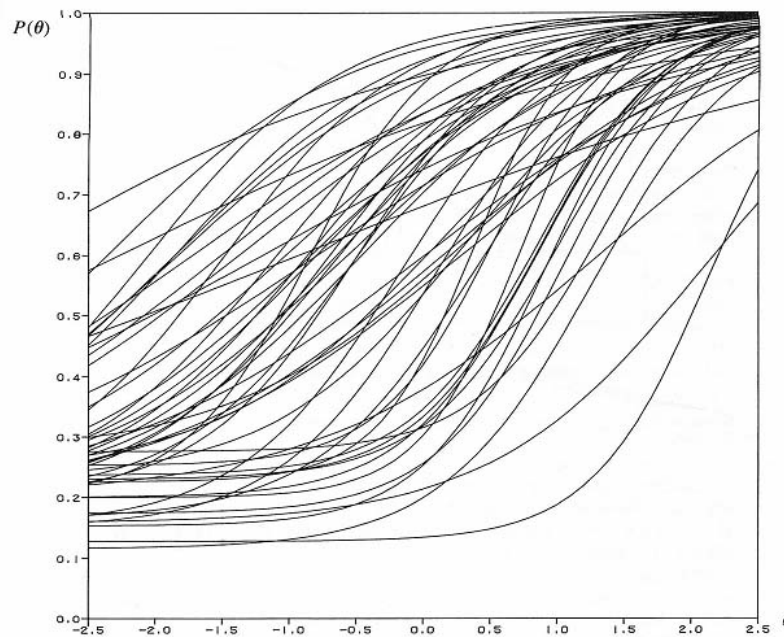
$\theta$ : latent trait

- How to sequentially estimate  $\theta$ ?
- Heuristically,
  - if the answer is correct, the next item should be more difficult;
  - If the answer is incorrect, the next item should be easier.

$$P\{\theta = b\} = 1/2$$

where  $b$  is item difficulty level

9



Item response functions for SCAT II Verbal Test, Form 2B.

10

## The Maximum Information Criterion (MIC)

- Lord's (1980) MIC method, the most commonly used method.

$\theta_0$  : true latent trait

$\hat{\theta}_n$  : MLE after  $n$  items were administered

$R_n$  : Item Pool after  $n$  items selected

$I_l(\theta)$  : Item information function

$$j_{n+1} = \max_l \left\{ I_l(\hat{\theta}_n) : l \in R_n \right\}$$

11

## Theoretical Foundation of MIC

Let  $X_1, X_2, \dots, X_n$  be item response, then

$$\hat{\theta}_n \rightarrow \theta_0 \text{ as } n \rightarrow \infty \text{ with } \text{var}(\hat{\theta}_n) \rightarrow \frac{1}{I(\theta_0)}$$

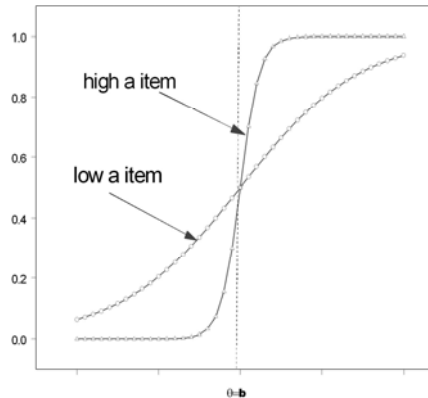
$$\text{where } I(\theta) = \sum_{l=1}^n I_l(\theta)$$

Thus, the closeness of  $\hat{\theta}_n$  to  $\theta_0$  is governed by  $I(\theta)$ .

12

## Are Low- $a$ -Items Bad Items?

- If we only use high- $a$  items then, what's the reason to include other items in the pool?
- High- $a$ -items are good items if we know true ability, otherwise may not be good.
- Low- $a$ -items can be more informative if we don't know the true ability.
  - Therefore, at beginning of the test, they can be more effectively used.



