A Bayesian/IRT Index of Objective Performance for Tests with Mixed Item Types[1]

Wendy M. Yen
Robert C. Sykes
Kyoko Ito
Marc Julian

CTB/McGraw-Hill

This paper was presented at the Annual Meeting of the
National Council on Measurement in Education in
Chicago, March 1997

[1]In order to provide more ready access to this unpublished paper, it is reprinted here in the form in which it was originally presented in 1997.

Introduction

Scores for subsets of items in achievement tests are frequently of interest to teachers in evaluating their students' strengths and weaknesses. These subsets of items traditionally have been called objectives. Although currently these subsets of items may more often reflect strands or outcomes, the term "objective" will be retained for this paper.

The standard errors of objective scores can be used to construct confidence intervals that will permit an assessment of the degree of uncertainty associated with an individual score or a classification of a student based upon the score. If the number of items in an objective is small, the standard errors may be so large that a meaningful interpretation of the score is not possible.

The increasingly common presence of constructed response (c.r.) items in educational assessments has resulted in a reduction of the number of items in an examination, given in a fixed testing period, compared to the number of items in an exclusively selected response (s.r.) or multiple choice test. Consequently tests with mixed item types or those composed of exclusively c.r. items may have fewer items per objective than more traditional tests composed of solely s.r. items.

On the other hand the total number of points per objective may be as great or greater than that for objectives consisting solely of s.r. items. Also, because c.r. items are not affected by guessing, they can contain more information than s.r. items. Thus, it is not known whether the scores estimated for objectives from mixed-item type tests will demonstrate less or greater precision or stability than those scores from traditional s.r. tests.

The purpose of this paper is to develop an objective score that will be sufficiently accurate to be useful to teachers when objectives and tests contain one or more item type. An important second goal is to establish the standard error of this objective score. The standard error is intended to produce a realistic estimate of the accuracy of scores for objectives from tests that vary in their proportions of s.r. and c.r. items. The procedure for obtaining these objective scores is developed by generalizing a procedure that has been successfully used to produce more stable scores for objectives in exclusively s.r. examinations (Yen, 1987).

Derivation of the Procedure

The Objective Performance Index (OPI) is an estimated true score (estimated proportion of total

or maximum points possible) for the items in an objective based on the performance of a given examinee. Assume a *k*-item test composed of J objectives with a maximum possible raw score of n. Assume further that each item contributes to at most one objective and the k_j items in objective j contribute a maximum of n_j points. Define X_j as the observed raw score on objective j. The true score (percent of maximum points possible) for objective *j* is:

$$T_j \equiv E(X_j / n_j).$$

It is assumed that there is information available about the examinee in addition to the objective score, and this information provides a prior distribution for T_j. This prior distribution of T_j for a given examinee is assumed to be $\beta(r_j, s_j)$:

$$g(T_j) = \frac{(r_j + s_j - 1)! T_j^{\,r_j - 1} (1 - T_j)^{s_j - 1}}{(r_j - 1)!(s_j - 1)!} \qquad (1)$$

for $0 \le T_j \le 1; r_j, s_j > 0$. Estimates of r_j and s_j are derived from Item Response Theory (IRT; Lord, 1980).

It is assumed that X_j follows a binomial distribution, given T_j:

$$p(X_j = x_j \,|\, T_j) = Binomial(n_j, T_j = \sum_{i=1}^{k_j} T_{ij} / n_j),$$

where T_{ij} is the expected value of the raw score for item i in objective j for a given θ.

Given these assumptions the posterior distribution of T_j, given x_j, is:

$$g(T_j \,|\, X_j = x_j) = \beta(p_j, q_j), \qquad (2)$$

with

$$p_j = r_j + x_j \qquad (3)$$

and

$$q_j = s_j + n_j - x_j. \qquad (4)$$

The OPI is defined to be the mean of this posterior distribution:

$$\tilde{T}_j = \frac{p_j}{p_j + q_j}.$$

Following Novick and Jackson (1974, p.119), a mastery band is created to be the C% central

credibility interval for T_j. It is obtained by identifying the values that place $\frac{1}{2}$(100 - C)% of the $\beta(p_j, q_j)$ density in each tail of the distribution.

Estimation of the Prior Distribution of T_j

The k items in each test are scaled together using a generalized IRT model (3pl/2ppc) that fits a three parameter logistic model (3pl) to the s.r. items and a generalized partial credit model (2ppc) to the c.r. items (Yen, 1993).

The 3pl model is:

$$P_i(\theta) = P(X_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7A_i(\theta - B_i)]} \tag{5}$$

where A_i is the discrimination, B_i is the location, and c_i is the guessing parameter for item i.

A generalization of Master's (1982) Partial Credit model (two-parameter partial credit model; 2ppc) was used for the c.r. items. For a c.r. item with l_i score levels assigned integer scores that range from 0 to $l_i - 1$:

$$P_{im}(\theta) = P(X_i = m - 1|\theta) = \frac{\exp(z_{im})}{\sum_{g=1}^{l_i} \exp(z_{ig})},\ m = 1,...l_i \tag{6}$$

where

$$z_{ig} = \alpha_i(m-1)\theta - \sum_{h=0}^{m-1} \gamma_{ih},$$

and $\gamma_{i0} = 0$. α_i is the item discrimination. γ_{ih} is related to the difficulty of the item levels: the trace lines for adjacent score levels intersect at γ_{ih}/α_i. The 2ppc model, the same as Muraki's (1992) "generalized partial credit model," has been shown to fit response data obtained from a wide variety of mixed-item type achievement tests (Fitzpatrick, Link, Yen, Burket, Ito, & Sykes, 1996).

Item parameters are estimated from a large sample and held fixed in subsequent scoring to obtain OPI values. $T_{ij}(\theta)$ is the expected score for item i in objective j and θ is the common trait value to which the items are scaled:

$$T_{ij}(\theta) = \sum_{m=1}^{l_i} (m-1) P_{ijm}(\theta).$$

T_j, the expected proportion of maximum score for objective j, is:

$$T_j = \frac{1}{n_j}\left[\sum_{i=1}^{k_j} T_{ij}(\theta)\right].$$

The expected score for item i and estimated proportion of maximum score for objective j are obtained by substituting the estimate of the trait $(\hat{\theta})$ for the true trait value.

The theoretical random variation in item response vectors and resulting $\hat{\theta}$ values for a given examinee produces the distribution $g(\hat{T}_j | \hat{\theta})$ with mean $\mu(\hat{T}_j | \theta)$ and variance $\sigma^2(\hat{T}_j | \theta)$. This distribution is used to estimate a prior distribution for T_j.

Given that T_j is assumed to be distributed as a Beta variable (equation 1), the mean $[\mu(\hat{T}_j | \theta)]$ and variance $[\sigma^2(\hat{T}_j | \theta)]$ of this distribution can be expressed in terms of its parameters, r_j and s_j. Using IRT, the mean and variance of the prior can also be expressed in terms of functions of item parameters. Thus, the parameters of the prior distribution, and therefore the posterior distribution, can be expressed as functions of the IRT parameters.

