A Bayesian/IRT Index of Objective Performance[1]

Wendy M. Yen
CTB/McGraw-Hill

Presented at the meeting of the Psychometric Society,
Montreal, June 1987

[1] In order to provide more ready access to this unpublished paper, it is reprinted here in the form in which it was originally presented in 1987.

Abstract

This paper presents a combination Bayesian/Item Response Theory procedure for pooling performance on a particular objective with information about an examinee's overall test performance in order to produce more stable objective scores. The procedure, including the calculation of a posterior distribution, is described. A split-half cross validation study finds that a credibility interval based on the posterior distribution is sufficiently accurate to be useful for scoring reports for teachers.

A Bayesian/IRT Index of Objective Performance

In many achievement tests items can be grouped into objectives and scores reported separately by objective. This information can be used by teachers to judge a student's strengths and weaknesses with respect to particular objectives. When the number of items per objective is small, the standard errors of these objective scores can be quite large. In order to produce a more stable objective score a combination Bayesian/Item Response Theory (IRT) procedure is derived that pools performance on a particular objective with information from the examinee's overall test performance. The resulting objective score is called the Objective Performance Index (OPI). The procedure, which includes the calculation of a credibility interval, is described and then an example and a split-half cross validation study of the procedure are presented.

Derivation of the Procedure

The Objective Performance Index is an estimated true score (estimated proportion-correct score) for the items in an objective based on the performance of a given examinee. [The procedure for calculating the OPI was first described in a brief technical appendix by Yen (1982).] In the following description it is assumed that only one examinee and one test (e.g., Reading Vocabulary) are being considered at a time, and for the sake of simplicity the subscripts for the examinee and test are not presented.

It is assumed that each *n*-item test is composed of *J* objectives, with n_j items in objective j; an item contributes to at most one objective and some items do not contribute to any objective. Let X_j be the observed number-correct score on objective j, and $T_j \equiv E(X_j / n_j)$. It is assumed that if there were information available about an examinee in addition to the objective score, a prior distribution for T_j could be specified. This additional or prior information might be the examinee's grade in school or performance on another test. The subsection "Estimating the Prior Distribution of T_j" describes the prior information used in the OPI procedure.

It is assumed that the prior distribution of T_j for a given examinee is $\beta(r_j, s_j)$. That is,

$$g(T_j) = \frac{(r_j + s_j - 1)! T_j^{r_j - 1} (1 - T_j)^{s_j - 1}}{(r_j - 1)!(s_j - 1)!} \text{ for } 0 \le T_j \le 1; r_j, s_j > 0. \qquad (1)$$

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

$$p(X_j = x_j \mid T_j) = \binom{n_j}{x_j} T_j^{x_j} (1 - T_j)^{n_j - x_j} \text{ for } 0 \le x_j \le n_j;\ 0 \le T_j \le 1. \quad (2)$$

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

$$g(T_j \mid X_j = x_j) = \beta(p_j, q_j), \quad (3)$$

with

$$p_j = r_j + x_j \quad (4)$$

and

$$q_j = s_j + n_j - x_j. \quad (5)$$

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

$$\tilde{T}_j = \frac{p_j}{p_j + q_j}. \quad (6)$$

Following Novick and Jackson (1974, p. 119), a mastery band is created to be the C% central credibility interval for T_j. This band is obtained by identifying the values that place $\frac{1}{2}(100 - C)\%$ of the $\beta(p_j, q_j)$ distribution in each tail.

Estimating the Prior Distribution of T_j

The *n* items in each test are scaled together using the three-parameter IRT model, and estimated item parameters are obtained. It is assumed that the item parameters are estimated using a large sample (e.g., a norm sample) and that these parameters are held fixed in subsequent scoring to obtain OPI values. $P_{ij}(\hat{\theta})$ is the item response function for item *i* in objective *j*, and $\hat{\theta}$ is the common estimated trait to which the items are scaled.

$$P_{ij}(\hat{\theta}) = c_{ij} + \frac{1 - c_{ij}}{1 + \exp[-1.7 a_{ij}(\hat{\theta} - b_{ij})]} \quad (7)$$

and

$$T_j = \frac{1}{n_j}\sum_{i=1}^{n_j} P_{ij}(\theta). \tag{8}$$

The estimated proportion-correct score for objective *j* given $\hat{\theta}$ is

$$\hat{T}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} P_{ij}(\hat{\theta}). \tag{9}$$

