ACT Research Report Series 97-5

Unidimensional Approximations for a Computerized Test When the Item Pool and Latent Space are Multidimensional

Judith A. Spray

Abdel-fattah A. Abdel-fattah

Chi-Yu Huang

C. Allen Lau

ACT

July 1997

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

Unidimensional Approximations for a Computerized Classification Test When the Item Pool and Latent Space Are Multidimensional

Judith A. Spray
Abdel-fattah A. Abdel-fattah
Chi-Yu Huang
ACT

C. Allen Lau
The Psychological Corporation

ABSTRACT

The primary concern or focus of a certification or licensure test is to obtain valid criterion-referenced information regarding a candidate's competency to practice. When the test is administered by computer, a valid pass/fail decision can be made with fewer items than an equivalent paper/pencil test by targeting items at the passing score and using a likelihood ratio approach such as the one utilized in the sequential probability ratio test or SPRT. When administered on a computer, the SPRT is frequently referred to as a computerized classification test or CCT (to distinguish it from the usual computerized adaptive test or CAT). If the CCT is IRT-based, an assumption of unidimensionality is usually required, and the concern is when the item pool is not essentially unidimensional. This study investigated the effects that a multidimensional item pool and latent ability space have on the accuracy of the decisions made using CCT. The results show that the procedure may be fairly robust to such assumption violations.

UNIDIMENSIONAL APPROXIMATIONS FOR A COMPUTERIZED CLASSIFICATION TEST WHEN THE ITEM POOL AND LATENT SPACE ARE MULTIDIMENSIONAL[1]

The primary concern or focus of a certification or licensure test is to obtain valid criterion-referenced information regarding a candidate's competency to practice. When the test is administered by computer, a valid pass/fail decision can be made with fewer items than an equivalent paper/pencil test by targeting items at the passing score and using a likelihood ratio approach such as the one utilized in the sequential probability ratio test or SPRT. When administered on a computer, the SPRT is frequently referred to as a computerized classification test or CCT (to distinguish it from the usual computerized adaptive test or CAT) and a high degree of accuracy of the classification decision can be obtained (Spray & Reckase, 1996). The concern is when the item pool is not essentially unidimensional, because many of the professional certification and licensure examinations are constructed from complex blueprints with content domains, cognitive levels, and practice levels as typical blueprint dimensions. Even when the blueprint consists only of different content categories, they are frequently quite diverse. For example, it is common to see a professional practice blueprint contain very specific categories covering the professional subject matter as well as more general areas such as *Professional Issues* or *Administration.*

When we first began to study the possible effects of a multidimensional item pool on the classification accuracy of CCT using SPRT procedures, we anticipated that we would be able to produce a multidimensional SPRT procedure analogous to the unidimensional one, so that, if it was determined that an item pool was, indeed

multidimensional, we could implement the modified procedure to guarantee minimum classification errors. Having access to multidimensional IRT estimation procedures ensured that we could calibrate a multidimensional item pool and then simply implement the modified SPRT CCT algorithms designed to handle the multidimensionality of the item pool.

