ACT Research Report Series 2000-4

Estimating Item Parameters from Classical Indices for Item Pool Development with a Computerized Classification Test

Chi-Yu Huang

John C. Kalohn

Chuan-Ju Lin

Judith Spray

ACT

March 2000

For additional copies write:
ACT Research Report Series
PO Box 168
Iowa City, Iowa 52243-0168

Estimating Item Parameters from Classical Indices for Item Pool Development with a Computerized Classification Test

Chi-Yu Huang
John C. Kalohn
Chuan-Ju Lin
Judith Spray

Abstract

Item pools supporting computer-based tests are not always completely calibrated. Occasionally, only a small subset of the items in the pool may have actual calibrations, while the remainder of the items may only have classical item statistics, (e.g., *p*-values, point-biserial correlation coefficients, or biserial correlation coefficients). Transformations can be applied to the classical statistics to obtain rough estimates of the item parameters from a 3-parameter logistic IRT model. These estimates, in turn, can be improved by linking them to items with actual calibrations from a program such as *BILOG*. The resulting item-parameter estimates can then be used in a computerized classification test (CCT). An evaluation of the results of using such estimated parameters in simulated CCTs is presented in this paper.

Estimating Item Parameters from Classical Indices for Item Pool Development with a Computerized Classification Test[1]

Moving a testing program from paper/pencil to computerized testing may require that an item pool replace some set of fixed test forms. For many types of computer-based tests (CBTs), an item pool that has been calibrated and scaled to a latent metric is desired. In practice, however, having a complete set of item responses for calibration purposes on all items in the pool may be an unreachable goal for some testing programs. Only one or two recently administered paper-pencil test forms might be calibrated and the rest of the item pool may just consist of classical item parameters such as *p*-values and biserial correlation coefficients for each single item. If these testing programs only require a simple classification decision to be made (e.g., pass/fail), it may be possible to use some methods of approximation when calibrating the item pool and still achieve valid classification results. The purpose of this paper is to describe a procedure which links IRT-calibrated items based on a small portion of an item pool to the remainder of a classically based item pool. The major research question of this study was, "Do these pseudo-calibrations perform as well as actual IRT calibrations obtained from programs such as *BILOG* in one particular CBT application, namely that of a computerized classification test (CCT)?"

[1] Portions of this paper were presented at the 1999 annual meeting of the Psychometric Society in Lawrence, KS. The co-authors of the paper are listed alphabetically.

Description of the Problem

Assume that a 360-item pool for a computerized classification test (CCT) consists of 60 calibrated items from one previously administered paper/pencil test. This set of items will be referred to as the *standard reference set* or SRS. The remainder of the items in the item pool possess their classical item statistics, *p*-values and either point-biserial or biserial correlation coefficients (ρ_{pbs} and ρ_{bs}, respectively, or abbreviated as r and R). The research question to be answered is, "Can item-parameter estimates be obtained on the 300 items that only have classical statistics, and then can these estimates, along with the calibrated SRS, be used to administer a CCT using the sequential probability ratio test (or SPRT) method?" Because the methods described by Urry (1974) and Schmidt (1977) were used to transform an item's classical statistics into estimates of the *a*- and *b*-parameters from the three parameter logistic model (3-PLM), it is helpful to review those procedures.

The Urry-Schmidt Transformations

Urry (1974) proposed that the transformations first described by Birnbaum in Lord and Novick (1968, chapter 16), be corrected for guessing by incorporating a lower bound for the probability of a correct response. Schmidt (1977) refined this method by adjusting for the unreliability of the estimate of the latent trait, θ, in the estimation of R.

Under the assumption that $\theta \sim N(0,1)$, and the free response (i.e., no guessing) items on a test of length n measure the unidimensional trait, θ, and the response functions of those items can each be described by the usual normal ogive response function, P_i or

$$P_i = (2\pi)^{-1/2} \int_{\gamma_i}^{\infty} \exp\left[-\frac{t^2}{2}\right] dt, \qquad (1)$$

where γ_i is the point of cut on the continuous and normal distribution underlying the binary item, then the discrimination parameter, a_i, can be estimated by

$$a_i = \frac{R_i}{\sqrt{1 - R_i^2}}. \qquad (2)$$

The difficulty parameter, b_i, is estimated by

$$b_i = \frac{\gamma_i}{R_i}. \qquad (3)$$

When the items are in a multiple-choice format, the effects of guessing must be incorporated into the estimates given above. Urry (1974) suggested that if c is the usual guessing parameter, so that $P_i^* = c_i + (1-c_i) P_i$, then R_i must be corrected for guessing before the expressions given by equations (2) or (3) above can be used. Urry showed that

$$r_i^* = \frac{(1 - c_i) R_i \phi(\gamma_i)}{\sqrt{(P_i^* Q_i^*)}}, \qquad (4)$$

where $Q^* = 1 - P^*$, and r_i^* is the point-biserial coefficient after correcting for guessing. Then, solving for R_i gives

$$R_i = \frac{r_i^* \sqrt{P_i^* Q_i^*}}{(1 - c_i)\phi(\gamma_i)}. \qquad (5)$$

the estimates for a_i and b_i can then be obtained from equations (2) and (3). The value of c_i can be obtained by using the reciprocal of the number of alternatives in the multiple-choice item format.