The theoretical random variation in item response vectors and resulting $\hat{\theta}$ values for a given examinee produces the distribution $g(\hat{T}_j \mid \theta)$ with mean $\mu(\hat{T}_j \mid \theta)$ and variance $\sigma^2(\hat{T}_j \mid \theta)$; this distribution is used to estimate a prior distribution for T_j. In (1) it is assumed that T_j is distributed $\beta(r_j, s_j)$. Expressing the mean and variance of this distribution in terms of the parameters of the beta distribution produces (Novick & Jackson, 1974, p. 113)

$$\mu(\hat{T}_j \mid \theta) = \frac{r_j}{r_j + s_j} \tag{10}$$

and

$$\sigma^2(\hat{T}_j \mid \theta) = \frac{r_j s_j}{(r_j + s_j)^2 (r_j + s_j + 1)}. \tag{11}$$

Solving (10) and (11) for r_j and s_j produces

$$r_j = \mu(\hat{T}_j \mid \theta) n_j^* \tag{12}$$

and

$$s_j = [1 - \mu(\hat{T}_j \mid \theta)] n_j^*, \tag{13}$$

where

$$n_j^* = \frac{\mu(\hat{T}_j \mid \theta)[1 - \mu(\hat{T}_j \mid \theta)]}{\sigma^2(\hat{T}_j \mid \theta)} - 1. \tag{14}$$

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

$$\mu(\hat{T}_j \mid \theta) \approx \frac{1}{n_j}\sum_{i=1}^{n_j} P_{ij}(\theta). \tag{15}$$

Because T_j is a monotonic transformation of θ,

$$\sigma^2(\hat{T}_j \mid \theta) = \sigma^2(\hat{T}_j \mid T_j). \tag{16}$$

From Lord (1980, p. 71)

$$\sigma^2(\hat{T}_j \mid T_j) \approx I(T_j, \hat{T}_j)^{-1}, \tag{17}$$

where $I(T_j, \hat{T}_j)$ is the information that $\hat{T}_j$ contributes about T_j. Using (16), (17), and Lord (1980, p. 85),

$$I(T_j, \hat{T}_j) = \frac{I(\theta, \hat{T}_j)}{[\partial T_j / \partial \theta]^2}. \tag{18}$$

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

$$= \frac{1}{n_j}\sum_{i=1}^{n_j}\frac{\partial P_{ij}(\theta)}{\partial \theta}$$

$$= \frac{1}{n_j}\sum_{i=1}^{n_j} P'_{ij}(\theta). \tag{19}$$

$$P'_{ij}(\theta) = \frac{1.7a_{ij}[1 - P_{ij}(\theta)][P_{ij}(\theta) - c_{ij}]}{(1 - c_{ij})}. \tag{20}$$

From Lord (1980, p. 79),

$$I(\theta, \hat{T}_j) \approx I(\theta, \hat{\theta}). \tag{21}$$

If θ is estimated using the maximum likelihood procedure based on the examinee's pattern of item responses, then from Lord (1980, p. 74)

$$I(\theta, \hat{\theta}) = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\frac{\left[P'_{ij}(\theta)\right]^2}{[P_{ij}(\theta)][1 - P_{ij}(\theta)]}. \tag{22}$$

If θ is estimated using the maximum likelihood procedure based on the examinee's number-correct score on the test,

$$I(\theta,\hat{\theta}) = \frac{\left[\sum_{j=1}^{J}\sum_{i=1}^{n_j} P'_{ij}(\theta)\right]^2}{\sum_{j=1}^{J}\sum_{i=1}^{n_j}[P_{ij}(\theta)][1-P_{ij}(\theta)]}. \qquad (23)$$

In (22) and (23) the simplifying assumption has been made that every item contributes to an objective; if there are items that do not contribute to any objective but do participate in the estimation of θ, the information contributed by those items would be added to (22) or (23).