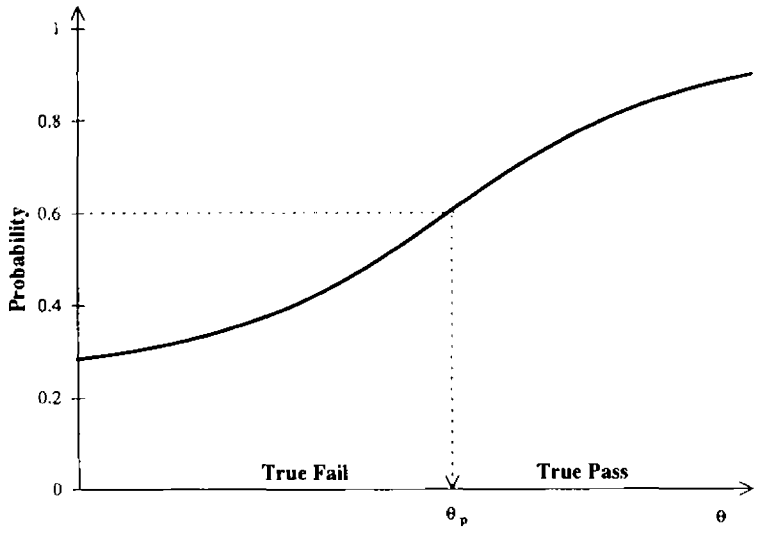
The Unidimensional Case

In the unidimensional case, assume that the item pool fits say, the 3-PL IRT model, $P(Y=1|\theta,a,b,c)$. A passing score is established by using some subset of the item pool called a *standard reference set*. The standard reference set of m items within the item pool mirrors the examination blueprint in terms of major category definitions, proportions of items included within each domain, item difficulty, and other pertinent characteristics. A passing score or passing *rate*, p, is obtained on the standard reference set by some established method (e.g., the Angoff procedure) and an equivalent latent passing score is obtained in the usual way by solving for θ in the relationship,

$$p = \frac{1}{m}\sum_{i=1}^{m} P_i(Y_i = 1|\theta,a_i,b_i,c_i). \qquad (1)$$

The value of θ that satisfies this relationship, θ_p, divides the unidimensional latent space into two mutually exclusive regions: candidates with latent ability, θ_i, who truly are minimally competent ($\theta_i \geq \theta_p$) versus those who are truly not minimally competent ($\theta_i < \theta_p$). See Figure 1.

Figure 1: Unidimensional passing score.



The SPRT CCT procedure selects items from the item pool for administration based or ranked on some statistical or psychometric criterion such as maximum item information at θ_p, after considering the usual selection adjustments involving content constraints, item exposure control, and so on. After each item has been administered to a candidate, the response is used to update a simple likelihood ratio. As soon as the weight of response evidence (i.e., the response string, $y_1, y_2, \ldots, y_n$) clearly supports one decision over the other, testing ceases after the n^{th} item has been administered and the candidate is classified accordingly.

The likelihood ratio, $\mathbf{L}(y_1, y_2, \ldots|\theta_0,\theta_1)$, is computed at two distinct points, θ_0 and θ_1, along the (unidimensional) latent ability space with $\theta_0 < \theta_p < \theta_1$, or

$$\mathbf{L}(y_1,y_2,\ldots|\theta_0,\theta_1) = \frac{\pi_1(\theta_1)\pi_2(\theta_1)\ldots}{\pi_1(\theta_0)\pi_2(\theta_0)\ldots}, \qquad (2)$$

where $\pi_i = P_i^Y(1-P_i)^{1-Y}$, i = 1,2,.... The likelihood ratio is then compared to boundaries *A* and *B* that are functions of the classification error rates, α (false positive) and β (false negative), to be tolerated within the test (Spray & Reckase, 1996; Spray & Reckase, 1987; Wald, 1947). Wald (1947) showed that $A \geq \beta/(1-\alpha)$ and $B \leq (1-\beta)/\alpha$.

If $\mathbf{L}(y_1, y_2, \ldots|\theta_0,\theta_1) \geq A$, the examinee is classified as passing.

If $\mathbf{L}(y_1, y_2, \ldots|\theta_0,\theta_1) \leq B$, the examinee is classified as not passing.

If $B < \mathbf{L}(y_1, y_2, \ldots|\theta_0,\theta_1) < A$, no decision is made and another item response must be observed if a decision is to be made within the specified false positive and false negative error rates.

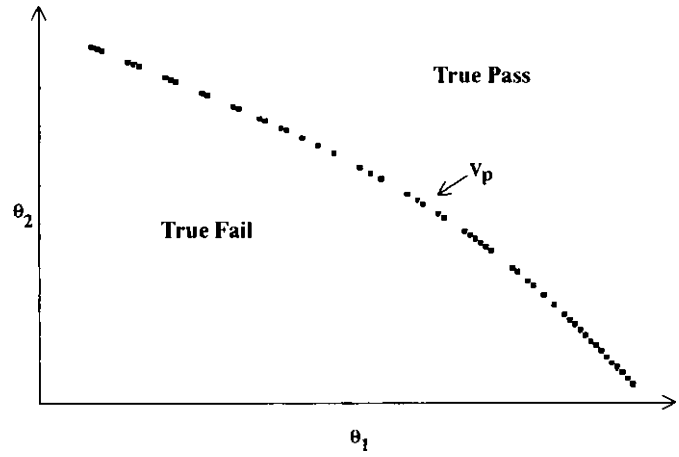
The Multidimensional Case

If the items within the pool are fit by a multidimensional model, say the linear or compensatory 3-PL MIRT model with *q* distinct dimensions and the usual item parameters (a, *d* and *c*) and the passing rate is established as before, then

$$p = \frac{1}{m}\sum_{i=1}^{m} P_i(Y_i = 1|\theta_1,\theta_2,\ldots,\theta_q,\mathbf{a}_i,d_i,c_i). \qquad (3)$$