Schmidt (1977) suggested that the Urry estimates could be improved by first noting that the procedure tended to systematically underestimate a_i and overestimate b_i. He suggested that the problem lay in the unreliability of the total test score, usually used as an estimate of the latent trait parameter, θ. Schmidt then suggested that the equation for R_i be modified to correct for this unreliability (i.e., correct for attenuation). This gave

$$R_i^* = \frac{r_i^* \sqrt{P_i^* Q_i^*}}{(1 - c_i)\phi(\gamma_i)\sqrt{r_{XX}}}, \qquad (6)$$

where r_{XX} represented the KR-20 reliability (pp. 615-616).

In the present experiment, six test forms of previously calibrated sets, with each form containing 60 items, were used in the experiments to follow. It was assumed, for the purposes of these experiments, that the first 60-item test was the SRS in the item pool, and that the remaining 300 items were not calibrated.

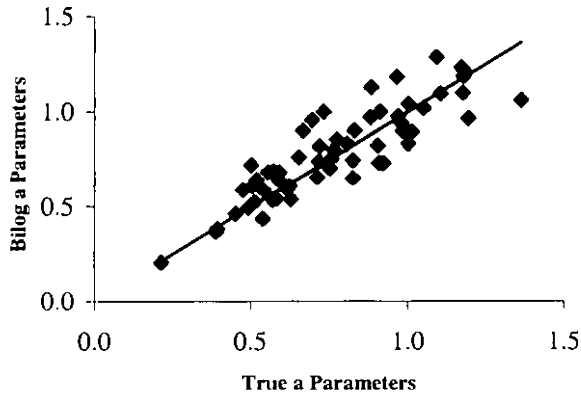
The Dataset

Standard Reference Item Set

The SRS set of item parameters were assumed to be the known or true set and used to generate 0/1 responses for 2000 simulated examinees with θ~N(0,1). The Urry transformations, given above, were then used to obtain estimates of the item parameters; it was assumed that c = .20 for all items. The SRS parameters were also calibrated using the traditional *BILOG* approach

for purposes of comparison. These parameter estimates have been plotted against the true a and b values in Figures 1 and 2, respectively, from *BILOG* as well as from the Urry transformations. The magnitude and direction of bias in the Urry transformations are obvious from these two plots.

***FIGURE 1*: The Comparison of Estimated *BILOG* a and Urry a Parameters**



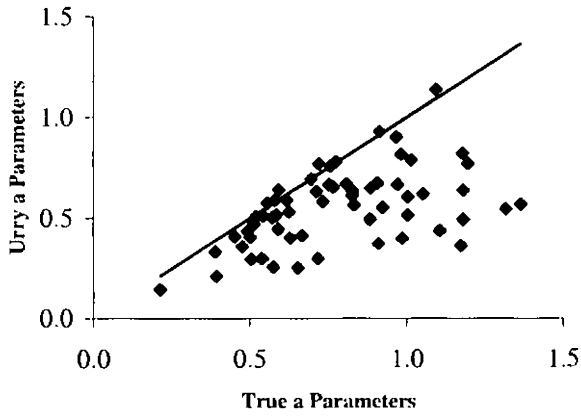
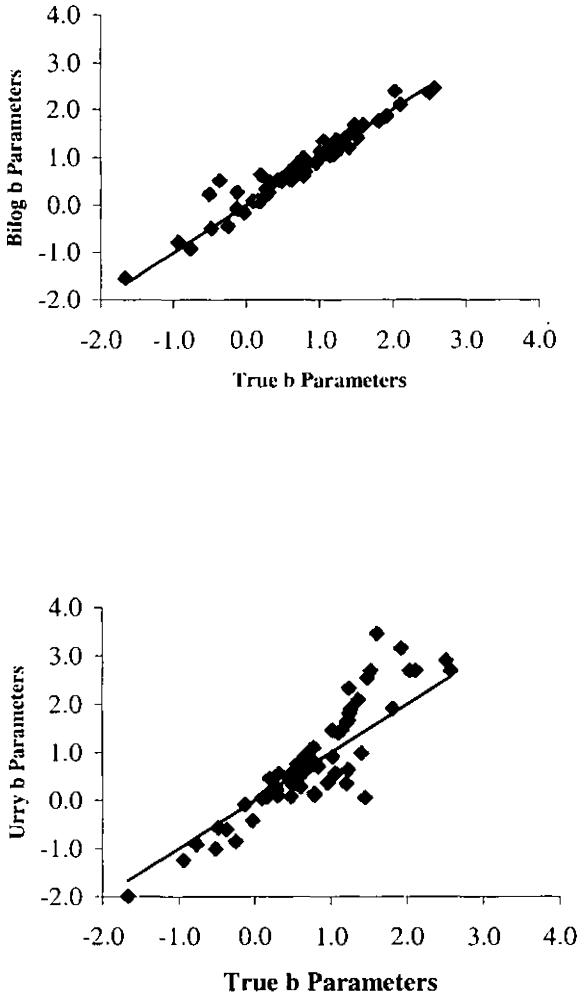


FIGURE 2: **The Comparison of Estimated *BILOG* b and Urry b Parameters**



As predicted, the *a*-parameter estimates were systematically smaller than the true *a*-parameters, while the *b*-parameter estimates slightly overestimated the true difficulty values. By applying the Schmidt correction for unreliability, the amount of bias in the *a*-parameter estimates was somewhat mitigated (see Figure 3). However, there was little effect on the estimation of *b*s, (see Figure 4). Because the Schmidt correction appeared to improve the estimates overall, the Urry estimates with the Schmidt correction for attenuated reliability were used for the remainder of this work. These will be known as the Urry-Schmidt (US) estimates.