Thus,

$$\sigma^2(\hat{T}_j \mid \theta) \approx \frac{\left[\frac{1}{n_j}\sum_{i=1}^{n_j} P'_{ij}(\theta)\right]^2}{I(\theta,\hat{\theta})}, \qquad (24)$$

and the parameters of the prior beta distribution for T_j in (12) to (14) can be expressed in terms of the parameters of the three-parameter IRT model by using (9) to estimate (15) and by using (20), (22) or (23), and (24). Using (4) and (5), the parameters of the posterior distribution of T_j also can be expressed in terms of their IRT parameters:

$$p_j = \hat{T}_j n_j^* + x_j, \qquad (25)$$

$$q_j = [1-\hat{T}_j]n_j^* + n_j - x_j, \qquad (26)$$

and

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

$$= \frac{\hat{T}_j n_j^* + x_j}{n_j^* + n_j}. \qquad (27)$$

The OPI can also be written in terms of the relative contribution of the prior estimate, $\hat{T}_j$, and the observed proportion-correct score, x_j / n_j, as

$$\tilde{T}_j = w_j \hat{T}_j + (1-w_j)\frac{x_j}{n_j}, \qquad (28)$$

where w_j is the relative weight given to the prior estimate:

$$w_j = \frac{n_j^*}{n_j^* + n_j}. \qquad (29)$$

Thus, n_j^* can be interpreted as the contribution of the prior in terms of theoretical numbers of items. Note that as the standard error of the prior estimate [the square root of (24)] goes to zero, w_j approaches unity; on the other hand, if $n_j^* = 0,$ no weight is given to the prior estimate.

Check on Consistency

It has been assumed that the item responses can be described by $P_{ij}(\hat{\theta})$. Even if the IRT model appears to accurately describe item performance pooled over examinees, for a given examinee item responses grouped by objective may be multidimensional. For example, a particular examinee may be able to perform difficult addition but not easy subtraction. In such a case it is not appropriate to pool the prior estimate, $\hat{T}_j$, with x_j / n_j. In calculating the OPI the following statistic is used to identify examinees with unexpected performance on the objectives in a test:

$$Q = \sum_{j=1}^{J} \frac{n_j (\frac{x_j}{n_j} - \hat{T}_j)^2}{\hat{T}_j (1 - \hat{T}_j)}. \qquad (30)$$

If $Q \le \chi^2(J,.10)$, (25) to (27) are used to calculate the OPI and its mastery band. If $Q > \chi^2(J,.10)$, n_j^* is set equal to 0 before using (25) to (27).

Violation of Assumptions

In the present application T_j is an expected proportion-correct score on an objective. If the items in the objective do not permit guessing, it is reasonable to assume $0 \le T_j$. However, if correct guessing is possible, as it is with multiple-choice items, there will be a non-zero lower limit to T_j, and a three-parameter beta distribution in which T_j is greater than or equal to this lower limit (Johnson & Kotz, 1970, p. 37) would be more appropriate. Thus, the use of the two-

parameter beta distribution would tend to underestimate T_j among very low-scoring examinees; however, there does not appear to be any practical importance of such underestimation.

The OPI procedure assumes that $p(X_j = x_j \mid T_j)$ is an binomial distribution. This assumption is appropriate only when all the items in an objective have the same item response function. In most applications items will differ in their response functions, so a compound binomial distribution would be more appropriate. However, the binomial can be considered a good first-order approximation to the compound binomial distribution (Lord & Novick, 1968, p. 525; Yen, 1984).

The prior estimate of T_j, $\hat{T}_j$, is based on $\hat{\theta}$, which is based on performance on the entire test, including objective *j*. Thus, the prior estimate is not independent of x_j. It is theoretically possible to obtain $\hat{\theta}_j$ separately for each objective, using only those items not included in objective *j*. However, such repeated trait estimation would greatly increase the computational requirements of the procedure.

The smaller the ratio n_j / n, the less impact the dependence of $\hat{T}_j$ and x_j will have. The posterior variance of T_j given x_j is

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

$$\approx \frac{\tilde{T}_j (1 - \tilde{T}_j)}{n_j^* + n_j + 1}. \qquad (31)$$

The width of the credibility interval will be roughly proportional to the square root of (31). To the extent that n_j^* and n_j reflect overlapping information rather than independent information, the width of the credibility interval will be understated.

A simple modification in the OPI procedure can be made to adjust for the overlapping information. The information in (22) or (23) can be summed over only those items not appearing in the objective of interest. Using this adjustment, the weight for the prior will decrease and the posterior standard deviation will increase. A similar modification that requires less computation

is to multiply the information in (22) or (23) by $(n - n_j)/n$; use of this multiplier will be called the "adjusted OPI procedure."

