Building Affordable CD-CAT Systems for Schools To Address Today's Challenges In Assessment

Hua-Hua Chang

University of Illinois at Urbana-Champaign

Challenges in NCLB Testing

- Many items are too difficult to students
  - 70% math items may be too difficult
    - The influence of this kind of test taking experience on low-achieving students is not well-understood (e.g., Roderick & Engle, 2001, Ryan & Ryan, 2005; Ryan, Ryan, Arbuthnot, & Samuels, 2007).
- Test security of NCLB
    - **The # of security violations in P&P based NCLB testing in on the rise.**
    - Documented cases of such incidents have been uncovered in numerous states including New York, Texas, California, Illinois, and Massachusetts. (Jacob & Levitt, 2003, and Texas Education Agency, 2007).

What is Adaptive Testing?

- Originally called *tailored* tests (Lord, 1970)
  - Examinee are measured most effectively if items are neither too difficult nor too easy.
- Θ: latent trait. Heuristically,
  - if the answer is correct, the next item should be more difficult;
  - If the answer is incorrect, the next item should be easier.
- How adaptive test works?
  - An item pool, known item properties
  - Algorithm, computer, and network
  - The core is the item selection algorithm
  - Constraint Control

CAT Has Glowing Future in K-12 Context

- Why not use benchmark testing?
  - Adaptive Testing can do better.
- Quellmalz & Pellegrino (2009):
  - more than 27 states currently have operational or pilot versions of online tests, including Oregon, North Carolina, Utah, Idaho, Kansas, Wyoming, and Maryland.
  - The landscape of educational assessment is changing rapidly with the growth of computer-administered tests.

From Theory to Large-Scale Operation

- Issues to be addressed:
- Should CAT only use the best items?
- Is CAT more secure than paper/pencil test?
- How to control non-statistical constraints?
- How to get diagnostic information?
- How to make CAT affordable to many schools?

Objectives:

1. MAKING CAT DIAGNOSTIC TOOL
2. DELIVER THE TOOL TO SCHOOLS

How to get diagnostic information?

- Post-hoc approach (non-adaptive)
  - perform CD after students completed CAT
- Adaptive approach
  - Select the next item which provides the max info about the student’s strength and weakness
  - Need a model, item selection algorithm
  - Psychometric theory
  - Simulation study
  - Field test

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.

What should be reported to examinees?

Traditional Testing:

$$\theta$$

A single score

Cognitive Diagnosis:

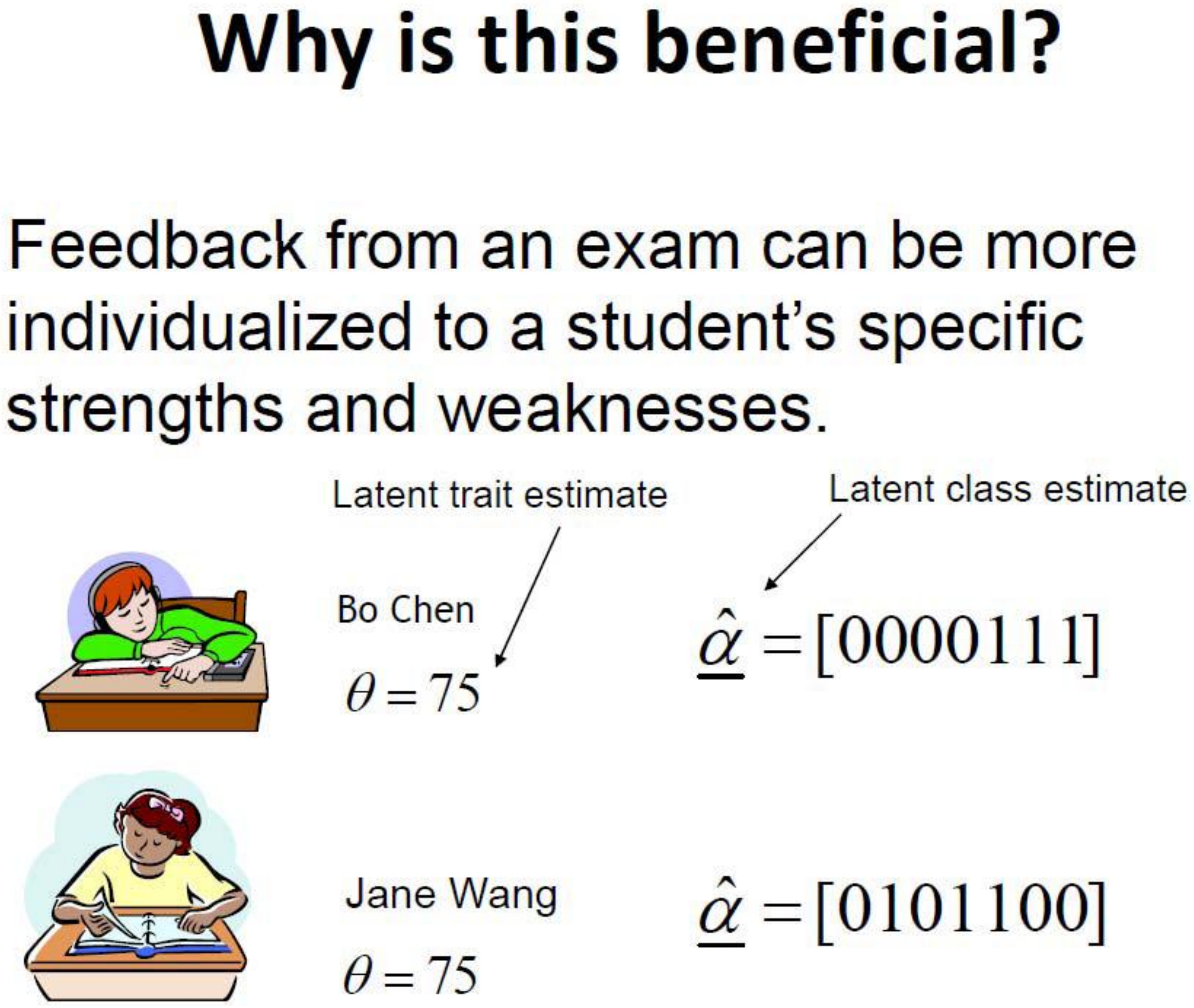
$$\underline{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_K]$$

A set of scores:

One for each attribute.

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

Why is this beneficial?



Why is this beneficial?

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

Latent trait estimate

Latent class estimate

Bo Chen

$\theta = 75$

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

Jane Wang

$\theta = 75$

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

The Item-Attribute Relationship

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

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

Cognitive Diagnostic Models

$$P(X_{\underset{\text{person}}{i}\underset{\text{Item}}{j}}=1|\overset{\text{vector}}{\underline{\alpha}}_i)$$

- Many models have been proposed
- DINA model (Macready &Dayton, 1977; Junker & Sijsma, 2001)
- Fusion model (Stout’s group)

The DINA Model

Student *i*
Item *j*

Deterministic Input; Noisy "And" Gate

(Macready & Dayton, 1977, 1989; Junker & Sijstma, 2001)

$$P(X_{ij} = 1 | \xi_{ij}) = (1 - s_j)^{\xi_{ij}} g_j^{(1-\xi_{ij})}$$

where

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

$s_j = P(X_{ij} = 0 | \xi_{ij} = 1)$ -- "slip" parameter

$g_j = P(X_{ij} = 1 | \xi_{ij} = 0)$ --"guess" parameter

How to select items sequentially?

- No direct analogy to “match theta with difficulty”
  - Regular CAT, b-parameter with θ
- Now α is a vector, called latent class

$$K:\ \text{\# of attributes}$$

$$\alpha_c:\ \text{pt in the latent space } (2^K)$$

$$\hat{\alpha}:\ \text{estimated } \alpha$$

$$P(X_i = x_i \mid \alpha):\ \text{IRF}$$

Developing Item Selection Algorithms for CD-CAT

1. Item selection based on theta
   – adaptive theta-estimates
   – non-adaptive alpha-estimates
2. Item selection based on alpha
   – adaptive alpha-estimates
   – non-adaptive theta-estimates
3. Item selection based on both theta & alpha
   – both estimates adaptive

Item selection based on alpha

- **KL information Approach** (Xu, Chang, & Douglas, 2004)

$$KL_{jc}(\hat{\alpha} \| \alpha_c) = \sum_{x=0}^{1} \log\left[\frac{P(X_j = x | \hat{\alpha})}{P(X_j = x | \alpha_c)}\right] P(X_j = x | \hat{\alpha})$$

Select item j to make the following as large as possible

$$KL = \sum_{c=1}^{2^K} KL_{jc}(\hat{\alpha} \| \alpha_c)$$

Item selection based on alpha (cont.)