FIGURE 3: **The Urry-Schmidt Estimated a Parameters**

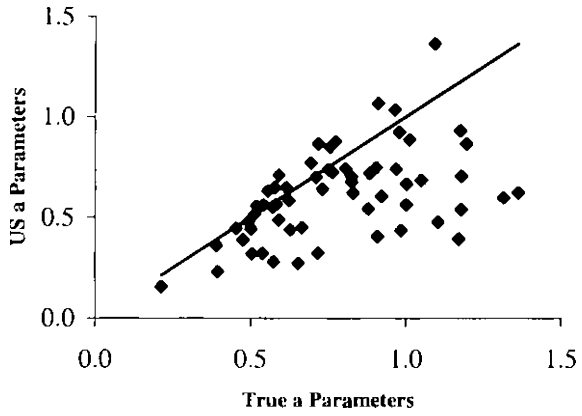
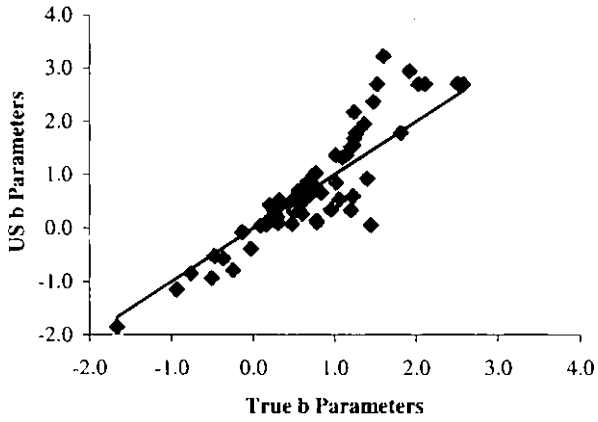


FIGURE 4: **The Urry-Schmidt Estimated b Parameters**



Defining a Linking Transformation

Recall that it was assumed that the item responses to the SRS items existed and could be calibrated. By submitting the generated 0/1 responses to the 60-item SRS from 2000 simulated examinees to the computer program, *BILOG*, it was possible to obtain item parameter estimates for these items (as plotted above in Figures 1 and 2). The item-parameter estimates obtained

from the Urry-Schmidt procedure (i.e., corrected for guessing and for unreliability) were then linked to those obtained from the *BILOG* program.

Four well known linking procedures were considered: (1) mean-mean (MM); (2) mean-sigma (MS); (3) Haebara (HAE); and (4) Stocking and Lord (SL). Each of these linking techniques produced a linear transformation from a θ_2 scale to a θ_1 scale of the general type (Kolen & Brennan, 1995), $\theta_1 = A\theta_2 + B$, where item parameters, a and b, are related by

$$a_1 = \frac{a_2}{A} , \tag{7}$$

$$b_1 = Ab_2 + B , \tag{8}$$

and

$$c_1 = c_2 . \tag{9}$$

Recall that the MM method requires that $A = \mu(a_2)/ \mu(a_1)$ and $B = \mu(b_1) - A\mu(b_2)$, while the MS has $A = \sigma(b_1)/ \sigma(b_2)$ and $B = \mu(b_1) - A\mu(b_2)$. The HAE and SL methods fall under the general category of characteristic curve procedures. The HAE method finds A and B values which minimize the sum of the squared difference of each item characteristic function on the two scales, while the SL procedure finds the values of A and B which minimize the squared difference of the test characteristic functions of the two scales. All procedures are, of course, based on common items (i.e., the SRS of 60 items).

Table 1 shows the results of the four procedures, in terms of the linear transformation coefficients, A and B, that linked the item-parameter estimates from the Urry-Schmidt transformations to the parameter estimate scale produced by *BILOG*.

TABLE 1: Linear Transformation Coefficients

Linking Method	A	B
MM	.764	.206
MS	.714	.244
HAE	.878	.044
SL	.860	.108

When each of the linking transformations given in Table 1 was applied to the Urry-Schmidt item-parameter estimates, new estimates were produced. Each of the linking procedures reduced the bias in the a-parameter estimates over the Urry-Schmidt estimates. Of the four linking procedures, the two characteristic curve methods produced the lowest root mean square errors, as seen in Table 2 below. Both the MM and MS methods produced b-parameter bias that was about the same as that observed from *BILOG* with smaller root mean squared errors than the two characteristic curve methods. The Urry-Schmidt estimates on the SRS are included in Table 2, as a point of comparison.