Summary

The following procedure is followed for each test.

Step 1. Estimate IRT parameters a_{ij}, b_{ij}, and c_{ij} from a large representative sample.

For each examinee the procedure is as follows.

Step 2. Treating the item parameter estimates as fixed values, obtain $\hat{\theta}$ based on overall test performance.

Step 3. For each objective calculate

$$P_{ij}(\hat{\theta}) = c_{ij} + \frac{1 - c_{ij}}{1 + \exp[-1.7a_{ij}(\hat{\theta} - b_{ij})]}$$

and

$$\hat{T}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} P_{ij}(\hat{\theta}).$$

Step 4. Obtain

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

Step 5. If $Q \le \chi^2(J, .10)$,

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

and

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

The OPI is defined to be

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

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

The value of $I(\theta,\hat{\theta})$ used in calculating $\sigma^2(\hat{T}_j \mid \theta)$ and n_j^* is based on (22) or (23) depending upon the scoring method used to obtain $\hat{\theta}$. For the adjusted OPI procedure this information is multiplied by $(n - n_j)/n$. The values of p_j and q_j are used to obtain the 67% credibility interval for T_j from the beta distribution.

If $Q > \chi^2(J,.10)$,

$$\tilde{T}_j = x_j / n_j,$$

$$p_j = x_j,$$

and

$$q_j = n_j - x_j.$$

Example

Figure 1 presents a sample score report for one second grade examinee based on the unadjusted OPI procedure. For all the tests except Language Expression $Q \le \chi^2(J,.10)$, and (25) to (27) were used to calculate the OPI. For Language Expression $Q > \chi^2(J,.10)$; this examinee had unexpectedly low knowledge about the appropriate use of verbs. In reporting the Language Expression OPI values, no prior information was used and the credibility bands were correspondingly wide.

