Differences in the Amount of Adaptation Exhibited by Various Computerized Adaptive Testing Designs

Mark D. Reckase
Unhee Ju
Sewon Kim
Michigan State University

Background

- Early versions of testing tended to be adaptive – an examiner asked a question, listened to a response, then asked follow up questions.
  - The follow-up questions were based on the responses from the examinee.
  - I did this with my students when I first began teaching.
- Binet (1905) formalized the examination process be providing rules to the examiner for scoring and selecting the next question – questions were prepared in advance.
- These are intelligent testing systems.
- The outgrowth of these intelligent testing systems is the computerized adaptive test (CAT) – the computer now does the questioning, scoring, and next item selection according to programmed rules.
- One characteristic of intelligent testing systems is that the set of questions is customized to each examinee.

How Much Customization Is There?

- With a human examiner, there may be total customization – each examinee may get a different set of questions.
- As the process becomes more formalized, the amount of customization may be reduced.
  - Ultimately, if a fixed test is given, there is no customization.
  - It would be difficult to argue that a fixed test is an intelligent testing system.
- The purpose of this presentation is to present some results on how much customization takes place in CATs.
  - It is our observation that some CATs are so constrained, or have such limited item pools, that there is little customization.
  - We believe that the amount of customization should be made evident so it can be determined how intelligent the testing system is.

Goals for this Presentation

- Present some indicators of the amount of adaptation/customization exhibited by an adaptive test.
- Provide some benchmark values for the indicators based on previous simulation studies.
- Show the results of applying the indicators to some common CAT test designs.
- Show some results for an operational CAT using a multistage test design.

Assumptions

- Basic assumption is that the goal of the adaptive test is to select the set of test items for a person that will give the best estimate of location on the reporting score scale.
- Hypothetical best case is:
  - Person’s location is known.
  - All items selected to have maximum information at the person’s location.
    - Difficulty matched to person’s location
    - Low variation in point of maximum information for the set of items administered to the person.

Proposed Indicators of Adaptation

- Each provides information about a different aspect of adaptation.
  - Correlation between the average difficulty or average point of maximum information of the items administered (ζ is used to represent either average difficulty or average point of maximum information) and the final proficiency estimate: $r(\zeta_j, \hat{\theta}_{\mathrm{j}})$. **Ordering of the sets of items administered.**
  - The ratio of the standard deviation of the average difficulties or points of maximum information to the standard deviation of the proficiency estimates: $s_{\zeta_j}/s_{\hat{\theta}_{\mathrm{j}}}$. **Spread of the sets of items administered.**
  - The proportion reduction in variation of the item difficulty or point of maximum information of items administered persons compared to the variation of difficulty in the item pool: $PRV = \frac{s_\zeta^2 - pooled\ s_{\zeta_j}^2}{s_\zeta^2}$. **Focus of the set of items administered.**

Benchmark Values

	$r(\bar{\zeta}_j, \hat{\theta}_\mathrm{j})$	$s_{\bar{\zeta}_j}/s_{\hat{\theta}_\mathrm{j}}$	$PRV = \dfrac{s_\zeta^2 - pooled\ s_{\zeta_j}^2}{s_\zeta^2}$
Benchmark	Low .90s	Mid .80s	About .80
NCLEX	.92	.96	.84

Comments about the Benchmarks

- Values were determined using an item-level CAT using the one-parameter logistic model (Rasch).
- The ratio indicator can exceed 1.0 if the item pool has many extreme items, but few in the middle. Distance from 1.0 is the main indicator.
- Statistics are sensitive to pool size, pool spread, and exposure control.
- These results were presented at the 2017 IMPS meeting and a paper has been submitted for publication.