TABLE 2: Bias and Root Mean Square Error of Estimates: SRS

Estimation Method	Bias(a)	Bias(b)	RMSE(a)	RMSE(b)
BILOG	.023	.056	.124	.204
US	-.166	.033	.290	.502
MM	.023	.056	.273	.367
MS	.080	.056	.296	.355
HAE	-.081	-.017	.264	.416
SL	-.067	.033	.263	.407

The Item Pool

The remaining 300 items in the item pool were characterized by their p-values and biserial correlation coefficients calculated from 0/1 data responses on 60-item tests that were assumed to be parallel to the 60-item SRS. The 0/1 data were generated from simulated populations of 2000 examinees with $\theta \sim N(0,1)$. The classical statistics were transformed to their

Urry-Schmidt item parameter estimates and then linked to the scale from the calibrated items in the pool (i.e., the 60-item SRS) by way of the four linking procedures described above. This yielded four additional item pools that consisted of 360 items; 60 items that had been calibrated using *BILOG* plus 300 items whose classical statistics had been transformed and then linked to the calibrated SRS. The average bias and root mean squared error for all of these item pools are provided in Table 3 below.

TABLE 3: Bias and Root Mean Square Error of Estimates: 360-Item Pool

Estimation Method	Bias(a)	Bias(b)	RMSE(a))	RMSE(b)
BILOG	.006	.037	.125	.162
US	-.172	.109	.335	.471
MM	.028	.133	.300	.318
MS	.078	.133	.319	.309
HAE	-.065	.072	.294	.344
SL	-.052	.113	.292	.346

All of the linking procedures improved upon the estimation of the item pool parameters over the Urry-Schmidt procedure alone. However, the performance of these estimates in an actual CCT simulation was yet to be determined. The next step was to use these different item pools in CCT simulations and compare the results, in terms of classification error rates and test length, to those of a true (i.e., known) pool and those in which the entire pool had been calibrated using *BILOG*.

CCT Simulations

ACT uses the sequential probability ratio test or SPRT procedure to make classification decisions within the framework of a computerized test. The procedure requires the computation of a likelihood ratio of two distinct events (e.g., pass or fail) following the administration of an

item from the pool. When the likelihood ratio becomes greater than some criterion value or less than some other criterion value, testing ceases and the examinee is classified into the appropriate category. In order to compute the likelihood ratio after each item administration, the probability of a correct response (or an incorrect response), given that the examinee has the ability to pass or fail, must be computed.

Item parameter estimates are used to make these calculations from the appropriate IRT model. The item parameter estimates are also used to calculate item information; items are selected for administration based on the amount of information an item has at the passing score. In general, the more informative an item is at the passing score, the greater will be its chances for selection. Item parameter estimates are also used to determine the passing score of the CCT. The process for determining the passing score from an item pool is described below.

Determining the Latent Passing Score

Item parameter estimates are used in the SPRT CCT to determine the latent value associated with the passing score for the test. This passing value is usually denoted as θ_p, where θ_p is the solution to the equation,

$$p = \frac{1}{n}\sum_{i=1}^{n} P\left(U_i = u_i = 1 \middle| \theta_p, \hat{a}_i, \hat{b}_i, \hat{c}_i\right), \qquad (10)$$

p is the passing score in terms of proportion-correct, u_i is the response to item i, and n is the number of items in the reference set used to determine the passing score. If the item parameter estimates are poor, the test may increase either false positive or false negative error rates because of the imprecision in determining the passing point, θ_p.

In the current study, the true latent passing scores were known and corresponded to either 66 % correct when $\theta_p = 1.0$, or 46 % correct, when $\theta_p = 0.\theta$. Depending upon the item parameter estimates in the pool, the value of θ_p may have been different from these true values. A description of each of the pools used in the simulations is provided below, and a table containing the θ_p value for each pool appears in Table 4. Note that the θ_p values for the four linking procedures corresponded to the θ_p used for the *BILOG* item pool. This was due to the fact that the *BILOG*-calibrated SRS was used as the reference or benchmark set for all of the item pools for determining the passing scores.

TABLE 4: Values of the Latent Passing Score

Item Pool	True $\theta_p = 0.0$	True $\theta_p = 1.0$
BILOG	0.06	1.04
US	-0.12	1.09
MM	0.06	1.04
MS	0.06	1.04
HAE	0.06	1.04
SL	0.06	1.04

Item Pools Used in Simulations

The item pools used in the CCT simulations were as follows:

1. Known Item Pool (360 items with known item parameters).
2. *BILOG* Item Pool (360 items with calibrated item parameter estimates).
3. US Item Pool (360 items with US transformed item parameter estimates).
4. MM Item Pool (60 items calibrated with *BILOG* and linked to the 300 item parameter estimates from US transformations using the MM method).
5. MS Item Pool (60 items calibrated with *BILOG* and linked to the 300 item parameter estimates from US transformations using the MS method).
6. HAE Item Pool (60 items calibrated with *BILOG* and linked to the 300 item parameter estimates from US transformations using the HAE method).
7. SL Item Pool (60 items calibrated with *BILOG* and linked to the 300 item parameters estimates from US transformations using the SL method.).