13

## Issues in Large Scale Applications

- From 2000 to 2003, GRE and GMAT did not produce reliable scores for several thousand test-takers.
  - ETS offered them a chance to retake the test at no charge (Carlson, 2000).
- Test Security Problems
  - Life items found on some websites
- In 2002, ETS suspended CAT-GRE in several countries
- In 2006 ETS announced IBT GRE

14

## Alternatives?

- **If the CAT is bad,**
  - what should we do? Kill him?
  - Should we take a U-Turn?
  - Or, Computer-delivered P&P tests?
- **Shall we go back to P&P?**

15

## Is P&P Much Safer Than CAT?

New challenges: Test Security In China

Examination booklets are shipped under protection of armed police force





## Security Specialists Detect Bluetooth Communication at a Test Site of National Center of Medical Examination



Device used for cheating →



Despite government's tremendous effort, the number of test security violation is on the rise.



## Will Internet-Based Testing (IBT) Work?



## Will IBT Work?

<b>CAT-GRE</b>	<b>IBT-GRE</b>
Adaptive	Linear
~ 600 test centers worldwide	> 2,500 test centers worldwide
Continuous	30 fixed administrations
Question pools in continuous use	New test form on every test date

## Test Administration Changes

<b>CAT-GRE</b>	<b>IBT-GRE</b>
Dedicated computer-based testing centers	Existing language labs, computer labs, etc., reserved for ETS testing purposes
Local start times	Staggered start times

Nowadays it may not be easy to find many terminals at universities.  
In July 2007 ETS announced the cancellation of the IBT-GRE plan.

23

## A Potential New Mode for GRE

- Computer-delivered linear (CLT) tests
- Unique advantage of CAT is the adaptive feature
  - Get information as a person taking the test
  - Without “adaptivity”, you apparently don’t make a full use the information.
- Will CLT work?
- For a given set of items will CLT outperform CAT?

24

## How Can We Overcome?

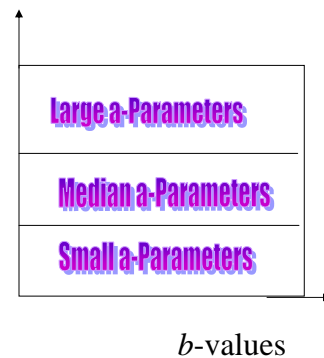
- How to Make CAT More Reliable?
- How to improve Test security without Sacrificing Estimation Efficiency?
- How to Use CAT in NCLB Accountability Testing?
- How to Extend CAT Application to Other Fields?
  - Psychological testing?
  - medical assessment, e.g., to measure Quality of Life, patient reported outcome.

25

## Controlling Exposure by Fixing the Selection Algorithm

Example:  $a$ -Stratified Design (Chang & Ying 1999)

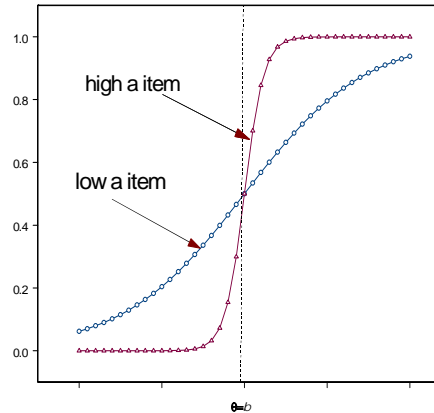
1. Partition item bank into  $K$  strata
2. Divide the test into  $K$  stages
3. In the  $k$ -th stage, select  $n_k$  items from  $k$ -th stratum
  - Within each stratum, the next item is selected so that its  $b$ -parameter is closest to the estimated trait
4. Repeat Step 3 from  $k=1, \dots, K$ .



26

## Davey & Nering (2002):

- Highly discriminating items are like a tightly focused spotlight that shines intensely but casts little light outside a narrow bean.
- Less discriminating items are more like floodlights that illuminate a wide area but not too brightly.
- The idea is to use the floodlights early on to search out and roughly locate the examinee, then switch to spotlights to inspect things more closely.



