[redacted]

in the Estimation of Item Difficulties:

Implications for a CAT Implementation

Kyoko Ito
Robert C. Sykes

CTB McGraw-Hill

Abstract

Responses to previously calibrated items administered in a computerized adaptive testing (CAT) mode may be used to recalibrate the items. This live-data simulation study investigated the possibility, and limitations, of on-line adaptive recalibration of precalibrated items.

Responses to the items of a Rasch-based paper-and-pencil licensure examination were used to simulate CAT and paper-and-pencil administrations. To simulate CAT conditions, CAT forms having varying difficulty levels were defined, and CAT samples of varying sizes were selected from the ranges of ability that roughly corresponded to the difficulty levels of the items. To simulate paper-and-pencil test conditions, two shorter versions of the conventional examination were constructed, and

Responses to the CAT items were extracted from the CAT samples and, for comparison purposes, from the representative samples as well. Similarly, responses to the shorter paper-and-

Introduction

Computerized adaptive tests (CAT) cannot be successfully administered without reliable item parameter estimates. In the first few years of operation, examinees can be scored using item parameter estimates obtained from previous paper-and-pencil administrations of the items, provided no mode effect exists. Under these circumstances, field-test items may also be calibrated "on-line" from administrations of linear computerized tests or linear portions of CAT (i.e., **on-line non-adaptive calibration**). Because these tests are not targeted on examinee ability, the resulting parameter estimates from linear

Previously scored items will likely have to be recalibrated after a period of operational administrations, for reasons such as the need to evaluate scale drift. An obvious way to recalibrate the items would be on-line non-adaptive recalibration. In this case, previously scored items will be treated just like field-test items; namely, items will not be selected based on their parameter estimates, but rather, they will be (quasi-)randomly administered to examinees. Because items are administered to examinees having a wide range of ability, on-line non-adaptive recalibration of precalibrated items should yield parameter estimates similar to those from a paper-and-pencil calibration.

ability.

All forms conformed to the test-plan specifications.

[redacted]

Characteristics of the Samples

The CAT samples were selected using examinees' ability estimates based on the total of 298 scored items in the PP. The three ability bands from which the CAT samples were drawn are shown below, along with the corresponding b-value ranges.

Items : B-Value Range	CAT Samples: Theta Range
-0.63 - 0.86 (Hard)	-0.5 - 0.0 (High)
-1.17 - -0.55 (Medium)	-1.0 - -0.5 (Middle)
-1.96 - -0.98 (Easy)	-1.7 - -1.05 (Low)

(unit : logit)

Targeting for on-line adaptive recalibration will result in more difficult items being calibrated on responses of more able examinees, and easier items calibrated on responses of less able examinees. For those examinations in which subpopulations differ in their mean ability, this also means the confounding of

[redacted]

For example, if members of group A, on average, were more able than members of group B, more difficult items would tend to be given to group A, and easier items to group B. This would

[redacted]

this examination were whites, which is the reference group for DIF analyses for the examination.

The means and standard deviations of theta estimates for the Representative samples (for both the First-Time U.S. and Ethnic Group) are provided in Table 2. As expected, the Ethnic Group Representative samples have substantially different mean ability

2,100), the Ethnic Group Representative samples had a mean theta of -0.810, as opposed to the 0.071 mean for the Representative First-Time U.S. samples. Furthermore, the Ethnic Group Representative samples consistently had somewhat greater dispersion of theta estimates.

Four sample sizes were considered: 100, 200, 400, and

Results

Results for the First-Time U.S.: The results for the First-Time U.S. group are presented in Tables 3, 4, and 5, respectively, for the CAT forms, Mini PP, and Shorter PP. The Easy CAT form was not analyzed for the First-Time U.S. group because of insufficient cases in the ability range corresponding to the b-value range of easy items.

Two points must be born in mind in evaluating the results. First, as mentioned earlier, the mean ability estimate for the entire reference group who had taken the full-length PP form was 0.6. Second, the benchmark b-values also came from this group of First-Time U.S. educated examinees. Consequently, the results for the First-Time U.S. group will be better than those for the Ethnic Group.