SPRT CTT Simulation Parameters

The CCT simulations were run using the SPRT procedure which requires certain test parameters or conditions to be established. These parameters plus additional information on the simulations included the following: (1) Examinees (i.e., θ) were randomly selected from a N (0,1). There were 100,000 examinees or replications of the SPRT CCT for each set of conditions. (2) There were seven item pools (see descriptions, above) and two passing criteria ($\theta_p = 1.\theta$; $\theta_p = 0.0$). (3) For these simulations, one of four possible content codes was arbitrarily assigned to every 4^{th} item (i.e., in other words, the first item = A; second item = B; third item = C; 4^{th} item = D; 5^{th} item = A, and so on). (4) The size of the indifference region around each

passing score was set at ± .40. (5) The nominal error rates for the test, α and β, were set at .05 each. (6) Test length minimum and maximum were 40 and 60, respectively. (7) The target item exposure control was set at .20, using unconditional Sympson-Hetter. (8) Items were selected based on maximum item information at the passing score.

Simulation Results

The results from the CCT simulations have been summarized in terms of outcomes. The outcomes considered important for evaluation in a CCT include passing and failing rates, false positive and false negative error rates of classification, total classification error rate, average test length and variability of the test length. The results for each of the seven item pools and each of two passing scores have been summarized below in Tables 5 and 6.

TABLE 5: CCT Summary for $\theta_p = 1.0$

Outcome	Known	*BILOG*	US	MM	MS	HAE	SL
Passing rate	.161	.157	.165	.160	.156	.154	.164
Failing rate	.839	.843	.835	.840	.844	.846	.836
False (+) rate	.022	.020	.025	.024	.021	.021	.025
False (–) rate	.020	.022	.020	.020	.022	.024	.020
Total error	.042	.042	.045	.044	.043	.045	.045
Ave length	41.5	41.5	42.2	41.5	41.4	41.8	41.8
SD length	5.0	5.0	6.0	5.0	4.8	5.4	5.4

The *Known* columns of both tables established the sampling error in these simulations because it was known that the expected passing rates for a $\theta \sim N(0,1)$ with $\theta_p = 1.0$ and $\theta_p = 0.0$ should be .159 and .500, respectively. Thus, we could state that the sampling error was between .002 and .005 for simulated error rates.

For the more difficult passing standard of θ_p = 1.0 it was difficult to really detect noticeable differences between the different methods in terms of test length, passing rates, and overall classification errors. The MS and HAE methods underestimated the passing rate, but so did the pool based solely on *BILOG* calibrations. For the easier passing standard of θ_p = 0.0, the results were a bit clearer. The SL method was obviously superior to the other procedures in every outcome category. See Table 6, below.

TABLE 6: CCT Summary for θ_p = 0.0

Outcome	*Known*	*BILOG*	US	MM	MS	HAE	SL
Passing rate	.495	.479	.524	.534	.556	.471	.496
Failing rate	.505	.521	.476	.466	.444	.529	.504
False (+) rate	.046	.040	.064	.071	.083	.037	.047
False (–) rate	.052	.059	.040	.036	.028	.063	.052
Total error	.098	.099	.103	.107	.111	.101	.099
Ave length	45.3	45.2	46.1	44.8	44.6	44.8	44.9
SD length	8.4	8.3	8.8	8.1	8.0	8.1	8.2

Effect of a Smaller Sample Size

The above results were based on fairly large samples of examinees. Recall that for all data generation, 2,000 values of θ were generated. When the sample size[2] was reduced to 500 and the entire study replicated, the results were as follows.

[2] The sample size of 500 was used only to generate the 0/1 response data to compute the classical statistics. The *BILOG* sample on which the original calibrations were obtained remained at 2,000.

TABLE 7: CCT Summary for $\theta_p = 1.0$ with a Sample Size = 500

Outcome	Known	*BILOG*	US	MM	MS	HAE	SL
Passing rate	.161	.157	.165	.162	.159	.155	.163
Failing rate	.839	.843	.835	.838	.841	.845	.837
False (+) rate	.022	.020	.025	.025	.022	.021	.026
False (–) rate	.020	.022	.020	.022	.023	.023	.021
Total error	.042	.042	.045	.047	.045	.044	.047
Ave length	41.5	41.5	42.2	41.6	41.5	41.8	41.8
SD length	5.0	5.0	6.0	5.1	4.9	5.4	5.5

TABLE 8: CCT Summary for $\theta_p = 0.0$ with a Sample Size = 500

Outcome	Known	*BILOG*	US	MM	MS	HAE	SL
Passing rate	.495	.479	.523	.536	.556	.473	.494
Failing rate	.505	.521	.477	.464	.444	.527	.506
False (+) rate	.046	.040	.064	.070	.084	.037	.049
False (–) rate	.052	.059	.041	.036	.030	.063	.052
Total error	.098	.099	.105	.106	.114	.100	.101
Ave length	45.3	45.2	46.3	44.9	44.8	44.9	45.1
SD length	8.4	8.3	8.8	8.1	8.1	8.1	8.3

These results were almost identical to those achieved on the large sample size and spoke to the stability of the transformations and the linking process.

Effect of the Change of c-Parameters

Because the *c*-parameter estimate in the US transformations is an artificial value, usually defined as 1/ (the number of alternatives), it is interesting to know whether using the average *c*-parameter estimate from the calibrated SRS would improve the US approximations and positively affect the outcome of the CCT simulation. To examine this effect, the average *c*-parameter estimate of .242 was used as the *c*-value in the US transformation.