27

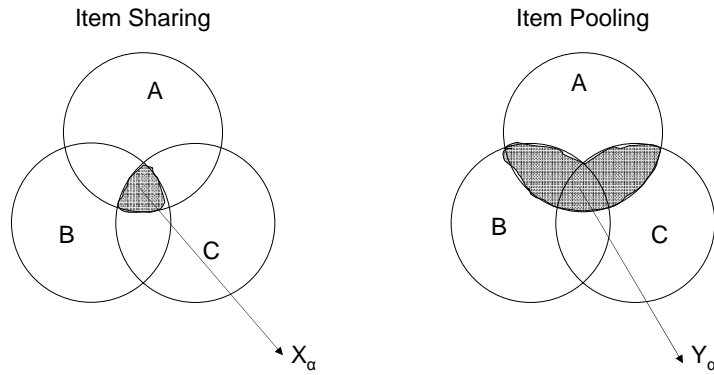
## In Order to Make a Secured CAT How Large the Item Pool Should Be?

- Assessment of Organized Item Thievery
  - Kaplan-ETS incident in 1994
- Research Question:
  - To compromise an item bank, how many thieves are needed?

28

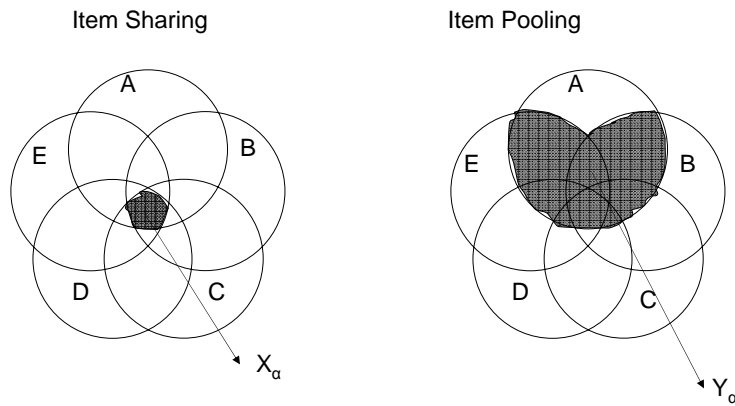
# Item Sharing vs. Item Pooling

Chang & Zhang, 2001



29

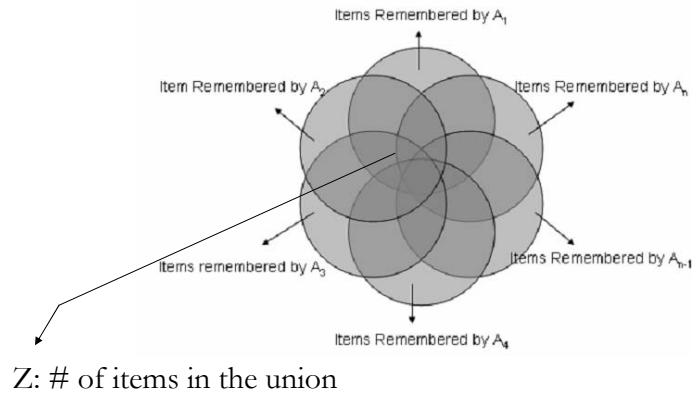
# Item Sharing vs. Item Pooling



30

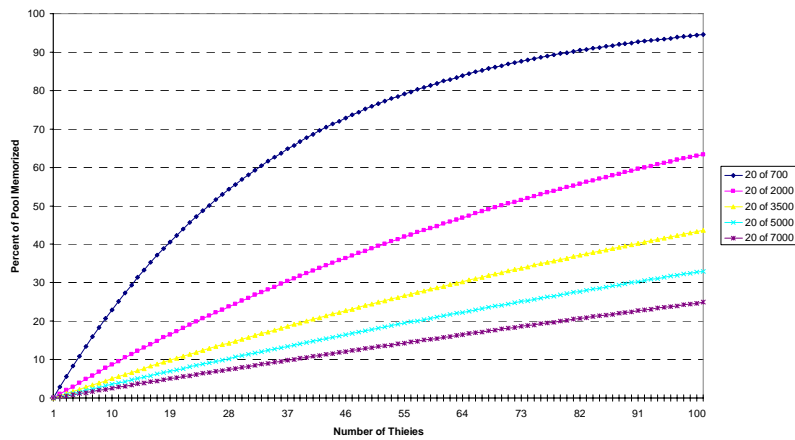
# How Many Thieves Are Needed?