- ***The Shannon Entropy Method (SHE)***
  - Minimize SHE

$$E[SHE_j(\alpha)] = E\left\{\sum_{k=1}^{K} P(\alpha_k \mid X_1, X_2, \ldots, X_{j+1}) \log\left[\frac{1}{P(\alpha_k \mid X_1, X_2, \ldots, X_{j+1})}\right]\right\}$$

$$= \sum_{x=0}^{1}\left\{\sum_{k=1}^{K} P(\alpha_k \mid X_1, X_2, \ldots, X_{j+1}) \log\left[\frac{1}{P(\alpha_k \mid X_1, X_2, \ldots, X_{j+1})}\right]\right\} P(X_{j+1} = x)$$

where

$$p(\alpha_k \mid X_1, \ldots, X_{j+1}) \propto P(X1, \ldots, X_{j+1} \mid \alpha_k) P(\alpha_k)$$

Estimate both θ and α Adaptively

- ***Shadow Test Approach*** *(McGlohen & Chang, 2004)*
- ***Dual Information Approach*** *(Cheng & Chang, 2007)*

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

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

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

Dual Information: $KL_j(\hat{\theta}_m, \hat{\alpha}_m) = wKL_j(\hat{\alpha}_m) + (1-w)KL_j(\hat{\theta}_m).$

Aggregate Ranked Information method (ARI)

- Wang, Chang & Wang (2011)

$$KL(\hat{\theta}) = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} \left\{ [p(\hat{\theta})\log\left[\frac{p(\hat{\theta})}{p(\theta)}\right] + [1-p(\hat{\theta})]\log\left[\frac{1-p(\hat{\theta})}{1-p(\theta)}\right] \right\} d\theta$$

$$\approx \sum_{i=-k}^{k} [p(\hat{\theta})\log\left[\frac{p(\hat{\theta})}{p(\hat{\theta}+i\Delta\theta)}\right] + [1-p(\hat{\theta})]\log\left[\frac{1-p(\hat{\theta})}{1-p(\hat{\theta}+i\Delta\theta)}\right] \Delta\theta$$

$$ARI = \lambda pe(KL(\hat{\alpha})) + (1-\lambda)pe(KL(\hat{\theta}))$$

Incorporate Multiple Constraints in CAT

- ***Goal: making programming so easy that most practitioners can implement constraint control by themselves***
- ***Example: A Weighted Priority Index (Cheng & Chang, 2008)***