Expressing the mean and variance of the prior distribution in terms of the parameters of the beta distribution produces (Novick & Jackson, 1974, p. 113):

$$\mu(\hat{T}_j | \theta) = \frac{r_j}{r_j + s_j}$$

and

$$\sigma^2(\hat{T}_j | \theta) = \frac{r_j s_j}{(r_j + s)^2 (r_j + s_j + 1)} \ .$$

Solving these equations for r_j and s_j produces

$$r = \mu(\hat{T}_j | \theta) n_j^*$$

and

$$s_j = [1 - \mu(\hat{T}_j | \theta)] n_j^*,$$

where

$$n_j^* = \frac{\mu(\hat{T}_j \mid \theta)[1-\mu(\hat{T}_j \mid \theta)]}{\sigma^2(\hat{T}_j \mid \theta)} - 1.$$

Using IRT, $\sigma^2(\hat{T}_j \mid \theta)$ can be expressed in terms of item parameters. From Lord (1983):

$$\mu(\hat{T}_j \mid \theta) \approx \frac{1}{n_j} \sum_{i=1}^{k_j} \hat{T}_{ij}(\theta).$$

Because T_j is a monotonic transformation of θ, and from Lord (1980, p.71):

$$\sigma^2(\hat{T}_j \mid \theta) = \sigma^2(\hat{T}_j \mid T_j) \approx I(T_j, \hat{T}_j)^{-1},$$

where $I(T_j, \hat{T}_j)$ is the information that $\hat{T}_j$ contributes about T_j. Given these results and those of Lord (1980, p.79 and p. 85) produces:

$$I(T_j, \hat{T}_j) = I(\theta, \hat{T}_j) / \left[\partial T_j / \partial \theta\right]^2,$$

and

$$I(\theta, \hat{T}_j) \approx I(\theta, \hat{\theta}). \tag{7}$$

Thus,

$$\sigma^2(\hat{T}_j \mid \theta) \approx \frac{\left[\frac{1}{n_j} \sum_{i=1}^{k_j} \hat{T}_{ij}'(\theta)\right]^2}{I(\theta, \hat{\theta})}$$

and the parameters of the prior beta distribution for T_j can be expressed in terms of the parameters of the IRT models. Further, the parameters of the posterior distribution of T_j also can be expressed in terms of the IRT parameters:

$$p_j = \hat{T}_j n_j^* + x_j,$$

and

$$q_j = [1-\hat{T}_j] n_j^* + n_j - x_j.$$

The OPI is

$$\widetilde{T}_j = \frac{p_j}{p_j + q_j}$$

$$= \frac{\hat{T}_j n_j^* + x_j}{n_j^* + n_j}.$$

The OPI can also be written in terms of the relative contribution of the prior estimate $\hat{T}_j$, and the observed proportion of maximum score (OPM), $\frac{x_j}{n_j}$, as:

$$\widetilde{T}_j = w_j \hat{T}_j + (1 - w_j)\frac{x_j}{n_j}.$$

w_j, a function of the mean and variance of the prior distribution, is the relative weight given to the prior estimate:

$$w_j = \frac{n_j^*}{n_j^* + n_j}.$$

The term, n_j^*, may be interpreted as the contribution of the prior in terms of theoretical numbers of score points.

Check on Consistency and Adjustment of Weight Given to Prior

The item responses are assumed to be described by $P_i(\hat{\theta})$ or $P_{im}(\hat{\theta})$, depending on the type of item. Even if the IRT model accurately described item performance over examinees, their item responses grouped by objective may be multidimensional. For example, a particular examinee may be able to perform difficult addition but not easy subtraction. Under these circumstances it is not appropriate to pool the prior estimate, $\hat{T}_j$, with x_j / n_j. In calculating the OPI the following statistic was used to identify examinees with unexpected performance on the objectives in a test:

$$Q = \sum_{j=1}^{J} n_j (\frac{x_j}{n_j} - \hat{T}_j)^2 / (\hat{T}_j(1 - \hat{T}_j)).$$

If $Q \le \chi^2(J,.10)$, the weight, w_j, is computed and the OPI produced. If $Q > \chi^2(J,.10)$, n_j^* and subsequently w_j are set equal to 0 and the OPM is used as the estimate of objective performance.

As previously noted, the prior is estimated using an ability estimate based on responses to all the items (including the items of objective j) and hence is not independent of X_j. An adjustment for the overlapping information that requires minimal computation is to multiply the test information in equation 7 by the factor: $(n - n_j)/n$. The application of this factor produces an "adjusted" OPI estimate that can be compared to the "unadjusted" estimate.

Possible Violations of the Assumptions

First, as described above, the Q statistic is used to evaluate the OPM relative to that predicted for the items in the objective on the basis of the student's overall trait estimate. If the chi-square is significant, the prior estimate is not used and the OPM obtained becomes the student's objective score.

Second, if the items in the objective do not permit guessing, it is reasonable to assume $\hat{T}_j$, the expected proportion of maximum score for an objective, will be greater or equal to zero. If correct guessing is possible, as it is with s.r. items, there will be a non-zero lower limit to $\hat{T}_j$, and a three-parameter beta distribution in which $\hat{T}_j$ is greater than or equal to this lower limit (Johnson & Kotz, 1979, p.37) would be more appropriate. The use of the two-parameter beta distribution would tend to underestimate T_j among very low-scoring examinees. Yen (1987), working with tests containing exclusively s.r. items, found that there does not appear to be a practical importance to this underestimation. The impact of any such effect would be reduced as the proportion of c.r. items in the test increases. The size of this effect is, nonetheless, evaluated with simulations.

Third, the OPI procedure assumes that $p(X_j | T_j)$ is a binomial distribution. This assumption is appropriate only when all the items in an objective have the same Bernoulli item response function. Not only do real items differ in difficulty, but when there are mixed item types X_j is not the sum of n_j independent Bernoulli variables. It is instead the total raw score. In essence the simplifying assumption has been made that each c.r. item with a maximum score of $l_j - 1$ is the sum of $l_j - 1$ independent Bernoulli variables. Thus, a complex compound distribution is theoretically more applicable than the binomial. Given the complexity of working with such a model, it appears

valuable to determine if the simpler model described here is sufficiently accurate to be useful.

Finally, because the prior estimate of $T_j, \hat{T}_j$, is based on performance on the entire test, including objective j, the prior estimate is not independent of X_j. The smaller the ratio n_j / n, the less impact this dependence will have. The effect of the overlapping information would be to understate the width of the credibility interval. The extent to which the size of the credibility interval is too small was examined by simulating objectives that contained varying proportions of the total test points.

Method

Simulations

Performance on mixed-item type examinations that differed in size, difficulty, and number and size of objectives were simulated. There were two types of simulations.

The first type, referred to as "simulations modeled on real tests," employed 3pl/2ppc item parameter estimates for eight actual mixed-item type achievement tests (several of the tests were slightly modified to vary the proportion of items contributed by c.r. items) in four different content areas (Math, Reading, Citizenship, and Science) administered in the primary and secondary grades. Using the existing parameter estimates and objective structures, simulated item responses were generated for 3000 simulees for each examination, assuming a N(0,1) distribution of thetas. Maximum likelihood estimates of thetas were then obtained through pattern scoring of the 3000 simulees based on all the items for each examination using PARDUX (Burket, 1991; 1995). Evaluations of the accuracy of the program with simulated data (Fitzpatrick, 1994) have found it to be at least as accurate as MULTILOG (Thissen, 1986).

Nine other examinations were simulated, with difficulties and proportions of total test points from c.r. items that complemented the eight "real test" simulations. Item parameters were generated for these "hypothetical" examinations by specifying the mean and standard deviation (s.d.) of each item parameter type (e.g. discrimination, first step difficulty, etc.). Simulated parameter values were drawn from populations of item parameter types that were assumed normally distributed. Item responses were then simulated and thetas estimated in the same manner as the real-test-modeled simulations.

The model used to generate the simulated item responses utilized the three-parameter logistic

(3pl) model for the s.r. items and the 2ppc model for the c.r. items.

Cross Validation

Construction of split-half forms. The consistency of the OPI values was investigated using a split-half procedure. Each of the 17 examinations was partitioned into half forms I and II. Items for the half forms were selected in order to attain split-half objectives that were as similar as possible in terms of:

1) the mean and s.d. of objective OPMs
2) the total number of points, and
3) the number of points contributed by c.r. items.

Table 1 contains the number of items (s.r. and c.r.) and points in the half forms and their objectives, as well as the complete forms from which they were selected. Tests are numbered from 1 to 17 in terms of the proportion of the total points constituted by their c.r. items. Table 2 displays the difficulty of the complete forms and their proportion of c.r. points. Note that Tests #9 through #17 are the hypothetic examinations.