## Items Can Be Pooled By Kaplan



31

Pool Size, Number of Items Memorized, and Number of Thieves



If pool size:7000 and each person can remember 20 items, then 100 thieves can pool 1750 items.

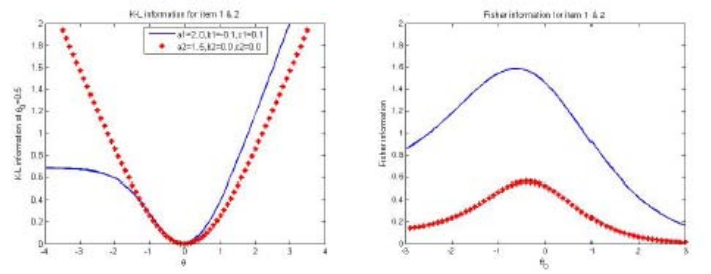
Memorizing 1750 items is very difficult!

However, this won't help very much, because there are still 5250 secured items in the pool!

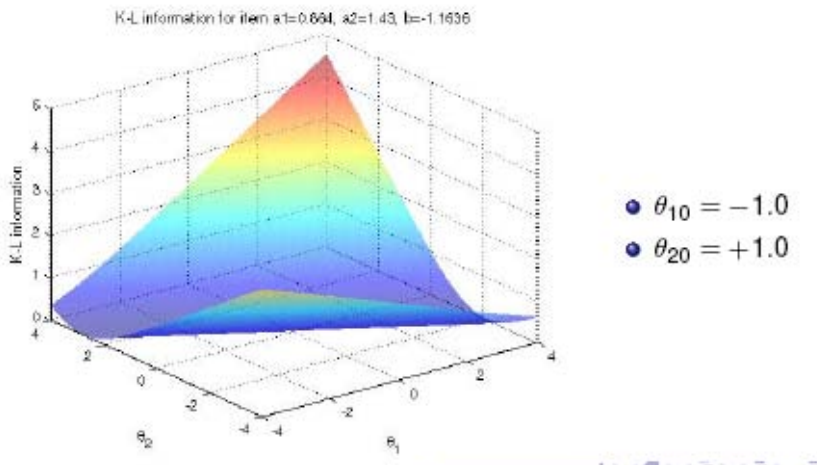
32

# Extension of Chang's Keynote Speech at the 11<sup>th</sup> IRT Workshop November 1995

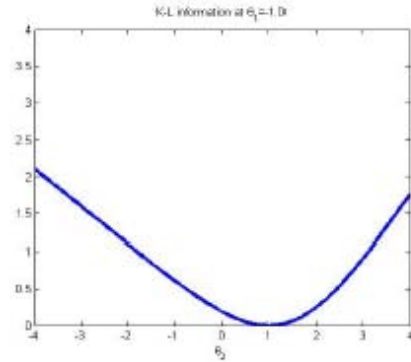
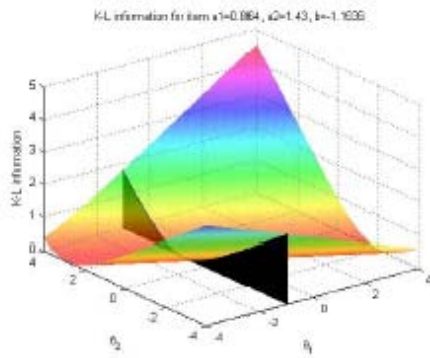
Introduction and Objectives Theory and Analytical derivations	Information Fisher Information & KL Information Application in Adaptive Tests Conclusion
<b>Illustration of the <i>KL</i> and <i>FI</i></b>	



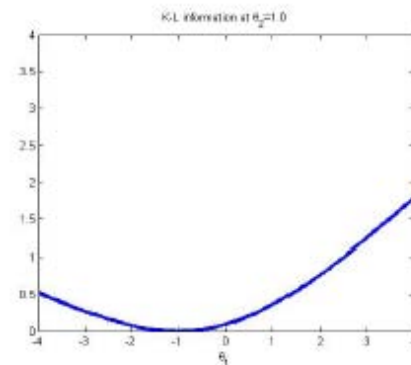
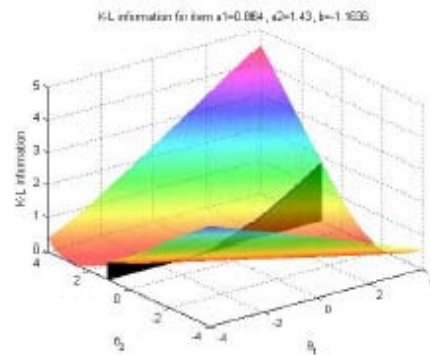
## Illustration of the KL information surface