$$f_{jm} = \frac{(u_m - x_m)}{u_m}.$$

$$f_{jk'} = \frac{\left(r_j - \frac{N - n_j}{N}\right)}{r_j},$$

A multiplier

$$p_j = I_j \prod_{k=1}^{K} (w_k f_{jk})^{c_{jk}}$$

Objective function

weight

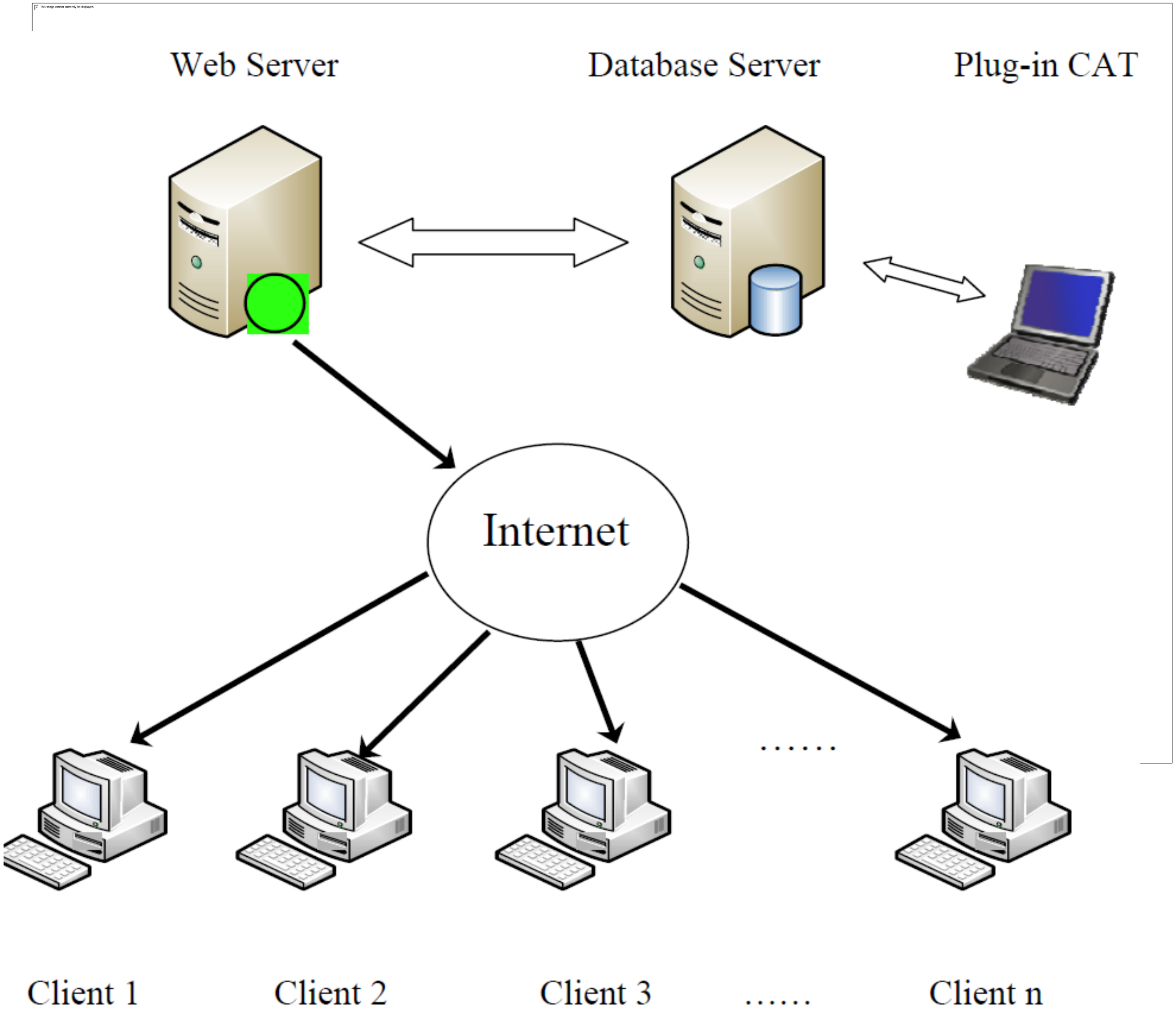
Why CD-CAT?

HOW TO HELP SCHOOLS TO OWN AND OPERATE CD-CAT?