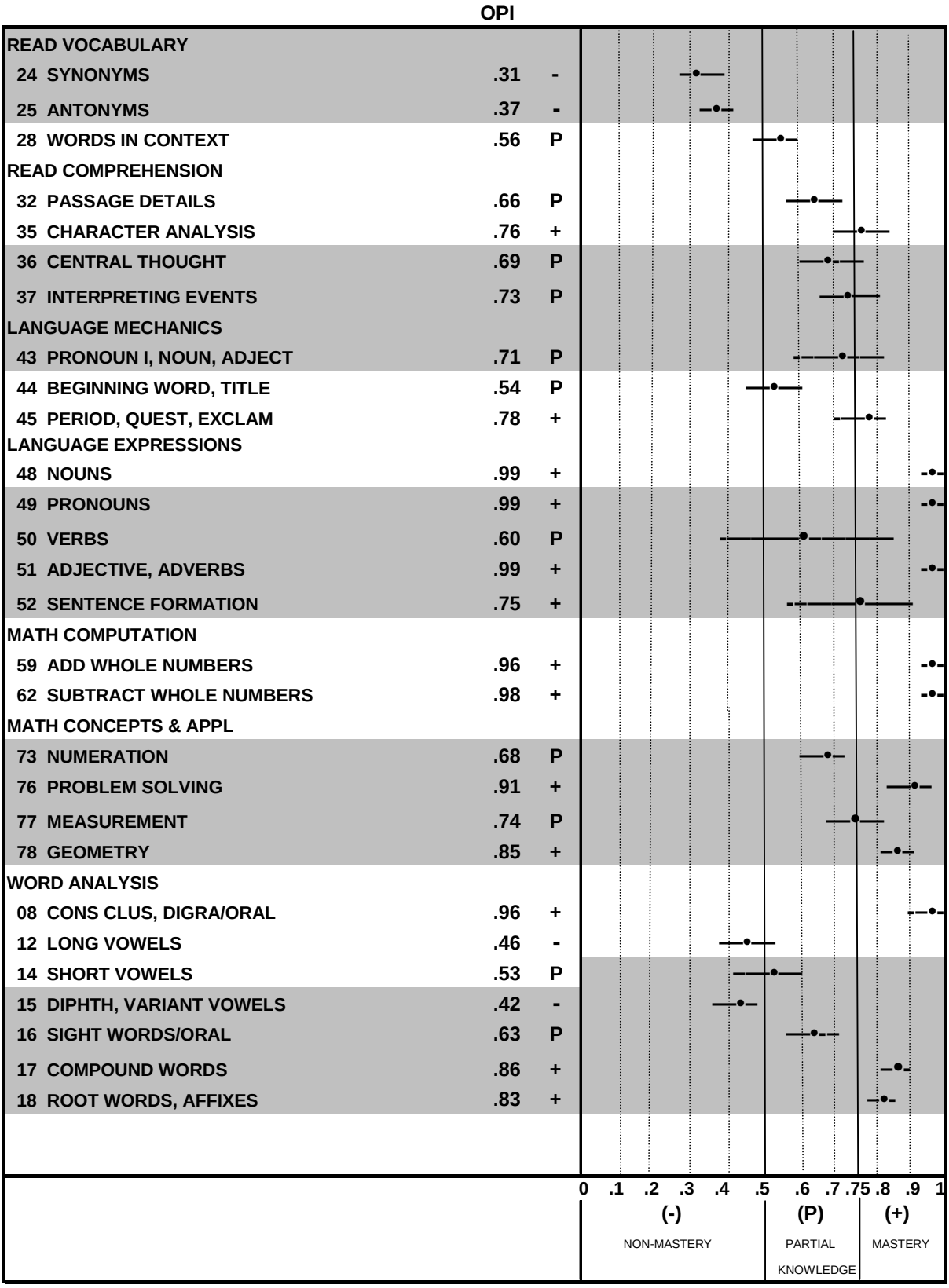

Figure 1 . Example of OPI scores for a second grade student.

Cross Validation

An analysis was conducted to examine the consistency of the OPI values over test halves. Examinees were second grade students tested in the Spring of 1987 with the *California Achievement Tests, Form E* (CAT/E; CTB/McGraw-Hill, 1985). The analyses for each test were based on the performance of 500 examinees who finished that test; embedded omitted items were treated as wrong answers. The tests examined were Word Analysis (WA), Reading Vocabulary (RV), Reading Comprehension (RC), Language Mechanics (LM), Language Expression (LE), Mathematics Computation (MC), and Mathematics Concepts and Applications (MA). Table 1 contains each test's number of items and objective structure. For this study the items were assigned to test halves labeled I and II. Using standardization item difficulties (CTB/McGraw-Hill, 1986), items were assigned to halves so that they would have close to equal difficulties. The numbers of items by objective by half also appear in Table 1.

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

Test	Objective No.	Objective Description	Whole	Half I	Half II
WA			45	22	22
	8	Cons clus, diagra/oral	5	3	2
	12	Long vowels	13	6	6
	14	Short vowels	7	3	4
	15	Diphth, variant vowels	7	4	3
	16	Sight words/oral	4	2	2
	17	Compound words	5	2	3
	18	Root words, affixes	4	2	2
RV			30	15	15
	24	Synonyms	20	10	10
	25	Antonyms	5	2	3
	28	Words in context	5	3	2
RC			30	15	15
	32	Passage details	12	6	6
	35	Character analysis	6	3	3
	36	Central thought	6	3	3
	37	Interpreting events	6	3	3
LM			30	15	15
	43	Pronoun I, noun, adject	10	5	5
	44	Beginning word, title	10	5	5
	45	Period, quest, exclam	10	5	5
LE			30	15	15
	48	Nouns	6	3	3
	49	Pronouns	5	3	2
	50	Verbs	5	2	3
	51	Adjectives, adverbs	6	3	3
	52	Sentence formation	8	4	4
MC			24	12	12
	59	Add whole numbers	12	6	6
	62	Subtract whole numbers	12	6	6
MA			36	18	18
	73	Numeration	16	8	8
	76	Problem Solving	10	5	5
	77	Measurement	6	3	3
	78	Geometry	4	2	2

Note. For test halves, objective scores were calculated and used in (30) only if they had at least 4 items.