If we define the function, $F(\theta_1,\theta_2,\ldots,\theta_q) = 1/m\ \Sigma P_i(Y_i=1|\theta_1,\theta_2,\ldots\theta_q,\mathbf{a},d,c)$, then any candidate with latent ability vector $(\theta_1, \theta_2, \ldots, \theta_q)$ that satisfies $F \geq p$, is truly minimally competent[2]. Values of $(\theta_1, \theta_2, \ldots, \theta_q)$ for which $F = p$, abbreviated as $\mathbf{V}_p$, define the curve in the latent space that divides the space into two mutually exclusive regions. See Figure 2 where $q = 2$, for example.

Figure 2: Two-dimensional passing function, Vp.



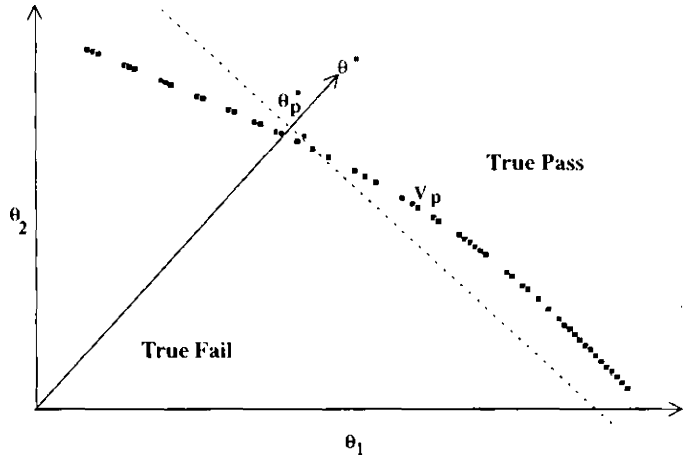
Problems in Extending the SPRT Procedure for a Multidimensional Item Pool

In the unidimensional case, as each item is administered to a candidate, the values of θ_0 and θ_1 and the item parameters can be used in the 3-PL function to yield unique values that are used to update the likelihood ratio. However, if the SPRT procedure were to be extended to the multidimensional situation, then we would need to define the likelihood ratio as before along two distinct curves, $\mathbf{V}_0$ and $\mathbf{V}_1$, approximately parallel to $\mathbf{V}_p$. The values of $\theta_1,\theta_2,...,\theta_q$ that satisfy $\mathbf{V}_0$ and $\mathbf{V}_1$ do not necessarily produce constant probability values for each item. Thus, the likelihood ratio cannot be updated by two unique values, π_0 and π_1, following the administration of each item. In addition there is no single point at which items can be ranked because multidimensional information computed along $\mathbf{V}_p$ varies. Although some approximations are possible, there did not appear to be a straightforward, exact extension of unidimensional SPRT to the multidimensional case.

Unidimensional Approximations to the Multidimensional Case

When an exact SPRT extension did not appear to be feasible, we began to investigate how estimated unidimensional item calibrations in an estimated unidimensional latent space would perform on multidimensional items in a multidimensional latent space. It was obvious that the degree of unidimensional robustness would depend almost entirely on two events: (1) the degree to which a unidimensional reference composite, θ^*, as estimated from IRT calibration software such as *BILOG*, would align with $\mathbf{V}_p$ in the true, multidimensional latent space in such a way as to minimize classification errors and (2) whether examinees with true, multidimensional ability vectors, $\mathbf{V}_i$ and thus with some true classification status relative to $\mathbf{V}_p$, would be sorted correctly along θ^* relative to the (unidimensionally approximated) latent passing score. See Figure 3.

Figure 3: Passing score contours and the unidimensional reference composite.



Stated another way, the accuracy of the unidimensional approximation would depend upon (1) the degree to which θ^* was perpendicular to $\mathbf{V}_p$ and (2) how close the

jth candidate's unidimensional latent approximation, θ_j^* came to sorting that candidate with true ability vector, $(\theta_{1j}, \theta_{2j}, ..., \theta_{qj})$, correctly.