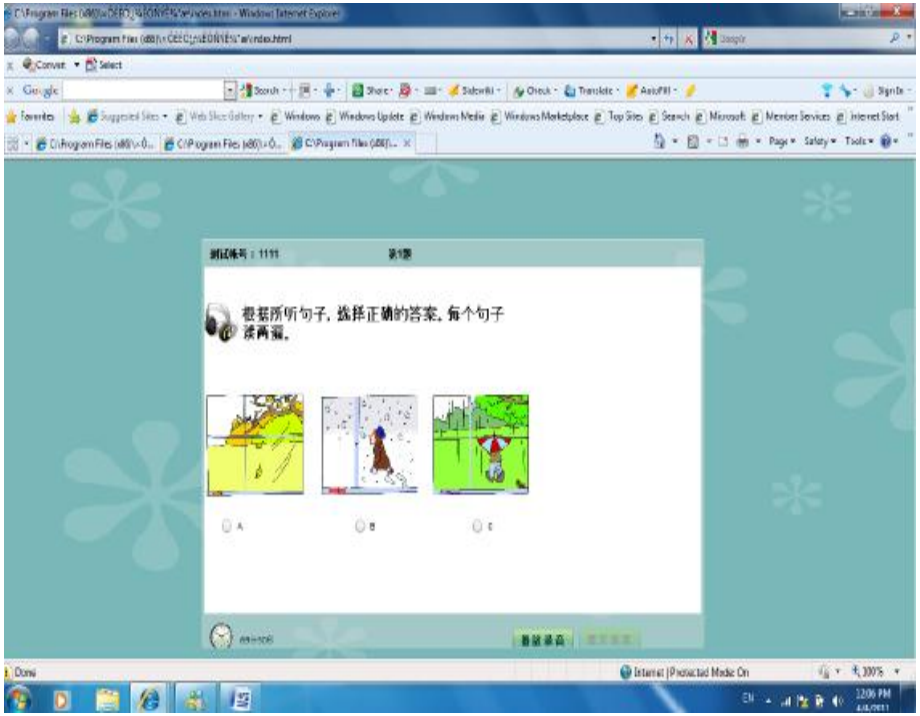
New Technologies
--- Schools can use existing PCs

- Client/Server Architecture (CS)
  - CAT software has to be installed on each client computer (large workload)
  - only applicable to Local Area Network (LAN)
- **Browser/Server Architecture (BS)**
  - database is still on the server
  - nearly all the tasks concerning development, maintenance and upgrade, are carried out on the server.
  - based on the Wide Area Network (WAN)