Table 1
Number of Items and Points by Objective for Whole and Half Tests

							Half Tests							
							I				II			
Test	Obj. #	# SR Items	# CR Items	(CR Pt. Range)	Total # Points	Prop CR Points	# SR Items	# CR Items	(CR Pt. Range)	Total # Points	# SR Items	# CR Items	(CR Pt. Range)	Total # Points
1		35	1	(3)	38	0.08	18	0		18	17	1	(3)	20
	1						7	0		7	8	0		8
	2						8	0		8	7	0		7
	3						3	0		3	2	1	(3)	5
2		34	2	(2-4)	40	0.15	17	1	(4)	21	17	1	(2)	19
	1						2	1	(4)	6	3	0		3
	2						4	0		4	3	1	(2)	5
	3						7	0		7	6	0		6
	4						4	0		4	5	0		5
3		30	3	(2-4)	38	0.21	14	2	(2)	18	16	1	(4)	20
	1						2	0		2	3	0		3
	2						2	0		2	4	0		4
	3						2	1	(2)	4	3	0		3
	4						4	0		4	1	1	(4)	5
	5						1	1	(2)	3	3	0		3
	6						3	0		3	2	0		2
4		42	8	(2)	58	0.28	21	4	(2)	29	21	4	(2)	29
	1						5	1	(2)	7	4	2	(2)	8
	2						4	1	(2)	6	5	0		5
	3						4	0		4	6	0		6
	4						5	0		5	3	1	(2)	5
	5						3	2	(2)	7	3	1	(2)	5
5		40	6	(3-5)	61	0.34	20	3	(2-5)	30	20	3	(3-4)	31
	1						5	2	(2-5)	12	5	1	(3)	8
	2						5	0		5	5	1	(4)	9
	3						5	1	(3)	8	5	1	(4)	9
	4						5	0		5	5	0		5
6		30	10	(2-4)	54	0.44	15	5	(2-4)	27	15	5	(2-4)	27
	1						4	1	(2)	6	1	1	(4)	5
	2						3	1	(2)	5	3	1	(2)	5
	3						6	2	(2-4)	12	9	2	(2)	13
	4						2	1	(2)	4	2	1	(2)	4
7		22	8	(2-4)	42	0.48	11	4	(2-4)	21	11	4	(2-4)	21
	1						2	1	(2)	4	1	1	(4)	5
	2						1	1	(2)	3	2	0		2
	3						5	1	(2)	7	6	2	(2)	10
	4						3	1	(4)	7	2	1	(2)	4
8		25	11	(2-4)	51	0.51	13	5	(2-4)	25	12	6	(2-4)	26
	1						5	1	(2)	7	5	1	(4)	9
	2						4	1	(2)	6	4	1	(2)	6
	3						2	2	(2)	6	2	2	(2)	6
	4						2	1	(4)	6	1	2	(2)	5
9		16	12	(1-3)	40	0.60	8	6	(1-3)	20	8	6	(1-3)	20
	1						2	2	(1-3)	6	2	2	(1-3)	6
	2						2	2	(1-3)	6	2	1	(2)	4
	3						2	1	(2)	4	2	1	(2)	4
	4						2	1	(1-3)	4	2	2	(1-3)	6
10		16	12	(1-3)	40	0.60	8	6	(1-3)	20	8	6	(1-3)	20
	1						2	2	(1-3)	6	2	1	(2)	4
	2						2	2	(1-3)	6	2	1	(2)	4
	3						2	1	(2)	4	2	2	(1-3)	6
	4						2	1	(2)	4	2	2	(1-3)	6

(table continued)

Table 1 (cont.)
Number of Items and Points by Objective for Whole and Half Tests

							Half Tests							
							I				II			
Test	Obj. #	# SR Items	# CR Items	(CR Pt. Range)	Total # Points	Prop CR Points	# SR Items	# CR Items	(CR Pt. Range)	Total # Points	# SR Items	# CR Items	(CR Pt. Range)	Total # Points
11		16	12	(1-3)	40	0.60	8	6	(1-3)	20	8	6	(1-3)	20
	1						2	2	(1-3)	6	2	1	(2)	4
	2						2	2	(1-3)	6	2	2	(1-3)	6
	3						2	1	(2)	4	2	1	(2)	4
	4						2	1	(2)	4	2	2	(1-3)	6
12		12	18	(1-3)	48	0.75	6	9	(1-3)	24	6	9	(1-3)	24
	1						2	3	(1-3)	8	2	3	(1-3)	8
	2						2	3	(1-3)	8	2	3	(1-3)	8
	3						2	3	(1-3)	8	2	3	(1-3)	8
13		12	18	(1-3)	48	0.75	6	9	(1-3)	24	6	9	(1-3)	24
	1						2	3	(1-3)	8	2	3	(1-3)	8
	2						2	3	(1-3)	8	2	3	(1-3)	8
	3						2	3	(1-3)	8	2	3	(1-3)	8
14		12	18	(1-3)	48	0.75	6	9	(1-3)	24	6	9	(1-3)	24
	1						2	3	(1-3)	8	2	3	(1-3)	8
	2						2	3	(1-3)	8	2	3	(1-3)	8
	3						2	3	(1-3)	8	2	3	(1-3)	8
15		0	24	(1-3)	48	1.00	0	12	(1-3)	24	0	12	(1-3)	24
	1						0	3	(1-3)	6	0	3	(1-3)	6
	2						0	3	(1-3)	6	0	3	(1-3)	6
	3						0	3	(1-3)	6	0	3	(1-3)	6
	4						0	3	(1-3)	6	0	3	(1-3)	6
16		0	24	(1-3)	48	1.00	0	12	(1-3)	24	0	12	(1-3)	24
	1						0	3	(1-3)	6	0	3	(1-3)	6
	2						0	3	(1-3)	6	0	3	(1-3)	6
	3						0	3	(1-3)	6	0	3	(1-3)	6
	4						0	3	(1-3)	6	0	3	(1-3)	6
17		0	24	(1-3)	48	1.00	0	12	(1-3)	24	0	12	(1-3)	24
	1						0	3	(1-3)	6	0	3	(1-3)	6
	2						0	3	(1-3)	6	0	3	(1-3)	6
	3						0	3	(1-3)	6	0	3	(1-3)	6
	4						0	3	(1-3)	6	0	3	(1-3)	6

Table 2
Means and Standard Deviations of OPM's for the Whole Tests:
Ordered by Proportion of CR Points in the Tests

Test	Total # Points	Prop. CR Points	OPM	
			Mean	s.d.
1	38	0.08	0.57	0.20
2	40	0.15	0.48	0.18
3	38	0.21	0.75	0.17
4	58	0.28	0.63	0.18
5	61	0.34	0.50	0.21
6	54	0.44	0.55	0.16
7	42	0.48	0.69	0.17
8	51	0.51	0.63	0.19
9	40	0.60	0.34	0.25
10	40	0.60	0.37	0.25
11	40	0.60	0.52	0.24
12	48	0.75	0.27	0.24
13	48	0.75	0.35	0.27
14	48	0.75	0.55	0.26
15	48	1.00	0.27	0.30
16	48	1.00	0.41	0.30
17	48	1.00	0.56	0.27

OPI values were calculated for the whole test and each half using the unadjusted and adjusted OPI procedures. Thetas were estimated in the same manner as the complete forms after partitioning the item responses and parameters into the half forms. Because of the importance of each pair of half objectives being approximately equivalent, OPI values were obtained only for those objectives whose split-halves:

1) had at least four points
2) differed by no more than one point in terms of maximum possible score, and
3) had mean OPMs and s.d.'s of OPMs that differed by no more than .05 across halves.

Tests 3 and 10 did not have any objectives meeting these criteria and hence were not included in the split-halves analyses.

All the items in a split-half test were used to estimate ability. The Q statistics were used to determine whether an OPI would be estimated. Since the power of the Q statistics increases as the number of items increases, more significant Q values were expected for the whole tests than the half tests.

Comparison of split-halves. The split-half objectives allowed the comparison of means, standard deviations, and correlations over halves. In addition, two analyses were conducted to examine the predicted posterior distribution of the OPI values.

In the first analysis the proportion of examinees whose 67% credibility interval for a split-half objective overlapped the credibility interval for the companion half objective was determined. Based on the normal approximation to the beta distribution, an expectation of the approximate proportion of overlapping intervals was obtained.

Let Y_1 and Y_2 be two independent normal variables with equal means and with variances equal to σ^2 and $K^2\sigma^2$, respectively, where K>0. It can be shown that the probability that $y_1 \pm .967\sigma$ overlaps $y_2 \pm .967K\sigma$ equals the probability that $|Z| < .967(K+1)/\sqrt{(K^2+1)}$, where Z is a standard normal variable. If the posterior distribution has the same standard deviation for both halves (i.e., $K = 1$), the normal approximation leads to the expectation that 83 percent of the

examinees would have credibility intervals that overlap for the two halves. This expected percent drops as the ratio of the posterior standard deviation varies from 1:1. If the ratio is 1:2, the expected overlap is 81, and if the ratio is 1:5, the expected percent overlap is 74. Given that the posterior standard deviations could vary substantially over halves, the expectation was that the percent or proportion overlap would be in the mid to high 70's for 67% credibility intervals.