Simulated CAT Forms: Table 3 shows that the correlations for the Hard form tended to be in the .80's and .90's, whereas those for the Medium Difficulty form tended to be between .50's and .70's. The MADs for the Hard form were in the .10's and those for the Medium Difficulty form were in the .10's and .20's. Thus, the Hard form produced b-values more similar to bank b-values.

The difference in correlations between the Hard versus Medium Difficulty forms would likely have been smaller if the Medium Difficulty form had contained items from as wide a range

in the theta range between -0.5 and 0.0 logit was more similar to

[redacted]

The correlations for the Medium Difficulty and Easy forms were, for a large part, in the .10's, and the MADs in the .50's. The

again verified the earlier finding with the First-Time U.S. that increasing test length from 75 items to 149 items did not improve

First-Time U.S. versus Ethnic Group: The results for the First-Time U.S. were compared with those from the Ethnic Group.

Second, for the "difficult-items-to-the-able: easy-items-to-the-less-able" on-line adaptive recalibration, only items of similar difficulty levels will be recalibrated simultaneously. In other words, difficult items will not be recalibrated with easy items, because the two sets of items will have been taken by two different groups of examinees. The resulting reduction in the

that was more heterogenous in ability but dissimilar to the reference group with regard to the mean theta.

Third, items to be (re)calibrated together must be relatively heterogenous in difficulty. The results improved substantially as the standard deviation of b-values for a set of items increased. The CAT forms simulated in this study had a b-value standard deviation about one-fifth to less than one half the standard deviation of the paper-and-pencil examination. Whether actual CAT forms will have a greater b-value standard deviation depends on various factors, such as stopping rules and the location of the initial item.

Items in a CAT form will typically be more similar in difficulty than items in a conventional form. However, by choosing an appropriate starting item (e.g., a relatively easy item for a relatively able reference group), many reference-group examinees will be administered a wider range of items.[1] Although this would mean a loss of efficiency, the amount of loss depends on several factors. If a large number of items can be

Since the present study was a live-data simulation, the findings need to be replicated with real CAT data. However, the

Table 1

Descriptive Statistics for the Forms Used

		Bank B-Value				Abs. 1st Lds.		Abs. 2nd Lds.	
Form	# Items	Mean	SD	Min	Max	Mean	SD	Mean	SD
Simulated CAT Forms:									
Hard	73	-.02	.34	-.63	.86	.19	.09	.10	.08
[redacted]									
Easy	68	-1.34	.23	-1.96	-.98	.19	.10	.10	.08
Mini PP	75	-0.98	.81	-3.03	.60	.21	.10	.10	.07
Shorter PP	149	-0.97	.80	-3.03	.63	.20	.10	.10	.08
PP	298	-0.97	.79	-3.06	.86	.20	.10	.10	.08

Table 2

[redacted]

Subpopulations

	Ethnic Group		1st-Time, US	
N	Mean	SD	Mean	SD
400	-.834	.479	.050	.376
	-.826	.487	.091	.391
	-.783	.502	.049	.394
Mean	**-.814**	**.489**	**.063**	**.387**
200	-.831	.442	.102	.355
	[redacted]	[redacted]	[redacted]	[redacted]
	-.842	.520	.084	.387
Mean	**-.818**	**.480**	**.082**	**.375**
100	-.786	.491	.102	.432
	-.807	.478	.042	.358
	-.731	.421	.104	.389
Mean	**-.775**	**.463**	**.083**	**.393**
2,100[1]	**-.810**	**.484**	**.071**	**.385**

[1] All cases combined. Because the smaller samples (i.e., N=400, 200, & 100) were drawn with no replacement, the combined total

Table 3

[redacted]

Difference (MAD) :
Form = Simulated CAT Forms
Sample = First-time U.S.[1]

	Item Difficulty							
	Medium (-1.17 - -0.55)				Hard (-0.63 - 0.86)			
	Ability Range				Ability Range			
N	CAT Sample: Middle[2]		Repre-sentative[3]		CAT Sample: High[4]		Repre-sentative[3]	
	r	MAD	r	MAD	r	MAD	r	MAD
100	.486	.235	.537	.214	.819	.176	.833	.165
[redacted]								
			.539	.228			.882	.161
Mean	**.511**	**.244**	**.554**	**.216**	**.807**	**.183**	**.861**	**.163**
Corrected[5]	**(.765)**		**(.799)**					
200	.610	.230	.671	.157	.867	.152	.884	.146
	.586	.198	.547	.167	.861	.155	.887	.149

Table 4

Correlation B/w Bank b's and Part-form b's (r) &
Their Mean Absolute Difference (MAD) :
Form = Mini PP
Sample = First-time U.S.