Hardware and Network Design



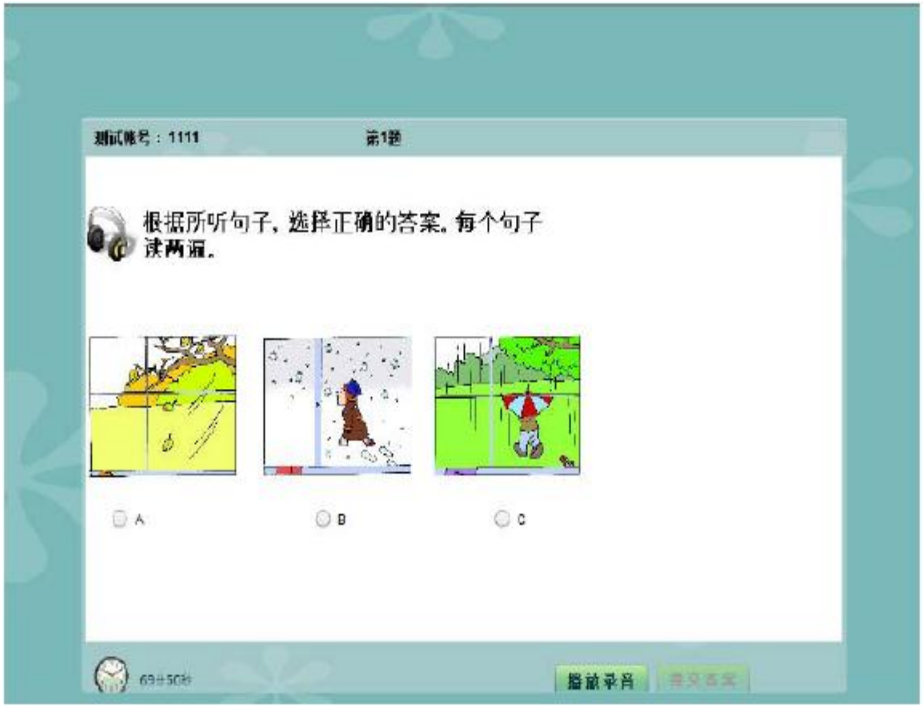
Use Web-Browser to Deliver Test



测试帐号：1111 第1题

根据所听句子，选择正确的答案。每个句子读两遍。

○ A ○ B ○ C



测试帐号：1111 第1题

根据所听句子，选择正确的答案。每个句子读两遍。

○ A ○ B ○ C

69分50秒

Develop a CD-CAT system to show its ability to improve teaching and learning

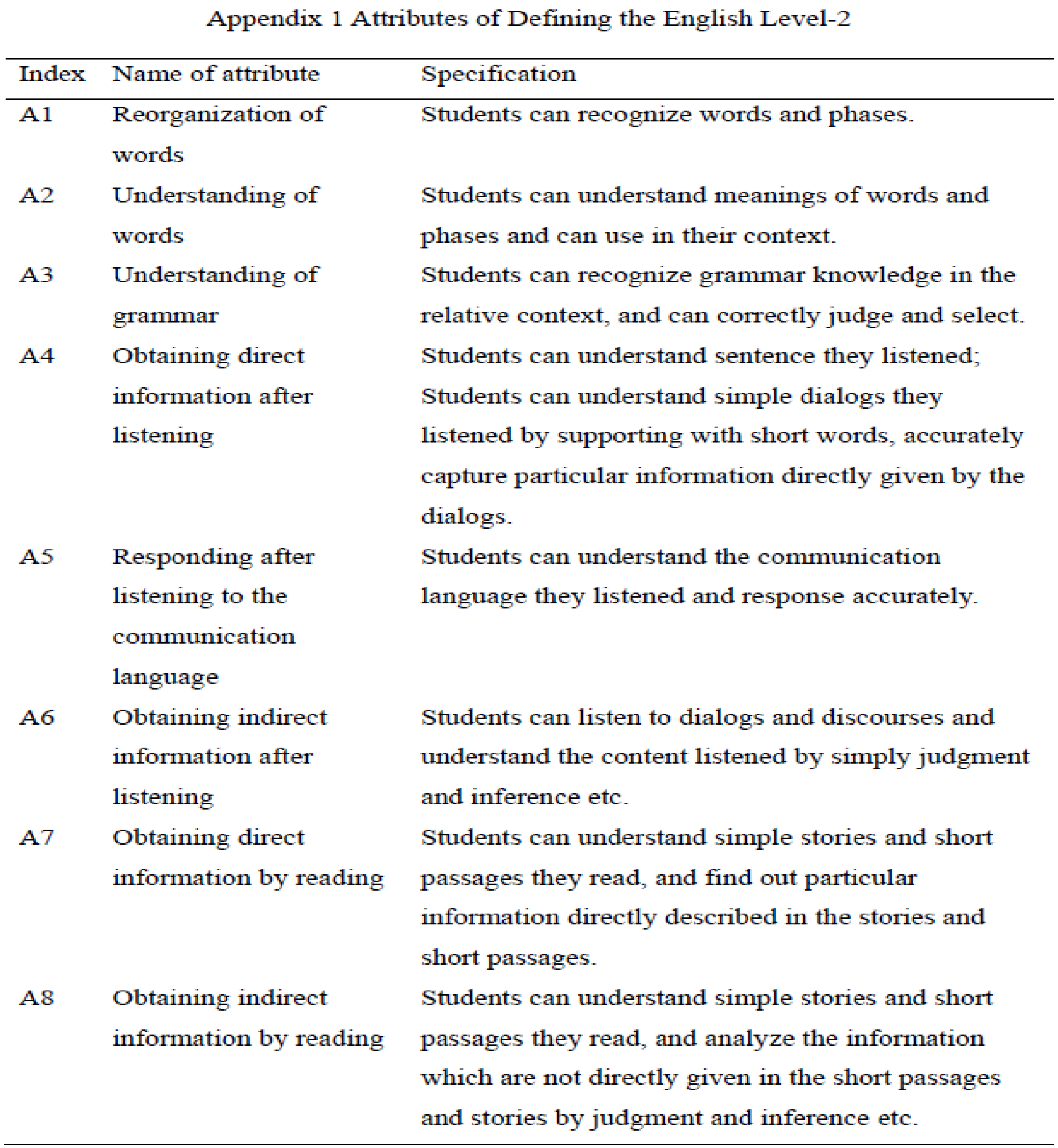
APPLICATION: THE CHINA PROJECT

Application (Liu, 2010): Level II English Proficiency Test

- Pretest and Calibration of Item bank
  - Pretest
    - 38,662 students from 78 schools, 12 counties participated
  - Analyzing pretest data
    1. Estimated the parameters of DINA model
    2. Estimated the parameters of 3PLM model
    3. Calibrate attributes of item again
    4. If it fits well then stop, otherwise revise q-matrix and got 3
  - Assembling the item bank with item parameters and specifications.