The Focus
# of the Research Reported Here

Comparison of the Amount of Adaptation
of a traditional Item-Level CAT,
an *a*-stratified CAT with *b*-blocking,
and two variations of a Multi-stage Test

Three Types of Test Designs

- Traditional CAT
  - Maximum information item selection
  - Maximum likelihood proficiency estimation
  - 40 item test length
- *a*-Stratified with *b*-blocking
  - Items sorted into four strata based on *a*-parameters for fixed ranges of *b*-parameters
  - Ten items selected from each strata with the lowest *a*-parameter strata used first.
  - Item selection same is traditional CAT
- Multi-stage test
  - Used a 1-3-3 three-stage design with fixed modules within each stage
  - First two stages designed to allocate equal numbers of examinees to modules – stage three designed to approximate uniform information.

The Challenge of a Fair Comparison

- Want to have a fair comparison between the three designs
  - Use the same item pool for item selection
  - Have similar goals for the assessments
- Designed an optimal item pool for the traditional CAT using the bin-and-union method
  - Selected items from a master pool to approximate the requirements of the optimal item pool
  - Optimal pool specifications called for 407 items spread over a wide range of difficulty
  - Operational pool had 319 items because had too few items in the master pool at extreme ranges of difficulty
- Sorted operational pool into four strata using procedure specified by Chang, Qian, and Ying (2001)
- Created the MST panel from the item pool to:
  - Have equal usage of the modules
  - Have uniform information over -3 to 3.

Simulation Design

- 3000 master pool generated to have the same multivariate distribution of *a*-, *b*-, and *c*-parameters as an actual item pool – matched marginal distributions and correlations between parameters.
  - Traditional CAT and stratified design used the operational 319 item pool selected from the 3000. This was replicated several times to determine of the particular selection of items made a difference.
  - The multi-stage test had 20 items in the first stage and 10 items at the second and third stage – used the best 80 items from the 319 item operational pool
- Examinees sampled from $N(0, 1)$.
  - 500 examinees
  - Process was replicated 100 times to get variation in the adaptation measures.

CAT Simulation

- Starting value 0.0.
- Select items to maximize information.
- Estimate proficiency using maximum likelihood.
  - Until correct and incorrect responses are present, increment the estimate by .7 or -.7.
- Stop at 40 items.

Multi-Stage Test
Equal Frequency over Modules

Stage	Difficulty Level	Number of Items	Routing Points	Mean *b*	SD *b*	Min	Max
First-stage		20	-0.44, 0.44	-0.05	0.77	-1.59	1.05
Second-stage	Easy	10	-0.44	-0.59	0.82	-1.77	0.92
	Medium	10	-0.44, 0.44	-0.49	1.19	-3.00	1.21
	Difficult	10	0.44	0.04	0.83	-1.63	0.66
Third-stage	Easy	10		-2.83	0.07	-2.96	-2.76
	Medium	10		-0.42	0.04	-0.44	-0.38
	Difficult	10		2.21	0.14	1.99	2.39

Checks on Simulations Proficiency Estimates

Test Design		True θ		Estimated θ			Correlation $(\theta,\ \hat{\theta})$
		Mean	SD	Mean	SD	SE	
Traditional CAT		0.01	1.02	0.012	1.042	0.194	0.980
Stratified CAT	Match b	0.01	1.02	0.020	1.050	0.244	0.970
	Max. Info	0.01	1.02	0.024	1.040	0.212	0.981
Multi-Stage Test 20-10-10		0.01	1.02	0.020	1.086	0.318	0.956
Multi-Stage Test 10-10-20		0.01	1.02	0.010	1.080	0.320	0.960

Results are based on five replications.