	Ability Range					
N	r	MAD	r	MAD	r	MAD
100	.915	.266	.944	.193	.951	.203
					.930	.252
Mean	.904	.284	.943	.204	.944	.222
200	.920	.265	.961	.182	.978	.148
	.925	.246	.969	.154	.973	.164
Mean	.923	.256	.965	.168	.976	.151

Table 5

Correlation B/w Bank b's and Part-form b's (r) &
Their Mean Absolute Difference (MAD) :
Form = Shorter PP
Sample = First-time U.S.

	Ability Range					
	CAT Sample:		CAT Sample:		Repre-	
N	r	MAD	r	MAD	r	MAD
100	.915	.266	.944	.207	.933	.234
	.891	.295	.940	.218	.948	.223
					.939	.239
Mean	**.903**	**.281**	**.942**	**.213**	**.940**	**.232**
200	.926	.253	.962	.175	.968	.172
	.926	.238	.969	.151	.967	.174
					.974	.142
Mean	**.926**	**.246**	**.966**	**.163**	**.970**	**.163**
400	.938	.227	.979	.135	.983	.125
	.931	.241	.973	.143	.984	.112
					.982	.127
Mean	**.935**	**.234**	**.976**	**.139**	**.983**	**.121**
1000	.941	.226	.983	.122	.989	.096

[1] Between -1.0 and -0.5.
[2] Between -0.5 and 0.0.

ble 6

Correlation B/w Bank b's and Part-form b's (r) & Their Mean Absolute Difference (MAD) :
Form = Simulated CAT Forms
Sample = Ethnic Group

	Item Difficulty					
	Easy (-1.96 - -.98)		Mdm (-1.17 - -.55)		Hard (-.63 - .86)	
	Ability Range		Ability Range		Ability Range	
N	CAT Sample: Low[1]	Rep.[2]	CAT Sample: Middle[3]	Rep.[2]	CAT Sample: High[4]	Rep.[2]
	r MAD	r MAD	r MAD	r MAD	r MAD	r MAD
100	.101 .581	.201 .552	.065 .582	.137 .477	.317 .494	.328 .491
	.125 .533	.224 .490	.138 .568	.108 .561	.342 .539	.296 .509
	.141 .553	.179 .530		.099 .537		.302 .487
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
200	.114 .562	.231 .500	.091 .525	.185 .507	.346 .468	.310 .512
	.112 .607	.273 .499	.101 .469	.106 .481	.367 .470	.334 .478
	.089 .559	.196 .554		.150 .502		.335 .498
Mean	**.105 .576**	**.233 .518**	**.096 .497**	**.147 .497**	**.357 .469**	**.326 .496**
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Table 7

Correlation B/w Bank b's and Part-form b's (r) &
Their Mean Absolute Difference (MAD) :
Form = Mini PP
Sample = Ethnic Group

	Ability Range					
	CAT Sample: Middle[1]		CAT Sample: High[2]		Repre-sentative	
N	r	MAD	r	MAD	r	MAD
100	.625	.570	.660	.427	.577	.462
	.652	.537	.695	.430	.669	.409
					.649	.404
200	.672	.535	.707	.389	.651	.425
	.674	.522	.691	.415	.662	.412
					.640	.455

Table 8

[redacted]

Form = Shorter PP
Sample = Ethnic Group

	Ability Range					
	CAT Sample:		CAT Sample:		Repre-	
N	r	MAD	r	MAD	r	MAD
[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]	[redacted]
	.616	.584	.708	.520	.667	.529