Item Writing

- About 40 Teachers in Beijing
- Process
  1. Psychometric Training
  2. Identify Attributes
  3. Writing Items
  4. Constructing Q-matrix
  5. Pre-testing and check FITTING
  6. Revise Q-matrix until fitting is ok; go to 5 if not
  7. stop



Appendix 1 Attributes of Defining the English Level-2

Index	Name of attribute	Specification
A1	Reorganization of words	Students can recognize words and phases.
A2	Understanding of words	Students can understand meanings of words and phases and can use in their context.
A3	Understanding of grammar	Students can recognize grammar knowledge in the relative context, and can correctly judge and select.
A4	Obtaining direct information after listening	Students can understand sentence they listened; Students can understand simple dialogs they listened by supporting with short words, accurately capture particular information directly given by the dialogs.
A5	Responding after listening to the communication language	Students can understand the communication language they listened and response accurately.
A6	Obtaining indirect information after listening	Students can listen to dialogs and discourses and understand the content listened by simply judgment and inference etc.
A7	Obtaining direct information by reading	Students can understand simple stories and short passages they read, and find out particular information directly described in the stories and short passages.
A8	Obtaining indirect information by reading	Students can understand simple stories and short passages they read, and analyze the information which are not directly given in the short passages and stories by judgment and inference etc.

Liu et al. (April, 2010):

- If the models do not fit well with the pretest data, adjusting Q matrix yields better fitting.
- Test developers and psychometricians should work together to tackle "bad-fit".
- “Model fitting" would be improved by fine-tuning the Q matrix after re-examining the cognitive process that examinees might use to solve the problems.
- Psychometric training increased quality of item writing, e.g., item discrimination would increse.

Field Test in 2010

- SHE with content constraints
- The adaptive test was web-based, consisting of 36 items and lasting for 40 minutes.
- *Number of Participants:* 584
  - 5th and 6th grade, from 8 schools in Beijing, China

Validity Study

- Evaluating the consistency of
  - CD-CAT system results with an existing English achievement test
    - a group of students took two exams
  - CD-CAT system results with Teachers’ evaluation outcomes.

CD scores vs. an achievement test

The Consistence between levels and # of mastered attributes

	# of mastered attributes									
Academic Performance Level	0	1	2	3	4	5	6	7	8	Total
Excellent	0	0	1	1	1	3	4	6	23	39
Good	0	0	1	2	8	5	7	7	3	33
Pass	1	1	3	5	3	1	0	0	1	15
Fail	0	1	2	0	0	0	0	0	0	3
Total	1	2	7	8	12	9	11	13	27	90

CD-CAT Results vs. Teachers'

- *Comparison of a CD scores with teachers' assessment*
  - *Participants from three classes:*
    - 91 6-grade students and 3 teachers were recruited to evaluate the diagnostic reports. one rural school and two urban schools.
  - Measurement
    - Students' diagnostic reports were presented to three teachers, they were asked to evaluate the accuracy of this report.

Validity Study: CD vs. Teachers

Evaluation on the CD-CAT feedback reports by teachers

Teacher	High consistency	medium consistency	low consistency	total
A	28(90.32)	3(9.68)	0(0.00)	31(100)
B	13(41.94)	16(51.61)	2(6.45)	31(100)
C	27(93.10)	1(3.45)	1(3.45)	29(100)
total	68(74.73)	20(21.98)	3(3.30)	91(100)

Progress in 2011

- 4000 items were developed
  - Why so many items?
  - Now each grade has 10 units and each unit has an item bank
- The B/S based delivery was tested with 500 PCs in Dalian!
- 30,000 students participated field tests
- A large scale validity study will be conducted
- See some pictures in the field testing...

Discussions

- Large scale field tests will also take place in Shanghai in the near future.
- CD-CAT can be built very economically.
- Though the DINA model was used, the results can be generalized to many other IRT and Cognitive Diagnostic Models!
- The method for on-line calibrating of pre-test items has been developed. In the future, paper/pencil based pretesting is not needed.
- What Are We Waiting For?

Thank you！