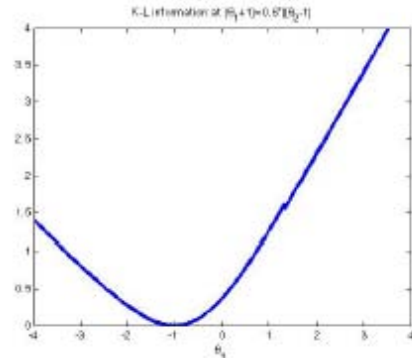
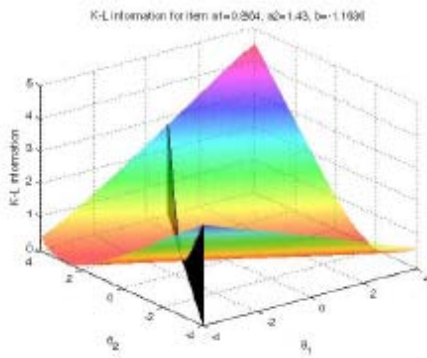
## Two Dimension case: KL vs. FI



## Two Dimension case: KL vs. FI



## Two Dimension case: KL vs. FI



## Fisher Information & KL Information

- Fisher Information Matrix

$$I_i(\theta) = g(\theta; \mathbf{a}_i, b_i, c_i) \begin{bmatrix} a_{i1}^2 & a_{i1}a_{i2} & \dots & a_{i1}a_{ip} \\ a_{i1}a_{i2} & a_{i2}^2 & \dots & a_{i2}a_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1}a_{ip} & a_{i2}a_{ip} & \dots & a_{ip}^2 \end{bmatrix} \quad (6)$$

$$g(\theta; \mathbf{a}_i, b_i, c_i) = \frac{Q_i(\theta)[P_i(\theta) - c_i]^2}{P_i(\theta)(1 - c_i)^2} \quad (7)$$

## Fisher Information & KL Information

- Fisher Information Matrix

$$I_i(\theta) = g(\theta; \mathbf{a}_i, b_i, c_i) \begin{bmatrix} a_{i1}^2 & a_{i1}a_{i2} & \dots & a_{i1}a_{ip} \\ a_{i1}a_{i2} & a_{i2}^2 & \dots & a_{i2}a_{ip} \\ \dots & \dots & \dots & \dots \\ a_{i1}a_{ip} & a_{i2}a_{ip} & \dots & a_{ip}^2 \end{bmatrix}$$

$$g(\theta; \mathbf{a}_i, b_i, c_i) = \frac{Q_i(\theta)[P_i(\theta) - c_i]^2}{P_i(\theta)(1 - c_i)^2}$$

- KL Information: The same as in UIRT

$$K(\theta || \theta_0) = E \left[ \ln \frac{L(\theta_0; \mathbf{u})}{L(\theta; \mathbf{u})} \right]$$

## Relationship of KL and FI in higher-dimension

### Theorem

Let  $\theta_0$  be the true ability vector of the examinee, and  $I(\theta_0)$  be the Fisher item information matrix evaluated at  $\theta_0$ . For any given  $\theta$ , let  $K(\theta_0 || \theta)$  be the KL item information. Then each entry of  $I(\theta_0)$  can be obtained by taking second derivatives of  $K(\theta_0 || \theta)$ .