Proficiency Estimates

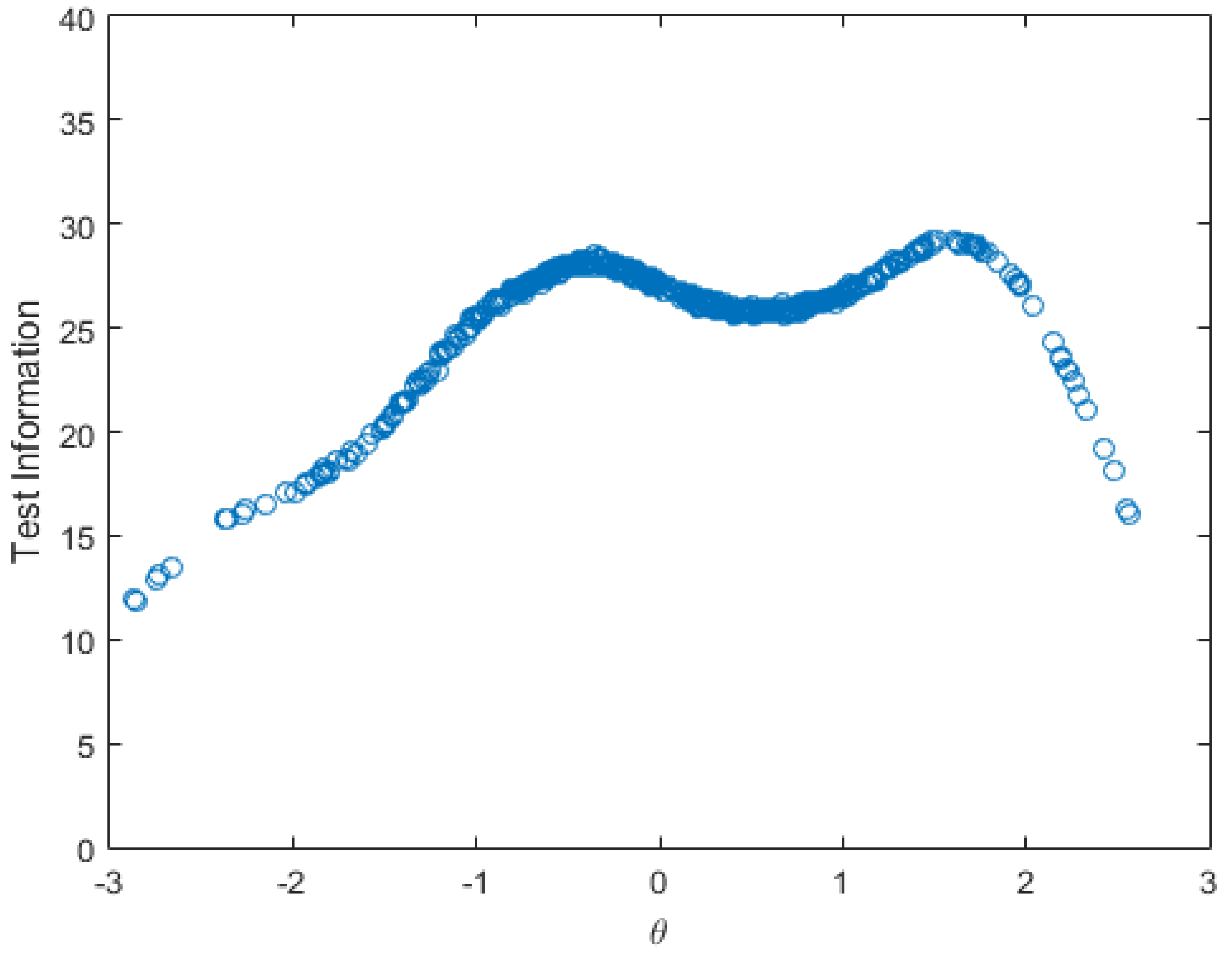
- Estimates function as expected – simulations are working well
  - High correlations – slightly lower for multi-stage test.
  - Smallest standard error for CAT.
  - Largest standard error for multi-stage test.
- For comparison, a fixed 40 item test gives a standard error of .42 and correlation of .93.