There are several factors which may impact on these two events. These include the degree or strength of the multidimensionality of an item and the ability of the examinee (and, thus, the interaction of the two) (Reckase, 1990), the degree of curvalinearity of $\mathbf{V}_p$ projected onto the latent space, and the discrimination of the item to sort candidates into the correct category. Based on these three factors, we set up conditions for SPRT CCT simulations. For these simulations we examined (1) the degree of multidimensionality of the item pool relative to the multidimensional latent space of the examinees; and (2) different values of the passing proportion, p. The primary question to be answered by the simulations was, *Are there certain conditions under which the unidimensional approximation to the multidimensional SPRT will still yield tolerable CCT errors?*

Method

An item pool was created from the multidimensional calibration of six forms of the ACT Assessment Mathematics Test[3] using the computer program, *NOHARM*[4]. There are 60 items on each form of the ACT Assessment Mathematics Test. Thus, the item pool was composed of six parallel test forms and totalled 360 items. Previous multidimensional analyses of this mathematics achievement test had identified two, distinct dimensions for each form of the test, and all items in the pool had been fit to a linear, multidimensional IRT model or *MIRT* model of the form,

$$P(Y=1|\theta_1,\theta_2,a_1,a_2,d,c) = c+(1-c)\ \phi[d+\mathbf{a}'\theta], \qquad (4)$$

where $P(Y=1|\theta_1,\theta_2,a_1,a_2,d,c)$ is the probability in a two-dimensional space of a correct response, a_1 and a_2 are item discrimination parameters, d is a scaler parameter related to item difficulty, c is a lower asymptote and ϕ is the normal distribution function. The resulting item parameter estimates were used as known (true) parameters for the remainder of the study. The means of these estimates were as follows:

$$\bar{a}_1 = .9320$$

$$\bar{a}_2 = .6383$$

$$\bar{d} = -.7946$$

$$\bar{c} = .1786$$

In order to simulate different levels or degrees of multidimensionality, the value of a_2 was premultiplied by a constant. For dimensionality condition ***A***, $a_2 = (1.5) \bullet a_2$; for dimensionality condition ***B***, $a_2 = (1.0) \bullet a_2$ (i.e., the original set of item parameters); for dimensionality condition ***C***, $a_2 = (.5) \bullet a_2$. Thus, the ***A*** set was theoretically more multidimensional than ***B***, and ***B*** was more multidimensional than ***C***. In addition to the strength of the second dimension of an item, the correlation (ρ) between θ_1 and θ_2 was considered (Reckase, 1990) and was either .00 or .50. These different sets of pool manipulations have been labeled A_0, B_0, C_0, and $A_{.5}$, $B_{.5}$, $C_{.5}$.

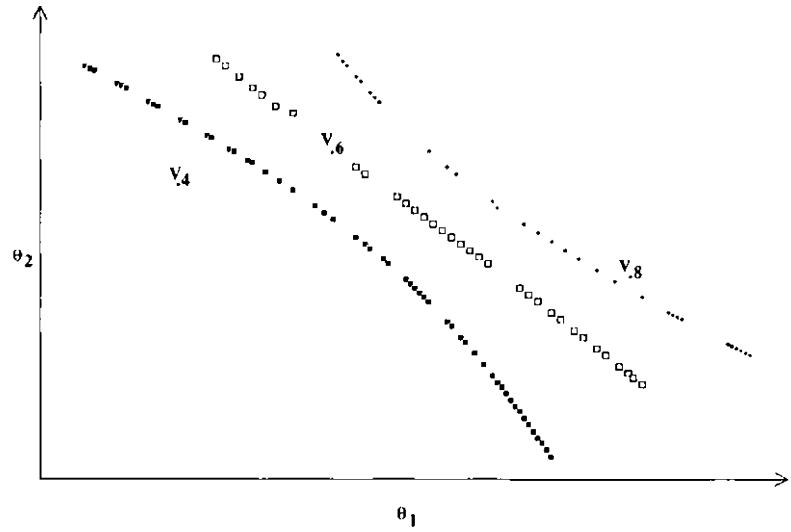
The effect that the passing level, p, has on classification accuracy was also of concern. Because the contour of the passing score function,

$$F(\theta_1,\theta_2,\ldots,\theta_q) = 1/m \; \Sigma P(Y=1|\theta_1,\theta_2,\ldots\theta_q,\mathbf{a},d,c) = p$$

on the two-dimensional latent space, (θ_1,θ_2) is not a straight line, the amount of curvature and the position of the contour line relative to the unidimensional passing score would

certainly affect classification accuracy. Three values of p were used in the simulations: .4 (the lowest passing standard), .6, and .8 (the highest passing standard). See Figure 4, for example, where the three contours from the original item pool ($\boldsymbol{B}$) have been projected onto the two-dimensional latent space.