### Proof.

$$\frac{\partial^2 K(\theta || \theta_0)}{\partial \theta_r^2} \Big|_{\theta = \theta_0} = I_{rr}(\theta_0) \quad (9)$$

$$\frac{\partial^2 K(\theta || \theta_0)}{\partial \theta_r \partial \theta_s} \Big|_{\theta = \theta_0} = I_{rs}(\theta_0) \quad (10)$$

□ ▷ ◁ ◂ ◃

## KL Information in MAT

KL Information Index in higher dimension (Veldkamp & van der Linden, 2002)

$$KI(\hat{\theta}_n) = \int_{\theta \in \mathcal{D}} K(\theta \| \theta_0) d\theta \quad (14)$$

We only consider two dimensions here, The KL Information Index (KI) reduces to

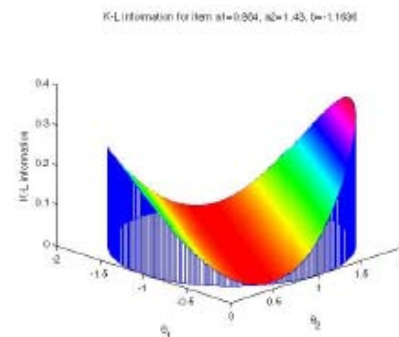
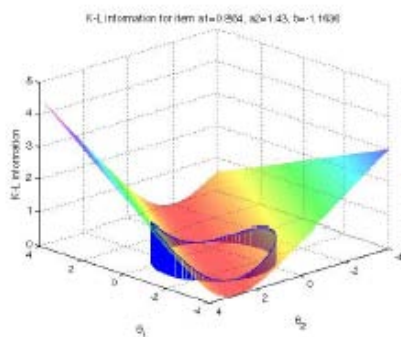
$$KI(\hat{\theta}_n) = \iint_{\mathcal{D}} K(\theta \| \theta_0) d\theta_1 d\theta_2 \quad (15)$$

The *Integration domain*

- 1 Square:  $\mathcal{D} = [\theta_{10} - \delta_1, \theta_{10} + \delta_1] \times [\theta_{20} - \delta_1, \theta_{20} + \delta_1]$
- 2 Circular:  $\mathcal{D} = \{(\theta_1, \theta_2) | \theta_1^2 + \theta_2^2 \leq r^2\}$ ;
- 3 Rectangular:  $\mathcal{D} = [\theta_{10} - \delta_1, \theta_{10} + \delta_2] \times [\theta_{20} - \delta_1, \theta_{20} + \delta_2]$
- 4 Elliptic:  $\mathcal{D} = \{(\theta_1, \theta_2) | \frac{\theta_1^2}{r_1^2} + \frac{\theta_2^2}{r_2^2} \leq 1\}$ ;

Navigation icons: back, forward, search, etc.

## Two-dimension case: KL Information Index (KI)



Navigation icons: back, forward, search, etc.

## Relationship of KI and Item Discriminations

### Theorem

Let  $\theta_0$  be the true ability of examinee and  $\mathbf{a}$  be the item discriminations. Define KL information Index as  $KI(\hat{\theta}_n) = \int_{\theta \in \mathcal{D}} K(\theta || \theta_0) \partial \theta$ , where  $\mathcal{D}$  is a central symmetric domain centered around  $\theta_0$ . For two-dimension case,  $KI(\hat{\theta}_n) \propto f(\mathbf{a})$  as  $r \rightarrow 0$ . In particular,  $f(\mathbf{a}) = a_1^2 + a_2^2$  when  $\mathcal{D}$  is a circle; and  $f(\mathbf{a}) = [(a_1 r_1)^2 + (a_2 r_2)^2]$  when  $\mathcal{D}$  is an ellipse.

- Euclidean norm of  $\mathbf{a}_i$ ,  $\sqrt{a_1^2 + a_2^2}$ , is Multidimensional Discrimination (MDISC) (Reckase & McKinley 1991)
- $MDISC = (\sum_{k=1}^p a_{ik}^2)^{1/2}$

## Cognitive Diagnostic CAT

### What is reported to examinees?

Traditional Testing:

Cognitive  
Diagnosis:

$$\theta$$

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

A single score

A set of scores:  
One for each attribute.