OPI Calculations

OPI values were calculated for the whole test and each half using the unadjusted and adjusted OPI procedures. Item parameters were obtained from the standardization of CAT/E (CTB/McGraw-Hill, 1986). Trait values were estimated on the basis of number-correct scores (Yen, 1984); the maximum and minimum possible trait values for each test half were assigned to be the same as those for the whole tests. For the test halves OPI values were obtained only for objectives that had at least 4 items. The items in objectives with fewer than 4 items contributed to the trait estimation, but did not enter into the *Q* statistic; thus, except for the Language Mechanics and Mathematics Computation tests where all the objectives in the halves had at least 4 items, the *Q* statistic was expected to be particularly low in value for the test halves. Also, since the power of the *Q* statistic increases as the number of items increases, more significant *Q* values were expected for the whole tests than the half tests.

Comparisons Between Halves

Means, standard deviations, and correlations were examined over halves. In addition two analyses were done to examine the predicted posterior distribution of the OPI values. In the first analysis the percent of examinees whose 67% credibility intervals overlapped for the two halves was found. Based on the normal approximation to the beta distribution, an expectation of the approximate percent of overlapping intervals was obtained. If the posterior distribution has the same standard deviation for both halves, the normal approximation leads to the expectation that 84 percent of the examinees would have credibility intervals that overlapped for the two halves. This expected percent drops as the ratio of the posterior standard deviations varies from 1:1; if the ratio is 1:2, the expected percent overlap is 77 percent, and if the ratio is 1:3, the expected percent overlap is 73. Given that the posterior standard deviations could vary substantially over halves, the expectation was that the percent overlap would be in the 70's if the credibility intervals were accurate.

The second analysis of the posterior distributions was a comparison of the predicted posterior standard deviation and the observed standard error of $\tilde{T}_j$. The prediction was the square

root of the average of (31) calculated for the two halves. The observed standard error was the square root of half the mean squared difference between $\tilde{T}_j$ for the two halves.

Results

Table 2 contains the means and standard deviations of x_j / n_j, $\hat{T}_j$, and $\tilde{T}_j$ by objective for the whole tests; that table also displays the proportion of examinees with significant *Q* values. The means and standard deviations of w_j and $\sigma(T_j \mid x_j)$ are in Table 3. The adjustment makes the expected reduction in w_j and increase in $\sigma(T_j \mid x_j)$, with little change in $\tilde{T}_j$ summary statistics.

Table 2
Means and Standard Deviations of
x_j / n_j, $\hat{T}_j$, and $\tilde{T}_j$ for Whole Tests

Test	Prop. Sig. Q	Obj. No.	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
WA	.23									
		8	.89	.20	.87	.18	.87	.18	.87	.18
		12	.53	.27	.52	.25	.53	.26	.53	.26
		14	.62	.31	.57	.28	.60	.29	.60	.29
		15	.49	.30	.48	.24	.49	.26	.49	.26
		16	.55	.31	.61	.23	.59	.26	.59	.26
		17	.76	.32	.76	.23	.75	.27	.75	.27
		18	.64	.38	.75	.22	.67	.31	.67	.31
RV	.06									
		24	.48	.30	.48	.28	.48	.28	.48	.29
		25	.47	.28	.48	.24	.48	.24	.48	.24
		28	.58	.31	.59	.24	.58	.26	.58	.26
RC	.02									
		32	.59	.30	.60	.28	.60	.29	.60	.29
		35	.64	.33	.64	.31	.64	.31	.64	.31
		36	.60	.30	.60	.27	.60	.28	.60	.28
		37	.61	.33	.61	.29	.61	.29	.61	.29
LM	.13									
		43	.58	.31	.58	.28	.58	.29	.58	.29
		44	.52	.29	.53	.26	.53	.27	.53	.27
		45	.64	.28	.64	.23	.64	.26	.64	.26
LE	.08									
		48	.79	.24	.78	.20	.78	.21	.78	.21
		49	.66	.33	.67	.25	.66	.27	.66	.27
		50	.60	.34	.57	.27	.58	.28	.58	.28
		51	.53	.33	.54	.30	.54	.31	.54	.31
		52	.69	.31	.71	.29	.70	.30	.70	.30
MC	.04									
		59	.58	.20	.57	.18	.57	.19	.58	.19
		62	.57	.25	.59	.22	.58	.23	.58	.23
MA	.08									
		73	.54	.23	.55	.19	.55	.20	.55	.21
		76	.70	.24	.69	.21	.69	.22	.69	.22
		77	.57	.26	.57	.20	.57	.21	.57	.21
		78	.80	.21	.77	.11	.77	.13	.77	.13