The second analysis of the posterior distributions was a comparison of the predicted posterior standard deviation and the observed standard error of T_j. The predicted value for objective j was:

$$\overline{\sigma}(T_j \mid x_j) = \left[\frac{1}{3000} \sum_{k=1}^{3000} \frac{1}{2} \sum_{m=I}^{II} \sigma^2 (T_{jkm} \mid x_{jkm}) \right]^{1/2}$$

where $\sigma^2 (T_{jkm} \mid x_{jkm})$ is obtained from the posterior variance of T_j, given x_j:

$$\sigma^2 (T_j \mid x_j) = \frac{p_j q_j}{(p_j + q_j)^2 (p_j + q_j + 1)}$$

$$\approx \frac{\tilde{T}_j (1 - \tilde{T}_j)}{n_j^* + n_j + 1}$$

for examinee k and test half m. The observed value for objective j was :

$$S.E. = \left[\frac{1}{3000} \sum_{k=1}^{3000} \frac{1}{2} (\tilde{T}_{kjI} - \tilde{T}_{kjII})^2 \right]^{1/2}$$

Results

Whole Tests

Table 3 contains the means and standard deviations of the whole test OPMs (X_j / n_j), priors ($\hat{T}_j$), and OPIs ($\tilde{T}_j$) by objective. The proportion of simulees that had significant Q values is also presented. The latter proportions range between .07 (Tests 1, 4, and 7) and .19 (Test 15). The nominal significance level of .10 for the χ^2 test predicts that 10% of the simulees will be rejected solely by chance in the absence of significant multidimensionality due to objectives. In general, the greater the proportion of c.r. items, the greater the percent rejections.

Table 3
Means and Standard Deviations of OPM's (x_j / n_j),
Priors $(\hat{T}_j)$ and OPI's $(\tilde{T}_j)$for Whole Tests

Test	Prop. Sig. Q	Obj. #	x_j / n_j Mean	x_j / n_j SD	$\hat{T}_j$ Mean	$\hat{T}_j$ SD	$\tilde{T}_j$ Unadjusted Mean	$\tilde{T}_j$ Unadjusted SD	$\tilde{T}_j$ Adjusted Mean	$\tilde{T}_j$ Adjusted SD
1	0.07									
		1	0.64	0.19	0.65	0.20	0.64	0.19	0.64	0.19
		1	0.61	0.23	0.63	0.24	0.61	0.23	0.61	0.23
		3	0.39	0.22	0.42	0.21	0.40	0.18	0.40	0.18
2	0.08									
		1	0.34	0.21	0.35	0.19	0.34	0.18	0.34	0.17
		2	0.52	0.22	0.50	0.21	0.50	0.20	0.50	0.20
		3	0.57	0.20	0.55	0.20	0.55	0.19	0.55	0.19
		4	0.46	0.20	0.47	0.17	0.46	0.16	0.46	0.16
3	0.09									
		1	0.82	0.20	0.83	0.16	0.81	0.16	0.81	0.16
		2	0.77	0.20	0.79	0.17	0.77	0.17	0.77	0.17
		3	0.73	0.22	0.75	0.18	0.73	0.18	0.73	0.18
		4	0.68	0.23	0.70	0.20	0.68	0.20	0.68	0.20
		5	0.72	0.21	0.75	0.16	0.73	0.16	0.73	0.16
		6	0.83	0.20	0.83	0.17	0.82	0.17	0.82	0.17
4	0.07									
		1	0.61	0.21	0.62	0.22	0.61	0.21	0.61	0.21
		2	0.56	0.21	0.57	0.19	0.56	0.19	0.56	0.19
		3	0.65	0.20	0.66	0.18	0.65	0.18	0.65	0.18
		4	0.61	0.22	0.61	0.20	0.60	0.20	0.60	0.20
		5	0.69	0.17	0.69	0.17	0.68	0.17	0.68	0.17
5	0.09									
		1	0.48	0.22	0.48	0.22	0.46	0.21	0.46	0.21
		2	0.44	0.19	0.44	0.17	0.43	0.16	0.43	0.16
		3	0.50	0.24	0.50	0.23	0.49	0.22	0.49	0.22
		4	0.60	0.24	0.60	0.23	0.59	0.22	0.59	0.22
6	0.09									
		1	0.53	0.22	0.53	0.19	0.52	0.19	0.52	0.19
		2	0.59	0.20	0.58	0.17	0.58	0.16	0.58	0.16
		3	0.53	0.19	0.54	0.19	0.53	0.18	0.53	0.18
		4	0.56	0.23	0.55	0.18	0.55	0.18	0.55	0.18
7	0.07									
		1	0.58	0.22	0.60	0.20	0.59	0.20	0.59	0.20
		2	0.62	0.26	0.64	0.20	0.63	0.20	0.63	0.20
		3	0.75	0.16	0.75	0.16	0.75	0.16	0.75	0.16
		4	0.74	0.20	0.74	0.16	0.74	0.16	0.74	0.16
8	0.10									
		1	0.70	0.20	0.71	0.19	0.71	0.19	0.71	0.19
		2	0.63	0.21	0.65	0.19	0.64	0.19	0.64	0.19
		3	0.61	0.22	0.62	0.20	0.62	0.20	0.61	0.20
		4	0.51	0.22	0.53	0.20	0.52	0.20	0.52	0.20

(table continued)

Table 3 (cont.)
Means and Standard Deviations of OPM's (x_j / n_j),
Priors ($\hat{T}_j$) and OPI's ($\tilde{T}_j$)for Whole Tests

Test	Prop. Sig. Q	Obj. #	x_j / n_j		$\hat{T}_j$		$\tilde{T}_j$ Unadjusted		$\tilde{T}_j$ Adjusted	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
9	0.11									
		1	0.38	0.30	0.38	0.28	0.39	0.28	0.39	0.28
		2	0.32	0.26	0.32	0.23	0.32	0.24	0.32	0.24
		3	0.26	0.23	0.26	0.19	0.27	0.19	0.26	0.19
		4	0.38	0.31	0.38	0.28	0.39	0.29	0.39	0.29
10	0.11									
		1	0.39	0.30	0.38	0.27	0.38	0.27	0.38	0.27
		2	0.35	0.27	0.35	0.25	0.35	0.25	0.35	0.25
		3	0.34	0.27	0.34	0.24	0.34	0.24	0.34	0.24
		4	0.45	0.28	0.44	0.25	0.44	0.26	0.44	0.26
11	0.11									
		1	0.55	0.27	0.55	0.25	0.56	0.26	0.56	0.26
		2	0.60	0.26	0.59	0.24	0.59	0.25	0.59	0.25
		3	0.49	0.27	0.49	0.23	0.50	0.23	0.50	0.23
		4	0.45	0.28	0.45	0.24	0.45	0.24	0.45	0.24
12	0.10									
		1	0.25	0.25	0.28	0.24	0.30	0.26	0.30	0.26
		2	0.28	0.26	0.31	0.25	0.33	0.26	0.33	0.27
		3	0.28	0.26	0.32	0.25	0.33	0.26	0.33	0.27
13	0.12									
		1	0.39	0.31	0.39	0.29	0.39	0.30	0.39	0.30
		2	0.38	0.29	0.38	0.27	0.38	0.28	0.38	0.28
		3	0.29	0.27	0.30	0.25	0.30	0.25	0.29	0.25
14	0.10									
		1	0.61	0.29	0.59	0.27	0.59	0.27	0.59	0.27
		2	0.52	0.28	0.51	0.25	0.51	0.26	0.51	0.26
		3	0.53	0.28	0.57	0.27	0.57	0.28	0.57	0.28
15	0.19									
		1	0.27	0.31	0.30	0.31	0.29	0.34	0.29	0.34
		2	0.31	0.34	0.35	0.33	0.31	0.35	0.31	0.35
		3	0.32	0.34	0.36	0.33	0.31	0.36	0.31	0.36
		4	0.22	0.29	0.26	0.29	0.29	0.36	0.29	0.36
16	0.16									
		1	0.40	0.32	0.40	0.30	0.40	0.31	0.40	0.31
		2	0.34	0.33	0.35	0.31	0.35	0.32	0.35	0.32
		3	0.44	0.34	0.45	0.32	0.45	0.33	0.45	0.33
		4	0.51	0.33	0.51	0.32	0.51	0.32	0.51	0.33
17	0.13									
		1	0.53	0.30	0.54	0.28	0.54	0.28	0.54	0.28
		2	0.57	0.29	0.57	0.27	0.58	0.27	0.58	0.27
		3	0.57	0.30	0.57	0.28	0.57	0.29	0.57	0.29
		4	0.61	0.31	0.61	0.29	0.62	0.29	0.62	0.29

The prior means are within .04 of the OPM means. The standard deviation of the priors and OPIs are, as expected, reduced relative to the standard deviations of the OPMs. The adjustment has little effect on the OPI summary statistics for any objective. OPI means are very similar to OPM means; the greatest deviation is .07 for the fourth objective of Test #15.