Figure 4: Three different passing functions.



Responses to all 360 items in the pool were then generated for 2,000 examinees drawn at random for each of these seven parameter sets according to the appropriate MIRT parameters and bivariate density for θ_1, θ_2. The latter was assumed to be bivariate normal with mean vector, $\mathbf{0}$, unit variances and covariance equal to ρ. Each set of 0/1 responses (2,000 by 360) was calibrated with the unidimensional computer program, *BILOG*.

For each set of unidimensional item calibrations, the (unidimensional) pool passing score was obtained by the method described previously (i.e., the unidimensionally estimated proportion-correct true score, $1/n \sum P(\theta^*)$, was set equal to p and then solved

for the passing score, θ_p^*, along the reference composite, θ^*). This value of the latent passing score was used in the SPRT CCT simulations, along with the (unidimensionally estimated) item parameters. For each simulation, 100,000 examinees were sampled from a bivariate normal population described above and were administered the SPRT CCT. The true classification status of each candidate was thus known because the location of $(\theta_{1i},\theta_{2i})$ relative to $\mathbf{V}_p$ was known. The response to each item in the simulation was made using the true, (multidimensional) IRT model. However, the selection of items for administration, the updating of the likelihood ratio, and the evaluation of the stopping point of the test were all determined using the unidimensional approximations.

Because actual CCTs are usually constrained in some way (e.g., by length constraints or exposure or content controls), the simulated CCTs were run under several constraint conditions: (1) the **No constraint** condition where items were administered solely on their rank on $I(\theta_p)$ at θ_p and the test was allowed to terminate according to the SPRT stopping rule; (2) the **Length constraint** condition where a 60-item minimum and a 120-item maximum were imposed on the test (i.e., no decision could be reached until at least 60 items were administered and the test was terminated after the administration of the 120th item)[5]; and (3) the **Length + Exposure constraint**. The latter constraint imposed the minimum and maximum lengths under the previous condition but included a simulated item exposure control in which items were selected for administration by a random draw from a stratum depth of 10 items. In other words, the first item administered to an examinee was selected from the top ten items ranked at θ_p. The second item selected for administration was chosen from the next ten items ranked at θ_p, and so on. This latter condition was thought to emulate actual constraints usually

observed in real CCTs. The nominal error rates for all of the tests were set at $\alpha = \beta = .05$.

The outcome measure obtained from the simulations was the proportion of classification errors made out of 100,000 decisions per simulation. These errors were further divided into false positive and false negative error rates.

Results

Tables 1–3 show the total classification errors, false positive and false negative errors respectively for all item pools, constraint conditions and passing proportions. Longer tests produced more accurate tests, which was expected. The exposure control, in general, had more of an effect for the lowest passing proportion (i.e., the easiest test). Usually item exposure controls produce less accurate results because less-than-optimal items are administered in an effort to control the number of times each item is exposed or administered. This was especially true for a passing proportion of .4 for each item pool but was less so for the remaining passing proportions.

See Tables 1–3 at end of report.

The degree of pool dimensionality made a slight difference. Overall averages for total classification error for pools ***A***, ***B***, and ***C*** in the unconstrained condition were .056, .042, and .042, respectively. However, under the **Length + Exposure constraint** condition, the overall averages by pool were .041, .039, and .039 for ***A***, ***B***, and ***C***. Thus, pool dimensionality made less difference under more realistic testing conditions. Total classification error rates under the most constrained condition (**+ Exposure**) across all

levels of multidimensionality and for all passing proportions averaged .039, which is approximately the error rates that have been observed with unidimensional pools (Spray & Reckase, 1987).

References

Abdel-fattah, A. A., Lau, C. A., & Spray, J. A. (1995, June). *The effect of model misspecification on classification decisions made using a computerized test: UIRT versus MIRT.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis, MN.