Multi-Stage Routing

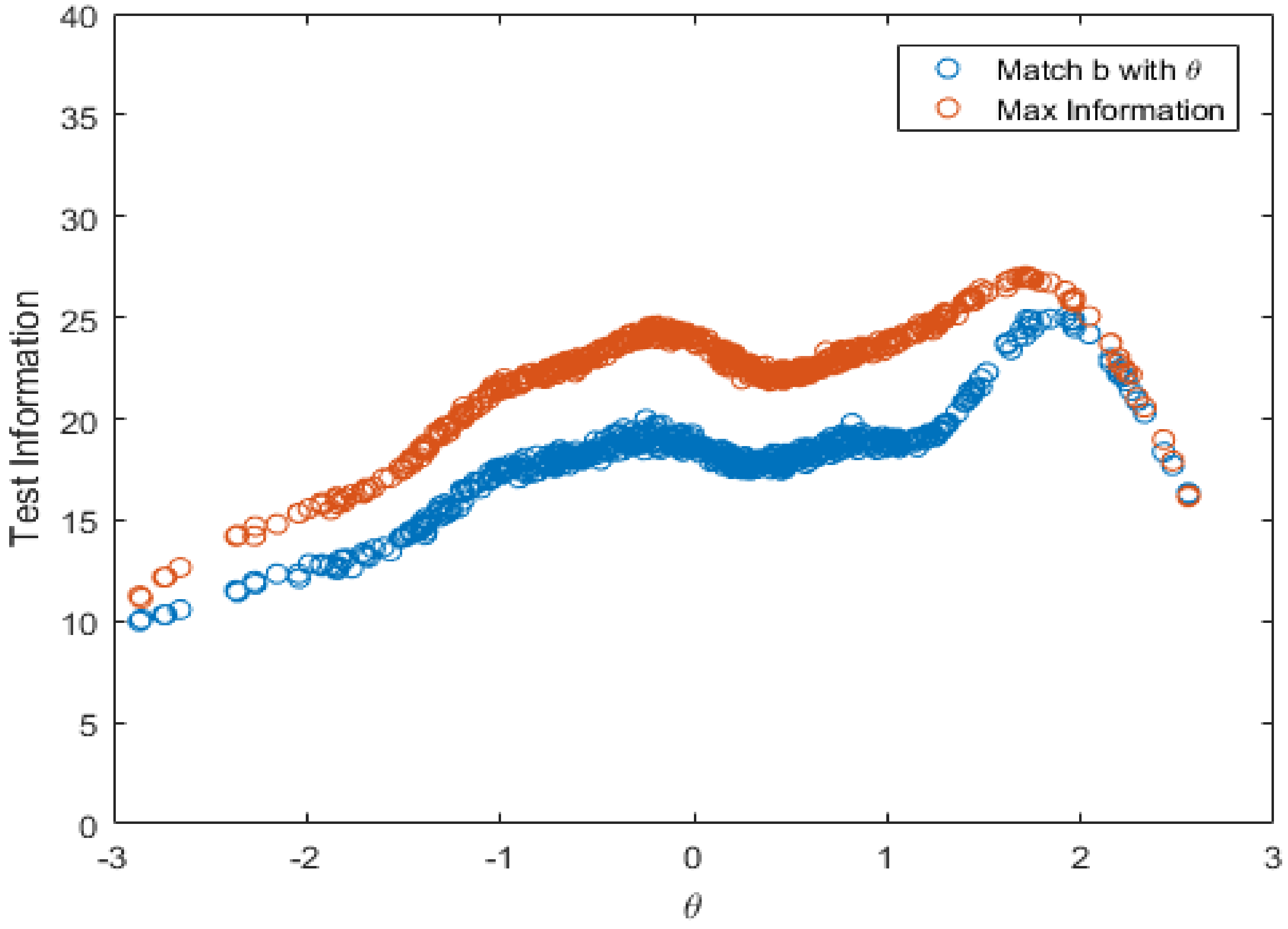
- Equal distribution routing gives approximately equal usage for the 20-10-10 number of items in modules.
- The overall level of accuracy of classification into modules was 76.68% at Stage 2 and 80.26% at Stage 3.