Means and standard deviations of the weights (w_j) and the predicted standard deviations of the objective posterior distributions [$\sigma(T_j \mid x_j)$] are provided in Table 4. The adjustment makes the expected reduction in w_j and increase in $\sigma(T_j \mid x_j)$. The increase in $\sigma(T_j \mid x_j)$ is small and only infrequently visible. The largest decrease in w_j occurs for those objectives that have the largest proportions of score points within a test (e.g. objective number 3 in both Test 6 and Test 7; see Table 1).

Half Tests

Table 5 contains the means, standard deviations, and correlations of the OPMs and priors across the selected split-half objectives, as well as the proportion of simulees with significant Q values for each half test. As expected, the proportions of significant Q values are, with only a few exceptions, lower for the halves than for the whole tests. More simulees have prior information incorporated into their OPIs for the test halves than for the whole tests.

The correlation of OPMs across the split-halves of an objective can be rather small, despite the controls placed on the construction of the split-halves. The attenuated correlations are no doubt due in large measure to the small size of some of the split-half objectives (e.g. the fourth objective in Test 6).

Tables 6 through 8 contain the same descriptive statistics for unadjusted and adjusted versions of the weights, OPIs, and predicted posterior standard deviations, respectively. The expected reductions in the weights under the adjusted, as opposed to unadjusted, procedure are clearly visible in Table 6. The expected increases in the predicted standard deviations of the posterior are not so readily apparent in Table 8, due to the restricted magnitude of the standard deviations. The correlations of the OPIs, both unadjusted and adjusted, across the split-halves of the selected objectives in Table 7 are substantially larger than those seen for the OPMs in Table 5.

Table 4
Means and Standard Deviations of w_j and $\sigma(T_j \mid x_j)$ for Whole Tests

		w_j				$\sigma(T_j \mid x_j)$			
		Unadjusted		Adjusted		Unadjusted		Adjusted	
Test	Obj. #	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1									
	1	0.78	0.05	0.69	0.06	0.05	0.01	0.06	0.01
	2	0.76	0.08	0.66	0.11	0.05	0.02	0.06	0.02
	3	0.86	0.10	0.84	0.12	0.05	0.02	0.06	0.02
2									
	1	0.87	0.06	0.83	0.08	0.05	0.02	0.06	0.02
	2	0.83	0.03	0.79	0.03	0.06	0.01	0.07	0.01
	3	0.80	0.06	0.72	0.07	0.06	0.01	0.07	0.01
	4	0.90	0.06	0.88	0.08	0.05	0.02	0.05	0.02
3									
	1	0.92	0.04	0.91	0.04	0.04	0.02	0.05	0.02
	2	0.88	0.02	0.86	0.02	0.05	0.01	0.05	0.02
	3	0.86	0.03	0.83	0.04	0.05	0.01	0.06	0.01
	4	0.82	0.02	0.78	0.03	0.06	0.01	0.06	0.01
	5	0.90	0.02	0.89	0.03	0.05	0.01	0.05	0.01
	6	0.91	0.03	0.89	0.04	0.05	0.02	0.05	0.02
4									
	1	0.82	0.04	0.78	0.04	0.05	0.01	0.05	0.01
	2	0.89	0.03	0.87	0.04	0.04	0.01	0.05	0.01
	3	0.91	0.02	0.89	0.03	0.04	0.01	0.05	0.01
	4	0.89	0.03	0.87	0.03	0.05	0.01	0.05	0.01
	5	0.90	0.03	0.87	0.03	0.04	0.01	0.04	0.01
5									
	1	0.77	0.10	0.69	0.11	0.05	0.01	0.05	0.01
	2	0.89	0.05	0.87	0.06	0.04	0.01	0.04	0.01
	3	0.79	0.08	0.73	0.09	0.05	0.01	0.06	0.01
	4	0.86	0.05	0.83	0.05	0.05	0.02	0.05	0.02
6									
	1	0.86	0.03	0.83	0.04	0.05	0.01	0.06	0.01
	2	0.90	0.01	0.88	0.02	0.05	0.01	0.05	0.01
	3	0.74	0.07	0.60	0.09	0.05	0.01	0.06	0.01
	4	0.90	0.02	0.89	0.02	0.05	0.01	0.05	0.01
7									
	1	0.84	0.03	0.81	0.03	0.06	0.01	0.06	0.01
	2	0.90	0.03	0.89	0.04	0.06	0.01	0.06	0.01
	3	0.80	0.03	0.70	0.04	0.04	0.01	0.05	0.01
	4	0.85	0.02	0.81	0.03	0.05	0.01	0.05	0.01
8									
	1	0.79	0.04	0.72	0.04	0.05	0.01	0.05	0.01
	2	0.83	0.03	0.79	0.03	0.05	0.01	0.06	0.01
	3	0.82	0.01	0.78	0.02	0.05	0.01	0.06	0.01
	4	0.83	0.06	0.79	0.07	0.06	0.01	0.06	0.01

(table continued)

Table 4 (cont.)
Means and Standard Deviations of w_j and $\sigma(T_j \mid x_j)$ for Whole Tests

		w_j				$\sigma(T_j \mid x_j)$			
		Unadjusted		Adjusted		Unadjusted		Adjusted	
Test	Obj. #	Mean	SD	Mean	SD	Mean	SD	Mean	SD
9									
	1	0.77	0.16	0.70	0.19	0.05	0.03	0.06	0.03
	2	0.83	0.11	0.79	0.14	0.05	0.03	0.05	0.03
	3	0.88	0.14	0.86	0.16	0.04	0.04	0.04	0.04
	4	0.77	0.15	0.72	0.17	0.06	0.03	0.06	0.03
10									
	1	0.78	0.11	0.73	0.14	0.06	0.03	0.06	0.03
	2	0.80	0.13	0.76	0.15	0.05	0.03	0.06	0.03
	3	0.83	0.11	0.78	0.13	0.05	0.03	0.06	0.03
	4	0.79	0.07	0.74	0.09	0.06	0.02	0.07	0.02
11									
	1	0.78	0.08	0.72	0.10	0.06	0.02	0.07	0.02
	2	0.78	0.09	0.72	0.11	0.06	0.02	0.06	0.02
	3	0.85	0.05	0.82	0.06	0.06	0.01	0.06	0.02
	4	0.86	0.05	0.84	0.06	0.06	0.02	0.06	0.02
12									
	1	0.68	0.26	0.61	0.29	0.05	0.04	0.06	0.04
	2	0.66	0.25	0.58	0.28	0.05	0.03	0.06	0.04
	3	0.65	0.26	0.57	0.29	0.05	0.03	0.06	0.04
13									
	1	0.69	0.17	0.60	0.20	0.05	0.03	0.06	0.03
	2	0.74	0.13	0.66	0.16	0.05	0.02	0.06	0.03
	3	0.74	0.20	0.67	0.24	0.05	0.03	0.05	0.03
14									
	1	0.75	0.11	0.67	0.13	0.05	0.02	0.06	0.02
	2	0.78	0.04	0.70	0.05	0.05	0.01	0.06	0.01
	3	0.75	0.06	0.66	0.08	0.05	0.02	0.06	0.02
15									
	1	0.57	0.29	0.52	0.29	0.05	0.04	0.05	0.04
	2	0.59	0.25	0.53	0.26	0.05	0.04	0.05	0.04
	3	0.57	0.21	0.51	0.21	0.05	0.04	0.05	0.04
	4	0.62	0.33	0.57	0.34	0.04	0.04	0.05	0.04
16									
	1	0.78	0.10	0.73	0.11	0.05	0.02	0.05	0.03
	2	0.78	0.14	0.73	0.16	0.05	0.03	0.05	0.03
	3	0.77	0.09	0.72	0.11	0.05	0.02	0.06	0.03
	4	0.77	0.12	0.72	0.14	0.05	0.02	0.06	0.03
17									
	1	0.76	0.08	0.71	0.09	0.06	0.02	0.06	0.02
	2	0.77	0.07	0.72	0.09	0.05	0.02	0.06	0.02
	3	0.76	0.08	0.70	0.10	0.06	0.02	0.06	0.02
	4	0.74	0.11	0.69	0.13	0.05	0.02	0.06	0.03