TABLE 9: CCT Summary for $\theta_p = 0.0$ when c= .242

Outcome	Known	*BILOG*	US	MM	MS	HAE	SL
Passing rate	.495	.479	.537	.464	.511	.482	.514
Failing rate	.505	.521	.463	.536	.489	.518	.486
False (+) rate	.046	.040	.077	.034	.056	.039	.059
False (–) rate	.052	.059	.033	.070	.045	.062	.045
Total error	.098	.099	.109	.103	.102	.101	.103
Ave length	45.3	45.2	46.6	44.6	44.3	44.9	45.0
SD length	8.4	8.3	9.0	8.0	7.8	8.2	8.2

Table 9 presents the CCT simulation results for passing standard of $\theta_p = 0.0$ when c = .242. Using the average *c*-parameter estimate in the US transformations did not show evidence of improved CCT results over the use of a fixed constant. However, the linking procedures did provide improvement over the US approximations alone.

Conclusions

As mentioned previously, there are three occasions in CCT where the quality of the item parameter estimates might affect the results of the test: (1) in the determination of the latent passing score for the test; (2) in the selection of items to be administered to each examinee; and (3) in the scoring of the test. Table 4 showed how the errors in parameter estimation lead to different passing score values of θ_p. All methods overestimated the difficulty of the passing standard except for the Urry-Schmidt transformations when $\theta_p = 0.0$. In general a more difficult passing score or standard should result in fewer examinees passing the test.

Item Selection

In terms of item selection, it was of interest to examine how the items within a pool ranked, in terms of their item information, at the passing score. Recall that for CCT, the most

informative items at the passing score are selected for possible administration (dependent, of course, on content specifications and item exposure rates). All of the items in each pool were rank-ordered on this criterion and the ranks were then correlated with those from the *Known* pool.

The results for θ_p = 1.0 showed that, not unexpectedly, the *BILOG* item ranks were highly correlated with the item ranks from the *Known* pool while the initial Urry-Schmidt transformations produced a lower correlation. See Table 10 below. The four linking methods, MM, MS, HAE, and SL, produced somewhat higher correlations, and therefore, would have been expected to perform better than the Urry-Schmidt method alone when compared to the *Known* pool.

For θ_p = 0.0, the *BILOG*-item ranks again correlated highly with the item ranks from the *Known* pool. Interestingly, the initial Urry-Schmidt transformations produced a correlation of .82 with the item ranks from the *Known* pool, while the four linking methods did not consistently increase this value. Under either passing standard, the five approximation methods (Urry-Schmidt plus the four linking procedures) basically ranked pool items at the passing score about the same. The inter-method correlations on item ranks ranged from .938 to .999 (see Tables 10 and 11 below for the θ_p = 1.0 and θ_p = 0.0 conditions, respectively).

TABLE 10: Pool Correlations on Item Ranks for θ_p = 1.0

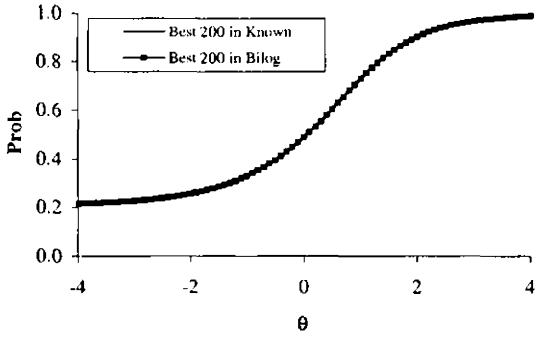
	Known	*BILOG*	US	MM	MS	HAE
BILOG	.956					
US	.688	.724				
MM	.734	.775	.944			
MS	.736	.776	.941	.997		
HAE	.742	.784	.938	.989	.978	
SL	.731	.773	.941	.993	.982	.999

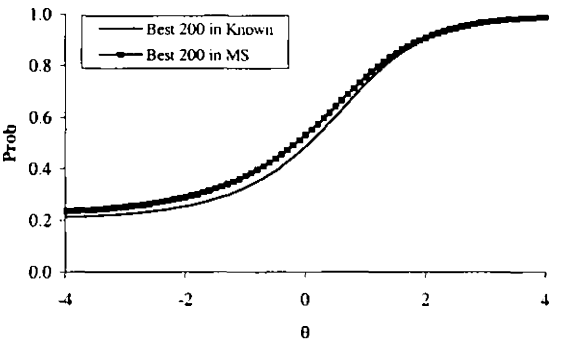
TABLE 11: Pool Correlations on Item Ranks for $\theta_p = 0.0$

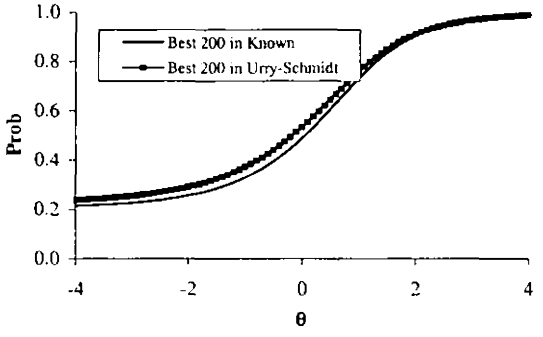
	Known	*BILOG*	US	MM	MS	HAE
BILOG	.959					
US	.817	.850				
MM	.798	.834	.968			
MS	.777	.805	.957	.997		
HAE	.864	.902	.969	.981	.966	
SL	.847	.884	.975	.990	.979	.998