Test Design	Stage	Module (%)		
		Easy	Medium	Hard
Equal Distribution	2nd	28.86	31.36	39.74
	3rd	29.18	35.43	35.28

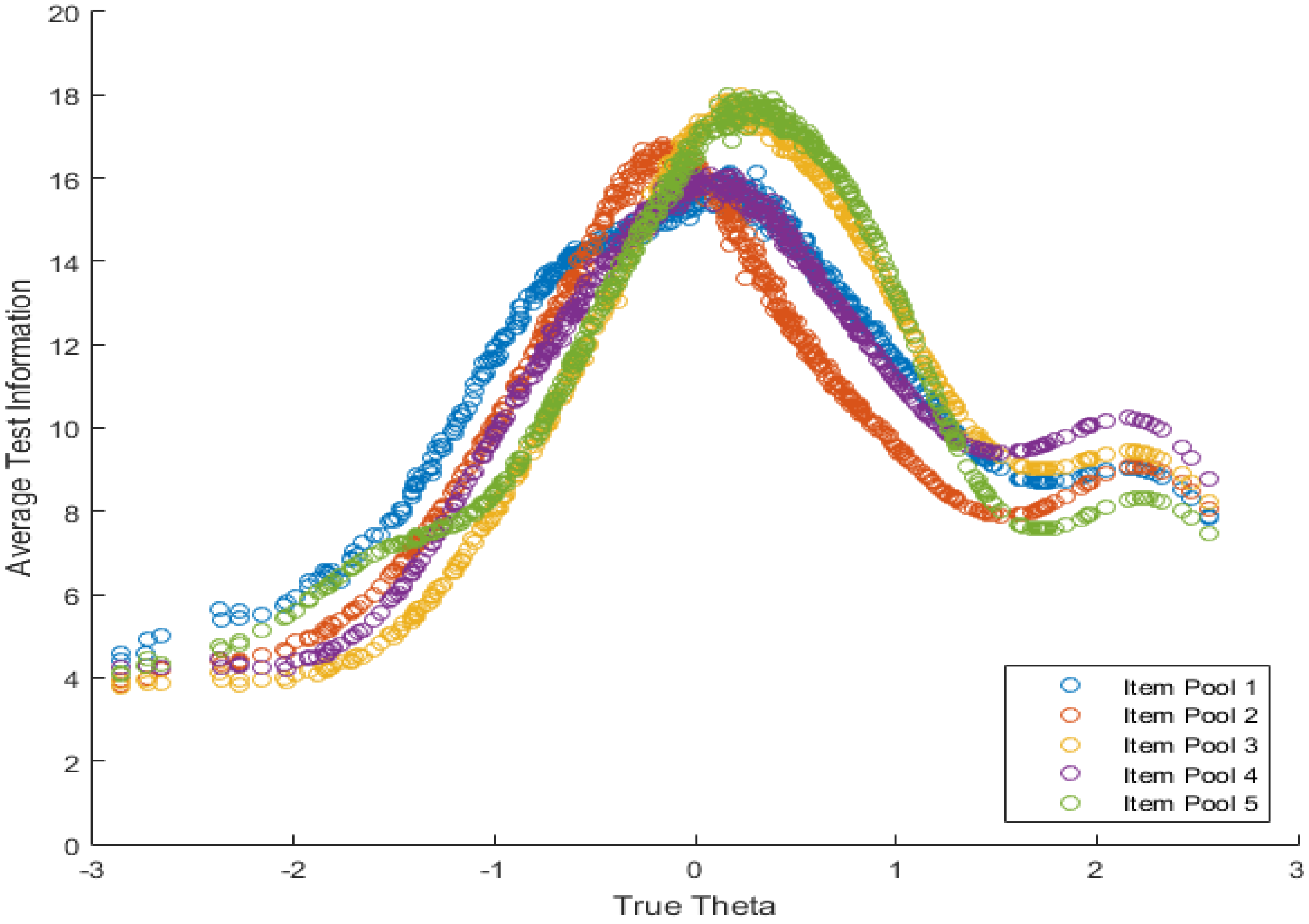
Information from Traditional CAT



Information from Stratified CAT



Information from Multi-stage Test 20-10-10 Design



Test Information Plots

- Traditional CAT gives information between 25 and 30 between -1 and 2.
- Stratified CAT gives slightly lower information over the range.
  - Maximum information item selection does better than closest *b*-parameter selection.
- Multi-stage test gives peaked information with maximum below 20.
  - The multiple curves show the five replications of item selection for the test design.
  - The information is lower for examinees in the tails of the distributions.

Adaptation Statistics for the Test Designs

		Statistics		
		$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV
Traditional CAT		0.99 (0.01)	0.88 (0.01)	0.89 (0.00)
Stratified CAT	Match b	0.94 (0.00)	0.91 (0.10)	0.84 (0.00)
	Max. Info	0.97 (0.01)	0.89 (0.01)	0.87 (0.00)
Multi-Stage Test 20-10-10		0.85 (0.02)	0.51 (0.01)	0.41 (0.01)
Multi-Stage Test 10-10-20		0.85 (0.02)	0.94 (0.02)	0.56 (0.01)

Note. Parenthesis = empirical standard deviation (i.e., standard error)

Adaptation Results

- CATs noticeably better than the multi-stage test.
- Traditional CAT slightly better than the Stratified CAT.
  - Stratified CAT has fewer items to select from because of stratification.
- For the multi-stage test, results are different depend of distribution of test items over modules. The PRV are well below the benchmark values for both designs.
  - Because of the use of modules of items, there is less focus of difficulty or information for the examinees.
  - Using longer modules at the end of the routing yields better adaptation.

Real Data Analysis

- Data from the Early Childhood Longitudinal Study – 1998/1999 Cohort
  - Third grade tests
  - Tests in Mathematics, Reading, and Science