Table 5
Proportions of Significant Q Values and Means, Standard Deviations,
Correlations of OPM's (x_j/n_j), and Priors $(\hat{T}_j)$ for Half Tests

				OPM (x_j/n_j)					$\hat{T}_j$				
	Prop. Sig. Q			Mean		SD			Mean		SD		
Test	I	II	Obj. #	I	II	I	II	r	I	II	I	II	r
1	0.03	0.03											
			1	0.64	0.65	0.23	0.23	0.50	0.64	0.64	0.17	0.19	0.76
			2	0.61	0.62	0.25	0.26	0.62	0.61	0.61	0.21	0.22	0.77
2	0.09	0.06											
			2	0.51	0.52	0.28	0.25	0.38	0.48	0.51	0.23	0.21	0.69
			3	0.58	0.56	0.22	0.25	0.42	0.58	0.56	0.22	0.26	0.68
			4	0.46	0.47	0.28	0.24	0.26	0.40	0.45	0.16	0.14	0.70
4	0.07	0.06											
			1	0.62	0.62	0.23	0.25	0.56	0.62	0.65	0.20	0.21	0.80
			2	0.57	0.56	0.26	0.25	0.41	0.56	0.59	0.19	0.17	0.81
5	0.05	0.05											
			3	0.50	0.50	0.25	0.28	0.60	0.50	0.50	0.20	0.24	0.85
			4	0.60	0.61	0.29	0.28	0.57	0.60	0.61	0.27	0.21	0.85
6	0.05	0.07											
			2	0.59	0.59	0.23	0.27	0.30	0.62	0.58	0.15	0.21	0.74
			3	0.54	0.52	0.21	0.21	0.58	0.53	0.51	0.18	0.19	0.76
			4	0.56	0.55	0.30	0.27	0.30	0.55	0.53	0.21	0.17	0.76
7	0.05	0.05											
			1	0.57	0.58	0.23	0.27	0.41	0.60	0.61	0.18	0.20	0.74
8	0.09	0.07											
			2	0.63	0.63	0.23	0.26	0.46	0.65	0.64	0.21	0.20	0.79
			3	0.61	0.61	0.26	0.26	0.43	0.64	0.63	0.18	0.20	0.78
			4	0.50	0.50	0.27	0.27	0.39	0.52	0.53	0.23	0.18	0.77
9	0.06	0.06											
			3	0.25	0.26	0.27	0.27	0.47	0.27	0.27	0.19	0.21	0.88
11	0.05	0.04											
			2	0.60	0.60	0.30	0.27	0.67	0.60	0.53	0.27	0.23	0.86
			3	0.49	0.48	0.33	0.29	0.57	0.49	0.58	0.25	0.25	0.86
12	0.08	0.10											
			1	0.27	0.24	0.27	0.26	0.74	0.27	0.24	0.24	0.24	0.91
			3	0.30	0.27	0.28	0.29	0.73	0.30	0.27	0.25	0.27	0.90
13	0.07	0.07											
			2	0.36	0.40	0.31	0.30	0.76	0.36	0.40	0.28	0.28	0.90
			3	0.28	0.30	0.28	0.30	0.70	0.29	0.31	0.24	0.27	0.90
14	0.05	0.06											
			1	0.61	0.62	0.30	0.32	0.75	0.61	0.61	0.27	0.30	0.90
			3	0.56	0.51	0.29	0.30	0.72	0.56	0.52	0.26	0.28	0.90
15	0.08	0.09											
			1	0.25	0.29	0.31	0.36	0.76	0.25	0.28	0.28	0.33	0.93
			2	0.33	0.29	0.36	0.35	0.81	0.33	0.29	0.33	0.33	0.93
			3	0.34	0.30	0.36	0.36	0.78	0.34	0.30	0.32	0.34	0.93
16	0.08	0.80											
			1	0.41	0.40	0.32	0.37	0.73	0.40	0.40	0.29	0.34	0.91
			3	0.44	0.45	0.37	0.36	0.75	0.44	0.45	0.33	0.32	0.92
			4	0.50	0.51	0.37	0.35	0.74	0.50	0.51	0.33	0.32	0.92
17	0.07	0.05											
			1	0.54	0.53	0.32	0.33	0.71	0.54	0.54	0.28	0.29	0.91
			2	0.58	0.57	0.33	0.31	0.68	0.57	0.57	0.28	0.27	0.91
			3	0.57	0.57	0.35	0.31	0.70	0.57	0.57	0.32	0.26	0.90
			4	0.60	0.62	0.34	0.33	0.73	0.60	0.62	0.30	0.29	0.91

Table 6
Means, Standaard Deviations, and Correlations of w_j for Half Tests

		w_j for Unadjusted $\tilde{T}_j$					w_j for Adjusted $\tilde{T}_j$				
		Mean		SD			Mean		SD		
Test	Obj.#	I	II	I	II	r	I	II	I	II	r
1											
	1	0.80	0.77	0.08	0.04	0.06	0.70	0.67	0.10	0.05	0.05
	2	0.72	0.76	0.09	0.06	0.36	0.58	0.67	0.13	0.08	0.35
2											
	2	0.86	0.84	0.08	0.04	0.02	0.83	0.79	0.09	0.05	0.03
	3	0.77	0.77	0.05	0.09	0.14	0.69	0.70	0.07	0.12	0.14
	4	0.93	0.93	0.08	0.05	0.63	0.91	0.91	0.09	0.06	0.63
4											
	1	0.85	0.83	0.04	0.03	0.29	0.82	0.78	0.05	0.03	0.21
	2	0.89	0.92	0.03	0.04	0.20	0.87	0.91	0.04	0.05	0.39
5											
	3	0.82	0.74	0.08	0.07	0.63	0.77	0.66	0.09	0.09	0.63
	4	0.83	0.86	0.06	0.03	0.21	0.80	0.84	0.07	0.03	0.21
6											
	2	0.91	0.88	0.04	0.04	-0.46	0.90	0.85	0.05	0.05	-0.47
	3	0.75	0.79	0.05	0.07	0.19	0.63	0.66	0.06	0.10	0.23
	4	0.87	0.93	0.04	0.03	0.56	0.85	0.92	0.04	0.04	0.56
7											
	1	0.87	0.84	0.06	0.04	0.37	0.84	0.80	0.07	0.05	0.38
8											
	2	0.82	0.83	0.03	0.04	0.08	0.78	0.78	0.03	0.05	0.08
	3	0.87	0.84	0.04	0.02	-0.18	0.84	0.80	0.05	0.02	-0.18
	4	0.80	0.89	0.07	0.04	0.48	0.75	0.87	0.08	0.04	0.47
9											
	3	0.89	0.85	0.12	0.19	0.72	0.87	0.83	0.15	0.21	0.74
11											
	2	0.74	0.82	0.14	0.04	0.35	0.67	0.76	0.17	0.06	0.35
	3	0.85	0.80	0.06	0.09	0.27	0.82	0.73	0.07	0.11	0.27
12											
	1	0.75	0.49	0.20	0.26	0.58	0.69	0.40	0.25	0.27	0.53
	3	0.74	0.43	0.18	0.28	0.49	0.67	0.35	0.22	0.27	0.41
13											
	2	0.72	0.77	0.16	0.13	0.72	0.63	0.69	0.19	0.16	0.71
	3	0.73	0.74	0.20	0.23	0.81	0.66	0.67	0.23	0.27	0.82
14											
	1	0.73	0.74	0.12	0.13	0.60	0.65	0.66	0.14	0.15	0.59
	3	0.75	0.78	0.05	0.08	0.25	0.67	0.71	0.07	0.10	0.26
15											
	1	0.41	0.58	0.35	0.36	0.16	0.37	0.54	0.34	0.36	0.21
	2	0.37	0.60	0.28	0.35	-0.09	0.32	0.55	0.25	0.37	-0.03
	3	0.40	0.58	0.28	0.34	-0.17	0.34	0.53	0.26	0.36	-0.11
16											
	1	0.77	0.78	0.16	0.16	-0.09	0.71	0.73	0.18	0.34	-0.06
	3	0.74	0.80	0.13	0.10	0.39	0.68	0.75	0.16	0.12	0.41
	4	0.74	0.80	0.16	0.11	0.47	0.68	0.76	0.18	0.14	0.47
17											
	1	0.78	0.74	0.10	0.08	0.44	0.73	0.68	0.12	0.09	0.43
	2	0.78	0.76	0.08	0.09	0.49	0.73	0.71	0.10	0.10	0.48
	3	0.75	0.77	0.14	0.08	0.20	0.69	0.71	0.17	0.09	0.20
	4	0.76	0.72	0.11	0.14	0.69	0.71	0.66	0.13	0.16	0.68