Abdel-fattah, A. A., Lau, C. A., & Spray, J. A. (1996, April). *Effect of altering passing score in CAT when unidimensionality is violated.* Paper presented at the American Educational Research Association annual meeting, New York, NY.

Lau, C. A. (1996). *Robustness of a unidimensional computerized mastery testing procedure with multidimensional testing data.* Unpublished doctoral dissertation, University of Iowa, Iowa City, IA.

Lau, C. A., Abdel-fattah, A. A., & Spray, J. A. (1996, April). *Using unidimensional IRT models for dichotomous classification via CAT with multidimensional data.* Poster session presented at American Educational Research Association annual meeting, New York, NY.

Reckase, M. D. (1990, April). *Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests.* Paper presented at the American Educational Research Association annual meeting, Boston, MA.

Spray, J. A., & Reckase, M. D. (1987). *The effect of item parameter estimation error on the decisions made using the sequential probability ratio test.* (Research Report. No. ONR87–1). Iowa City, IA: American College Testing.

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.

Wald, A. (1947). *Sequential analysis.* New York: Wiley.

Wang, M. (1986). *Fitting a unidimensional model to multidimensional item response data: The effects of latent space misspecification on the application of IRT.* Unpublished ONR research report and presentation. ONR Contractor's Annual Meeting, Gatlinburg, TN.

Table 1
Total Classification Error

MIRT Dimensionality	Constraints	Passing Proportion .4	.6	.8
A_0	None	.066	.071	.040
	Length	.063	.058	.032
	+ Exposure	.063	.047	.024
B_0	None	.050	.051	.024
	Length	.049	.037	.019
	+ Exposure	.062	.039	.015
C_0	None	.062	.045	.019
	Length	.061	.032	.014
	+ Exposure	.066	.041	.012
A_5	None	.058	.059	.041
	Length	.046	.041	.031
	+ Exposure	.050	.036	.023
B_5	None	.050	.054	.036
	Length	.042	.039	.025
	+ Exposure	.059	.037	.020
C_5	None	.052	.046	.028
	Length	.049	.032	.020
	+ Exposure	.061	.039	.015

Table 2
False Positive Classification Error

MIRT Dimensionality	Constraints	Passing Proportion .4	.6	.8
A_0	None	.036	.023	.005
	Length	.034	.020	.006
	+ Exposure	.024	.016	.008
B_0	None	.035	.016	.001
	Length	.037	.012	.001
	+ Exposure	.029	.016	.005
C_0	None	.050	.011	.000
	Length	.052	.011	.001
	+ Exposure	.036	.019	.004
$A_{.5}$	None	.032	.012	.003
	Length	.027	.011	.003
	+ Exposure	.021	.012	.006
$B_{.5}$	None	.019	.020	.002
	Length	.014	.011	.001
	+ Exposure	.015	.014	.006
$C_{.5}$	None	.035	.012	.001
	Length	.037	.008	.000
	+ Exposure	.028	.016	.004

Table 3
False Negative Classification Error

MIRT Dimensionality	Constraints	Passing Proportion .4	.6	.8
A_0	None	.030	.048	.035
	Length	.029	.038	.026
	+ Exposure	.039	.031	.016
B_0	None	.015	.035	.023
	Length	.013	.025	.018
	+ Exposure	.033	.023	.011
C_0	None	.012	.034	.019
	Length	.009	.021	.013
	+ Exposure	.030	.023	.008
$A_{.5}$	None	.027	.047	.038
	Length	.020	.030	.028
	+ Exposure	.029	.024	.017
$B_{.5}$	None	.031	.034	.034
	Length	.028	.028	.023
	+ Exposure	.044	.023	.014
$C_{.5}$	None	.017	.035	.028
	Length	.013	.024	.019
	+ Exposure	.033	.024	.010

Endnotes

1. An earlier version of this paper was presented at the 1997 AERA Annual Meeting in Chicago.

2. The degree of dimensionality of the item pool and of the latent space need not be identical. However, they were identical for this study.

3. We used the ACT Math test because it provided a fairly well established, two-dimensional item pool and because the entire item pool was directly proportional to the test blueprint, in terms of content areas and proportions of items (i.e., the pool consisted of six parallel forms).

4. The lower asymptote was estimated using the unidimensional program, *BILOG*.

5. If a maximum number of items has been administered and no decision has yet been reached, the CCT is terminated by forcing the classification according to some distance rule from a likelihood boundary. See Spray and Reckase (1996).