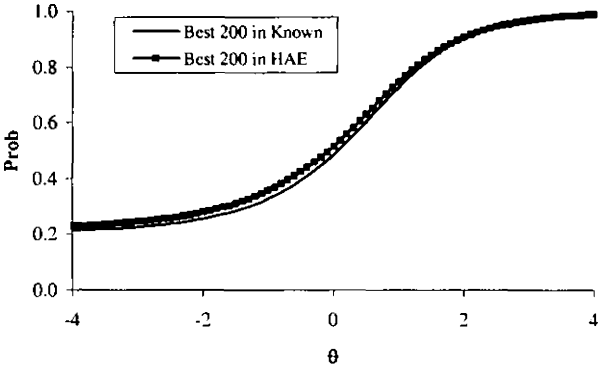
Recall that the item pool size was 360. With a 40-item minimum, mean test lengths of 42-45, and a target item exposure rate of .20, it was estimated that only the best (i.e., most informative at the passing score) 200 items in the pool were being administered on average. In order to study the difficulty level of the tests that were most likely administered, the best 200 items were selected from the *Known* item pool, based on their *true* item information values at the *true* θ_p of 0.0. Then the total characteristic function of the items that were *actually* selected for administration were plotted relative to the total characteristic function of the 200 best items. These plots can be seen in Figure 5 below and indicate that, except for the *BILOG* pool, the set of 200 items that were actually selected for administration were generally easier than those that should have been selected under the *Known* condition.

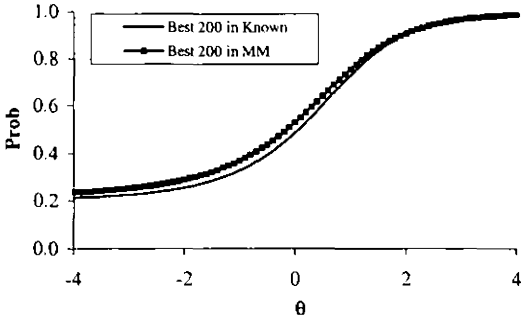
FIGURE 5. **Characteristic Functions for Best 200 Items in Known and Other 6 Pools**

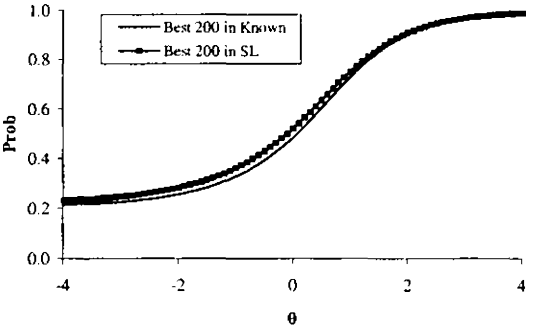












Scoring

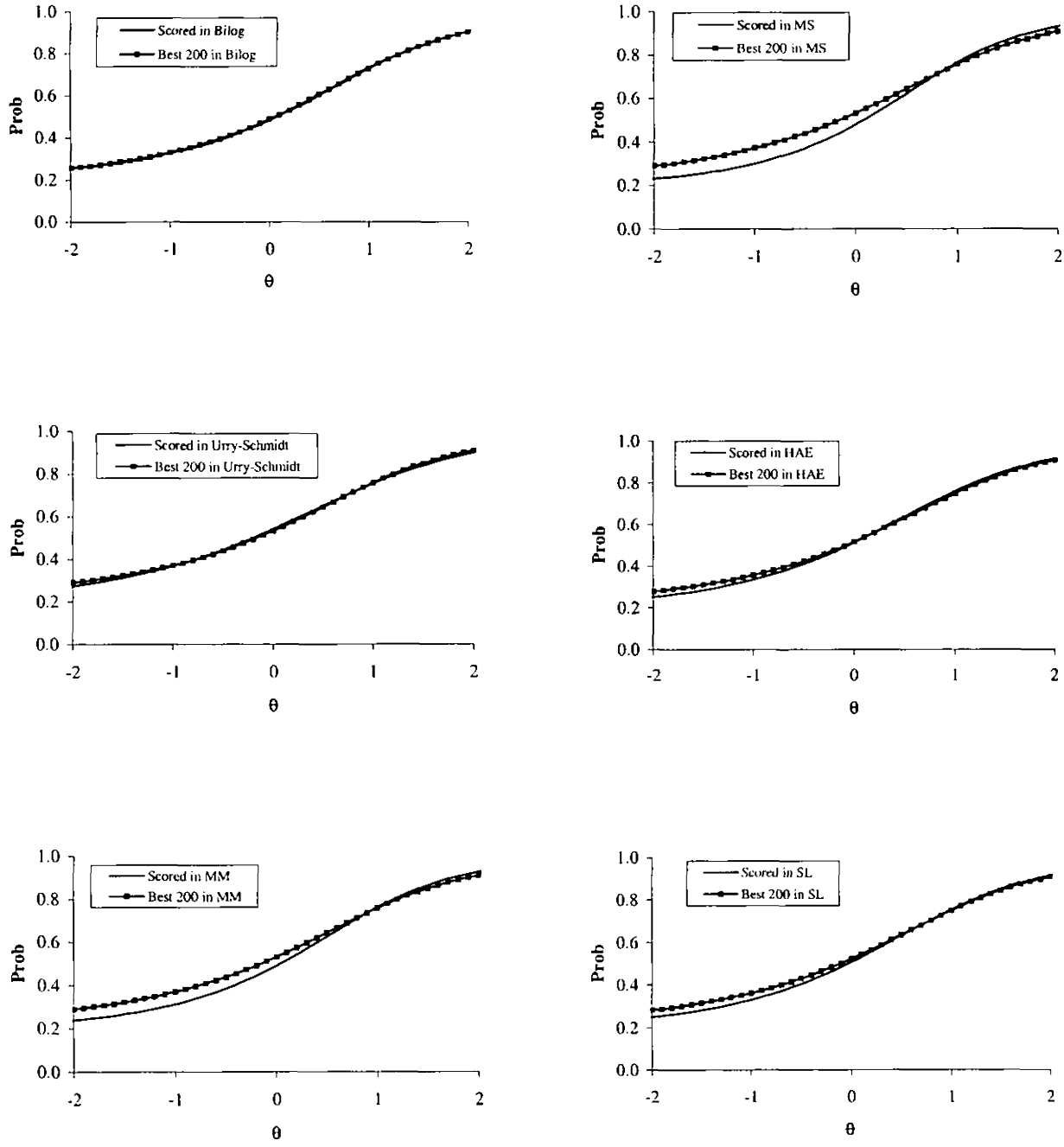
A third source of error from the item parameter estimates affected the length of the tests in a subtle way. To understand the source of this type of error, another traceline was introduced on Figure 6. This traceline represented the **scored** total characteristic function or the impact of the estimated item parameters on the way in which the test was scored. In the SPRT, a likelihood ratio of the form,