Table 7
Means, Standard Deviations, and Correlations of OPI's ($\tilde{T}_j$) for Half Tests

		Unadjusted $\tilde{T}_j$					Adjusted $\tilde{T}_j$				
		Mean		SD			Mean		SD		
Test	Obj.#	I	II	I	II	r	I	II	I	II	r
1											
	1	0.64	0.65	0.17	0.19	0.74	0.64	0.65	0.17	0.19	0.72
	2	0.61	0.62	0.21	0.23	0.76	0.61	0.61	0.22	0.23	0.74
2											
	2	0.47	0.50	0.22	0.20	0.66	0.48	0.50	0.22	0.20	0.65
	3	0.57	0.56	0.21	0.22	0.65	0.57	0.56	0.21	0.22	0.64
	4	0.40	0.45	0.16	0.13	0.67	0.40	0.45	0.16	0.13	0.67
4											
	1	0.61	0.64	0.19	0.21	0.78	0.61	0.63	0.19	0.21	0.78
	2	0.55	0.58	0.19	0.16	0.79	0.55	0.58	0.19	0.16	0.78
5											
	3	0.50	0.50	0.20	0.25	0.83	0.50	0.50	0.20	0.25	0.82
	4	0.60	0.61	0.23	0.21	0.83	0.60	0.61	0.23	0.21	0.83
6											
	2	0.62	0.57	0.15	0.21	0.73	0.62	0.57	0.15	0.21	0.72
	3	0.53	0.51	0.18	0.18	0.74	0.54	0.51	0.18	0.18	0.73
	4	0.55	0.53	0.21	0.17	0.75	0.55	0.53	0.21	0.17	0.74
7											
	1	0.59	0.60	0.18	0.20	0.72	0.59	0.60	0.18	0.20	0.71
8											
	2	0.64	0.64	0.21	0.20	0.76	0.64	0.64	0.20	0.21	0.76
	3	0.64	0.62	0.18	0.20	0.76	0.63	0.62	0.18	0.20	0.75
	4	0.51	0.52	0.22	0.18	0.74	0.51	0.52	0.22	0.18	0.73
9											
	3	0.27	0.27	0.19	0.22	0.78	0.27	0.27	0.19	0.22	0.76
11											
	2	0.60	0.53	0.27	0.24	0.85	0.60	0.53	0.28	0.24	0.84
	3	0.50	0.58	0.26	0.25	0.85	0.50	0.58	0.26	0.25	0.84
12											
	1	0.28	0.24	0.25	0.26	0.87	0.28	0.24	0.25	0.26	0.86
	3	0.31	0.27	0.25	0.29	0.86	0.31	0.27	0.26	0.29	0.85
13											
	2	0.36	0.40	0.29	0.28	0.90	0.36	0.40	0.29	0.28	0.89
	3	0.29	0.31	0.25	0.28	0.88	0.29	0.30	0.25	0.28	0.86
14											
	1	0.61	0.61	0.27	0.30	0.88	0.61	0.61	0.28	0.30	0.87
	3	0.56	0.51	0.26	0.28	0.89	0.56	0.51	0.27	0.28	0.88
15											
	1	0.24	0.28	0.29	0.35	0.89	0.24	0.28	0.29	0.35	0.88
	2	0.33	0.28	0.35	0.34	0.91	0.33	0.28	0.35	0.35	0.90
	3	0.33	0.30	0.34	0.35	0.90	0.33	0.30	0.34	0.36	0.89
16											
	1	0.40	0.40	0.29	0.35	0.89	0.40	0.40	0.30	0.35	0.89
	3	0.44	0.45	0.34	0.33	0.92	0.44	0.45	0.34	0.33	0.91
	4	0.50	0.51	0.34	0.32	0.91	0.50	0.51	0.34	0.32	0.91
17											
	1	0.55	0.54	0.29	0.29	0.89	0.55	0.54	0.29	0.29	0.88
	2	0.58	0.57	0.29	0.27	0.90	0.58	0.57	0.29	0.27	0.89
	3	0.57	0.57	0.32	0.27	0.89	0.57	0.57	0.32	0.27	0.88
	4	0.61	0.62	0.30	0.30	0.90	0.61	0.62	0.30	0.30	0.89

Table 8
Means, Standard Deviations and Correlations of $\sigma(T_j \mid x_j)$ for Half Tests

		Unadjusted $\sigma(T_j \mid x_j)$					Adjusted $\sigma(T_j \mid x_j)$				
		Mean		SD			Mean		SD		
Test	Obj.#	I	II	I	II	r	I	II	I	II	r
1											
	1	0.07	0.07	0.02	0.01	0.56	0.09	0.08	0.02	0.02	0.55
	2	0.08	0.08	0.02	0.02	0.48	0.10	0.09	0.03	0.03	0.47
2											
	2	0.08	0.08	0.03	0.02	0.18	0.09	0.09	0.03	0.02	0.18
	3	0.08	0.08	0.02	0.03	0.27	0.09	0.09	0.02	0.03	0.26
	4	0.05	0.05	0.03	0.02	0.60	0.06	0.06	0.03	0.02	0.60
4											
	1	0.06	0.06	0.01	0.01	0.38	0.07	0.07	0.02	0.01	0.38
	2	0.06	0.05	0.01	0.02	0.21	0.07	0.06	0.02	0.02	0.21
5											
	3	0.07	0.07	0.02	0.02	0.55	0.07	0.08	0.02	0.02	0.53
	4	0.08	0.07	0.03	0.02	0.52	0.08	0.08	0.03	0.02	0.52
6											
	2	0.06	0.07	0.01	0.02	-0.14	0.06	0.07	0.02	0.02	-0.13
	3	0.07	0.06	0.01	0.01	0.33	0.08	0.07	0.01	0.02	0.35
	4	0.08	0.06	0.02	0.01	0.40	0.08	0.06	0.02	0.01	0.40
7											
	1	0.08	0.08	0.02	0.02	0.32	0.09	0.09	0.02	0.02	0.32
8											
	2	0.07	0.07	0.02	0.02	0.46	0.08	0.08	0.02	0.02	0.46
	3	0.06	0.07	0.01	0.01	0.42	0.07	0.08	0.01	0.02	0.42
	4	0.08	0.07	0.02	0.01	0.27	0.09	0.07	0.02	0.01	0.25
9											
	3	0.05	0.06	0.05	0.06	0.81	0.06	0.06	0.05	0.06	0.82
11											
	2	0.08	0.07	0.04	0.02	0.52	0.09	0.09	0.04	0.02	0.52
	3	0.08	0.08	0.03	0.03	0.40	0.09	0.09	0.03	0.03	0.40
12											
	1	0.06	0.08	0.05	0.05	0.72	0.07	0.08	0.05	0.05	0.71
	3	0.06	0.08	0.04	0.04	0.65	0.07	0.09	0.05	0.05	0.63
13											
	2	0.07	0.07	0.04	0.03	0.69	0.08	0.08	0.04	0.04	0.68
	3	0.06	0.06	0.04	0.04	0.79	0.07	0.07	0.04	0.05	0.78
14											
	1	0.07	0.07	0.03	0.03	0.67	0.08	0.08	0.03	0.04	0.67
	3	0.07	0.07	0.02	0.02	0.58	0.08	0.08	0.02	0.03	0.57
15											
	1	0.06	0.06	0.06	0.06	0.68	0.07	0.06	0.06	0.06	0.68
	2	0.07	0.07	0.06	0.06	0.73	0.07	0.07	0.06	0.07	0.70
	3	0.07	0.07	0.05	0.06	0.72	0.07	0.07	0.06	0.07	0.68
16											
	1	0.07	0.06	0.03	0.04	0.44	0.08	0.07	0.04	0.04	0.47
	3	0.07	0.07	0.04	0.03	0.64	0.08	0.07	0.04	0.04	0.64
	4	0.07	0.07	0.04	0.03	0.57	0.08	0.07	0.04	0.04	0.57
17											
	1	0.07	0.08	0.03	0.03	0.57	0.08	0.09	0.03	0.03	0.57
	2	0.07	0.08	0.03	0.03	0.60	0.08	0.09	0.03	0.03	0.60
	3	0.07	0.08	0.04	0.02	0.51	0.08	0.09	0.04	0.03	0.51
	4	0.07	0.08	0.03	0.04	0.70	0.08	0.09	0.04	0.04	0.69