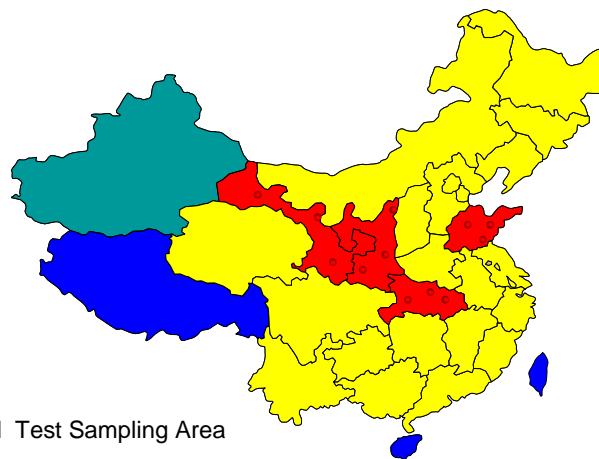
( $K$  is the total # of attributes.)

## CD-CAT Project in China

- 4-th grade and 8-th grade math and English
- CD-CAT items were written after item writers received psychometric training
- 4 item banks for each grade and each subject
- A field test was conducted for 120,000 students
- Several million students will use the system
- Provide feedback to teachers

45

## Distribution of the students in pretest



Red: Field Test Sampling Area

Yellow and red: Current Implementation

46

# Linking Design

Eg, this block has 10 anchor items,

	Anchor items			
	Group1	Group2	Group3	Group4
Test1				
Test2				
Test3				
Test4				
Test5				
Test6				
Test7				
Test8				
Test9				
Test10				
Test11				
Test12				
Anchor Test				

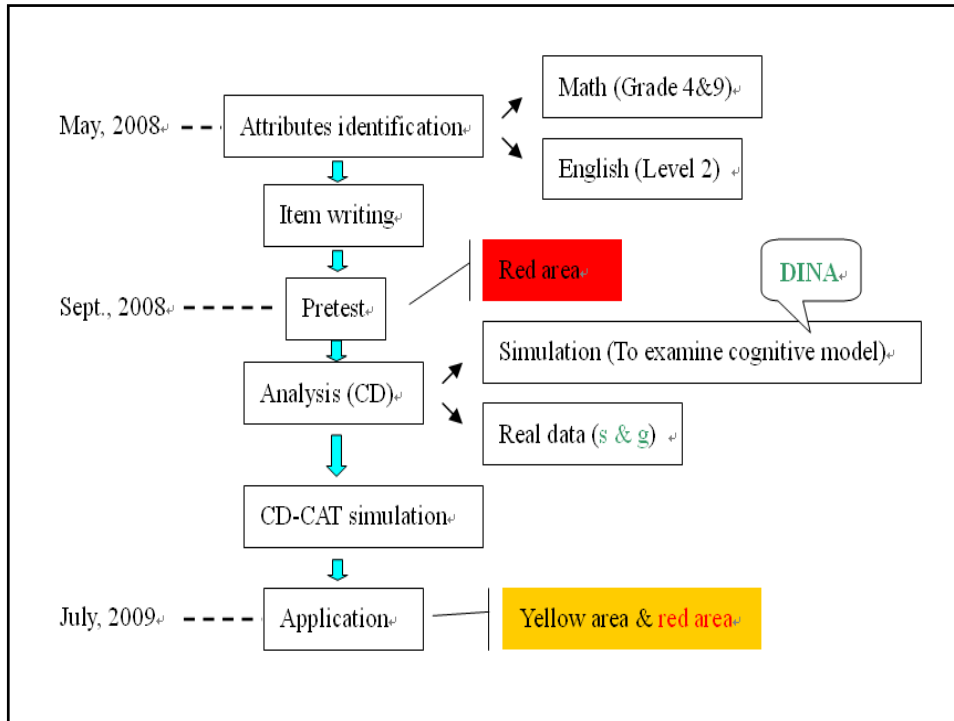
The locations of the anchor items in each booklet are the same (as they appear in anchor test).

47

# CD: Simulation Study

- Model: **DINA (EM) vs. HO-DINA(MCMC)**;
- Number of Attributes: **5, 6, 7, 8**;

48



## Conclusion

Despite its limitations, CAT undoubtedly has a great future because cutting-edge developments in technology will enable us to solve the problems encountered in current large-scale applications.

—Hua-Hua Chang and Zhiliang Ying