$$L = \frac{\pi_1}{\pi_0} , \qquad (11)$$

where π_0 is the binomial probability that the item will be answered given that $\theta = \theta_0$ and π_1, given that $\theta = \theta_1$, is calculated after each item response[3]. The item-parameter estimates are used to calculate the probabilities and, hence, a third source of error is introduced after each item is administered. If the error had little effect, it was expected that the scored traceline and the traceline of the items *actually* selected would have been almost identical.

[3] Here we assume that $\theta_1 > \theta_0$ and that the distance between these two points is the indifference region.

FIGURE 6. **Characteristic Functions for Best 200 Items in 6 Pools**



A steeper slope between these two points on the total characteristic functions at θ_0 and θ_1 implied that, when the test was scored, ***more credit or points*** would be added to the likelihood ratio and the examinee would be classified sooner. Conversely, a more shallow slope implied

that the test would have been scored lower than expected, adding to the length of the test. Table 12, below, provides estimates of the *Known* ratio between endpoints of the indifference region, representing the average amount by which the likelihood ratio was updated following a correct response. Endpoints of the indifference regions were computed at ± .40 around θ_p. The expected ratio was around 1.41. Ratios higher than this would suggest that the procedure scored the test more quickly; those under 1.41, more slowly. This was borne out from observing the average test lengths. The Urry-Schmidt procedure took, on average, 1 to 1.5 items more to complete.

In addition to influencing the length of the test, the ratios in Table 12 also offer an explanation as to why the false positive error rates for the MM and MS methods were inflated. As with all examinees, those near the true passing score were administered easier items than expected and were scored with this higher ratio on average, thus passing at a higher level. See Table 6 for the inflated false positive rates for MM and MS procedures. On the other hand, for HAE and SL, the scoring ratio applied to the examinees was similar in magnitude to what was expected, and inflated false positive error rates were not observed.

TABLE 12: Likelihood Ratio of Correct Responses in 6 Pools

Method	*Known*	Scored
BILOG	1.414	1.415
US	1.409	1.363
MM	1.414	1.517
MS	1.414	1.513
HAE	1.414	1.421
SL	1.414	1.428

Summary

The Urry-Schmidt transformations followed by one of the characteristic curve linking or scaling techniques produced item-parameter estimates that, when used in several CCT situations, resulted in testing outcomes quite close to those expected when the entire pool had been calibrated with a 3-PL model. Although the procedure tended to produce an easier set of items for administration, this bias was somewhat mitigated or offset by a higher-than-expected, estimated passing score.

It is suggested that when considering the use of such augmented parameter estimates as those produced by the linking procedures discussed in this paper, consideration be given to the three sources of possible error and their effects on the outcomes of the test. In addition to augmenting an existing item pool, these techniques may offer initial solutions to the calibration of CBT pretest items as well.

References

Kalohn, J. C., & Spray, J. A. (1999). The effect of model misspecification on classification decisions made using a computerized test. *Journal of Educational Measurement, 36,* 47-59.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: methods and practice.* Springer: New York.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Addison-Wesley: Reading, PA.

Schmidt, F. L. (1977). The Urry method of approximating the item parameters of latent trait theory. *Educational and Psychological Measurement, 37,* 613-620.

Spray, J. A., & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21,* 405-414.

Urry, V. W. (1974). Approximations to item parameters of mental test models and their uses. *Educational and Psychological Measurement, 34,* 253-269.