The proportions of simulees with overlapping credibility intervals, the predicted posterior standard deviations, and the observed standard errors are presented in Table 9. The unadjusted procedure produces proportion overlaps that range from .41 to .70, which is lower than the range expected from the normal approximation. The adjusted overlap proportions range from .42 to .74. The objectives exhibiting the greatest degree of overlap were from Tests 1 and 5, with adjusted overlap proportions ranging between .68 and .74. These objectives were of moderate difficulty (Table 3) with few c.r. items. The objectives demonstrating the smallest amount of (adjusted) overlap were from Test 15. The objectives in Test 15 were very difficult and composed solely of c.r. items.

Yen (1987), in her application of the original OPI procedure for tests composed of exclusively s.r. items, found unadjusted proportion overlaps that ranged between .65 to .75 and adjusted proportion overlaps between .72 and .79. Thus, the OPI procedure was more accurate when dealing solely with s.r. items than with mixtures of s.r. and c.r. items. It appears likely that the inaccuracy arose from the treatment of c.r. items with l_j score levels as combinations of $l_j - 1$ independent Bernoulli variables. Further adjustment of the procedure to deal with this inaccuracy bears exploration.

The observed standard errors are greater than the adjusted predicted posterior standard deviations by .01 to .03. Similar underestimation of standard errors was found by Yen (1987) for tests composed solely of s.r. items.

It should be noted that the level of inaccuracy in these standard errors is very unlikely to have any practical importance in the use of the OPI scores by teachers. For example, it is difficult to imagine a negative consequence of reporting a credibility interval for a student's objective performance of .41 to .59, rather than the more accurate .38 to .62. Furthermore, given these credibility intervals are typically reported graphically rather than numerically, it is unlikely that differences could be visually detected between actual and predictive confidence bands. Thus, the present procedure appears quite sufficiently accurate for its intended use. Of course, further enhancement of the procedure would be welcome.

Table 9
Proportion of Simulees with Overlapping Credibility Intervals, Observed OPI Standard Errors, and Predicted Posterior Standard Deviations

		Proportion Overlap		Unadjusted		Adjusted	
Test	Obj. #	Unadjusted	Adjusted	SE	$\bar{\sigma}(T_j \mid x_j)$	SE	$\bar{\sigma}(T_j \mid x_j)$
1							
	1	0.69	0.74	0.09	0.07	0.10	0.09
	2	0.68	0.73	0.11	0.08	0.11	0.10
2							
	2	0.53	0.56	0.12	0.08	0.12	0.09
	3	0.55	0.60	0.13	0.08	0.13	0.10
	4	0.45	0.49	0.09	0.06	0.09	0.06
4							
	1	0.57	0.61	0.09	0.06	0.10	0.07
	2	0.57	0.60	0.08	0.06	0.08	0.07
5							
	3	0.64	0.68	0.10	0.07	0.10	0.08
	4	0.70	0.72	0.09	0.08	0.09	0.08
6							
	2	0.52	0.55	0.10	0.07	0.10	0.07
	3	0.58	0.64	0.09	0.06	0.10	0.07
	4	0.59	0.61	0.10	0.07	0.10	0.08
7							
	1	0.63	0.66	0.10	0.08	0.10	0.09
8							
	2	0.58	0.62	0.10	0.07	0.10	0.08
	3	0.58	0.62	0.10	0.07	0.10	0.08
	4	0.55	0.58	0.11	0.07	0.11	0.08
9							
	3	0.54	0.56	0.09	0.08	0.09	0.08
11							
	2	0.57	0.61	0.11	0.08	0.12	0.09
	3	0.59	0.64	0.12	0.08	0.12	0.09
12							
	1	0.65	0.65	0.10	0.08	0.10	0.09
	3	0.59	0.52	0.11	0.09	0.11	0.09
13							
	2	0.60	0.64	0.10	0.08	0.10	0.09
	3	0.63	0.67	0.10	0.08	0.10	0.08
14							
	1	0.63	0.67	0.10	0.08	0.10	0.09
	3	0.61	0.65	0.10	0.07	0.10	0.08
15							
	1	0.41	0.42	0.12	0.09	0.12	0.09
	2	0.45	0.46	0.11	0.09	0.11	0.09
	3	0.41	0.42	0.11	0.09	0.12	0.09
16							
	1	0.46	0.48	0.11	0.08	0.11	0.08
	3	0.53	0.55	0.10	0.08	0.10	0.08
	4	0.54	0.57	0.10	0.08	0.10	0.08
17							
	1	0.64	0.67	0.10	0.08	0.10	0.09
	2	0.67	0.69	0.09	0.08	0.09	0.09
	3	0.59	0.62	0.11	0.08	0.11	0.09
	4	0.66	0.69	0.10	0.08	0.10	0.09

Conclusions

The two goals of this research were to develop a more accurate objective score for tests containing mixed item types and to estimate the standard error of this objective score.

The use of prior information (i.e., overall test performance) to calculate OPIs resulted in objective scores with substantially smaller standard errors than the OPMs, which do not utilize this information. The unadjusted OPI procedure overestimated the amount of independent information provided by the prior distribution and underestimated the length of the 67% credibility interval; this effect was more pronounced as the proportion of c.r. items increased. The adjusted OPI procedure more frequently produced credibility intervals that overlapped, although still a smaller proportion than that expected using the normal approximation.

The adjusted OPI procedure produced estimated standard errors that were lower than the empirical standard errors by .01 to .03. While this effect bears exploration and improvement, the current adjusted estimates appear sufficiently accurate for use in diagnostic score reports used by teachers.

References

Burket, G. R. (1991; 1995). *PARDUX.* Monterey, CA: CTB/McGraw-Hill.

Fitzpatrick, A. R. (1994). *Two studies comparing parameter estimates produced by PARDUX and BIGSTEPS.* Unpublished manuscript.

Fitzpatrick, A. R., Link, V. B., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. C. Scaling performance assessments: A comparison of one-parameter and two-parameter partial credit models. *Journal of Educational Measurement, 33,* 291-314.

Johnson, N. L., & Kotz, S. (1970). *Distributions in statistics: Continuous univariate distributions, Vol 2.* New York: John Wiley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale: Lawrence Erlbaum.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48,* 233-245.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* New York: McGraw-Hill.

Thissen, D. (1986). *MULTILOG: Multiple categorical item analysis and test scoring, Version 5.* Mooresville, IN: Scientific Software.

Yen, W. M. (1987). *A Bayesian/IRT index of objective performance.* Paper presented at the meeting of the Psychometric Society, Montreal.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.