  - Two-stage Test with 1-3 Design
  - Size of modules varied by test.
- Based on the three-parameter logistic model

Real Data Adaptation Results

Test	Item Statistic	Adaptation Statistic		
		$r(\bar{b}_j, \hat{\theta}_j)$	$s_{\bar{b}_j}/s_{\hat{\theta}_j}$	PRV
Mathematics	b-parameter	0.86	0.74	0.47
	Maximum information point	0.86	0.73	0.47
Reading	b-parameter	0.76	0.44	0.25
	Maximum Information Point	0.76	0.46	0.26
Science	b-parameter	0.83	0.50	0.42
	Maximum Information Point	0.83	0.49	0.40

Real Data Adaptation Results

- Mathematics most adaptive and reading least adaptive.
- Reading results poor compared to multi-stage simulation.
  - Reading tests are challenging because items come in sets with reading passages.
- Science results are better than reading, but still have evidence of poor adaptation.

Conclusions

- Statistics measures of adaptation give useful information about amount of adaptation.
- Multi-stage tests are less adaptive than item-level CATs – in some cases might merit the term “barely adaptive tests” (BATs).
- The operational tests were less adaptive than they could be with well designed modules and routing rules.

Conclusions

- There was an initial premise that adaptive tests are a type of intelligent testing system than mimics what a person would do who is evaluating a student with a one-on-one conversation.
- Traditional CATs and Stratified CATs seem to approximate that type of intelligent testing system – examinees tend to get unique sets of items.
- Multi-stage tests seem less like intelligent testing systems because there is limited opportunity to give unique sets of items and there is little opportunity to correct for miss-routing.
- How adaptive does a test have to be before it is considered as an intelligent system?
- Newer adaptive systems such as cognitive diagnostic tests and complex simulations are also expected to be adaptive. It may be possible to adapt the statistical indicators to those new forms of tests as well.