

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

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 Test¹

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.

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

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

The difficulty parameter, b_i , is estimated by

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

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

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

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

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

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

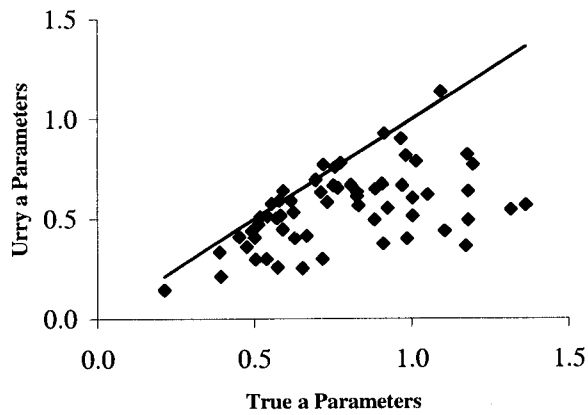
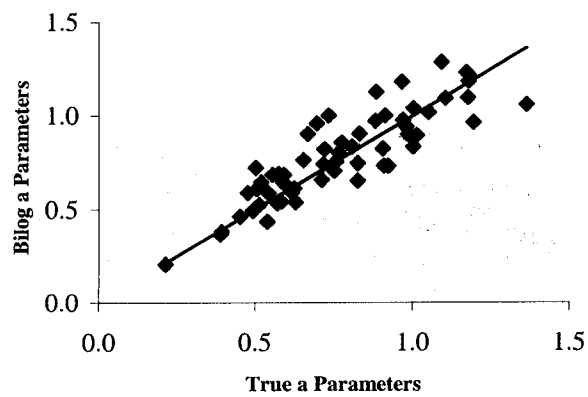


FIGURE 3: The Urry-Schmidt Estimated a Parameters

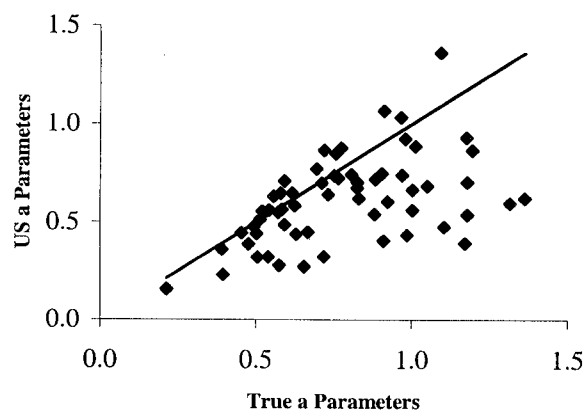
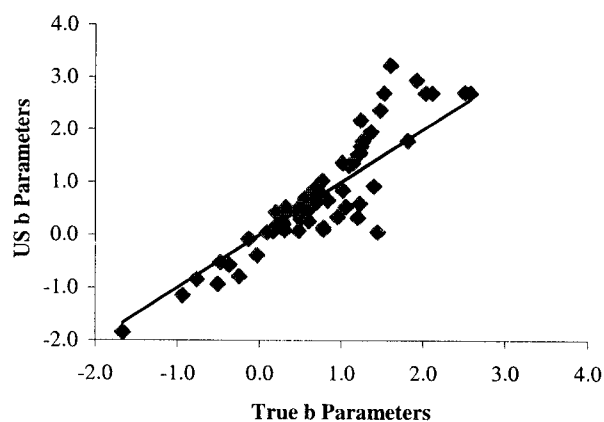


FIGURE 4: The Urry-Schmidt Estimated b Parameters



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

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)
<i>BILOG</i>	.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

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,

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

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 .

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 CCT 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 ($\theta_p = 1.0$; $\theta_p = 0.0$). (3) For these simulations, one of four possible content codes was arbitrarily assigned to every 4th item (i.e., in other words, the first item = A; second item = B; third item = C; 4th item = D; 5th item = A, and so on). (4) The size of the indifference region around each

For the more difficult passing standard of $\theta_p = 1.0$ 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 $\theta_p = 0.0$, 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 $\theta_p = 0.0$

Outcome	<i>Known</i>	<i>BILOG</i>	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 size² 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 9: CCT Summary for $\theta_p = 0.0$ when $c = .242$

Outcome	Known	<i>BILOG</i>	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

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

	<i>Known</i>	<i>BILOG</i>	US	MM	MS	HAE
<i>BILOG</i>	.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.

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,

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

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 response³. 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.

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 $\pm .40$ 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	<i>Known</i>	Scored
<i>BILOG</i>	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

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.