Table 3
Means and Standard Deviations of
w_j and $\sigma(T_j \mid x_j)$ for Whole Tests

		w_j				$\sigma(T_j \mid x_j)$			
	Obj.	Unadjusted		Adjusted		Unadjusted		Adjusted	
Test	No.	Mean	SD	Mean	SD	Mean	SD	Mean	SD
WA									
	8	.65	.40	.64	.40	.05	.06	.05	.06
	12	.63	.36	.59	.34	.06	.04	.06	.04
	14	.67	.37	.65	.37	.06	.05	.07	.05
	15	.68	.39	.67	.38	.06	.05	.06	.05
	16	.71	.39	.71	.39	.08	.06	.08	.06
	17	.66	.38	.65	.38	.06	.06	.06	.06
	18	.68	.39	.68	.39	.07	.06	.07	.06
RV									
	24	.62	.19	.37	.17	.05	.02	.07	.02
	25	.82	.22	.81	.22	.07	.03	.07	.03
	28	.75	.27	.73	.28	.08	.04	.08	.04
RC									
	32	.72	.10	.61	.09	.06	.02	.07	.02
	35	.78	.16	.74	.17	.07	.03	.07	.03
	36	.84	.11	.81	.11	.06	.02	.07	.02
	37	.86	.12	.83	.12	.06	.02	.06	.02
LM									
	43	.65	.27	.58	.25	.06	.04	.07	.04
	44	.68	.28	.61	.27	.06	.03	.07	.03
	45	.64	.29	.57	.28	.07	.03	.07	.04
LE									
	48	.80	.25	.78	.24	.05	.04	.06	.04
	49	.80	.24	.78	.24	.06	.04	.07	.04
	50	.82	.25	.80	.25	.06	.03	.06	.03
	51	.74	.26	.70	.26	.06	.04	.07	.04
	52	.70	.24	.64	.24	.06	.04	.07	.04
MC									
	59	.72	.15	.58	.14	.07	.01	.08	.01
	62	.67	.15	.51	.14	.07	.02	.09	.02
MA									
	73	.68	.21	.56	.18	.06	.02	.07	.01
	76	.68	.21	.61	.19	.07	.03	.08	.03
	77	.80	.23	.78	.23	.07	.03	.08	.03
	78	.88	.26	.87	.26	.05	.04	.05	.04

Table 4 contains the means, standard deviations, and correlations of x_j / n_j and $\hat{T}_j$ by objective for the half tests, as well as the proportion of examinees with significant *Q* values. The means for the halves are quite similar, indicating the successful selection of halves of similar difficulty, except for the first Mathematics Computation objective. As expected, the proportion of significant *Q* values is lower for halves than the whole tests; thus, more examinees have prior information incorporated into their OPI scores for the test halves than for the whole tests.

Table 4
Proportions of Significant Q Values and
Means, Standard Deviations, and Correlations
of x_j / n_j and $\hat{T}_j$ for Half Tests

	Prop. Sig. Q			x_j / n_j					$\hat{T}_j$				
				Mean		SD			Mean		SD		
Test	I	II	Obj. No.	I	II	I	II	r	I	II	I	II	r
WA	.06	.02											
			12	.53	.53	.31	.28	.68	.53	.52	.26	.24	.91
RV	.00	.00											
			24	.50	.46	.31	.31	.84	.50	.47	.28	.28	.88
RC	.01	.03											
			32	.59	.59	.31	.33	.75	.60	.60	.27	.30	.90
LM	.07	.09											
			43	.57	.58	.32	.34	.75	.59	.59	.27	.29	.87
			44	.51	.52	.32	.32	.67	.54	.52	.25	.27	.90
			45	.65	.64	.31	.31	.72	.64	.64	.24	.22	.83
LE	.05	.08											
			52	.69	.70	.33	.34	.73	.70	.72	.29	.30	.87
MC	.02	.02											
			59	.61	.55	.21	.25	.50	.60	.56	.17	.22	.70
			62	.58	.57	.29	.26	.65	.60	.58	.24	.22	.71
MA	.06	.04											
			73	.54	.53	.25	.24	.70	.56	.55	.20	.20	.81
			76	.71	.68	.26	.28	.59	.70	.68	.22	.22	.77

The means, standard deviations, and correlations across halves of w_j, $\tilde{T}_j$, and $\sigma(T_j \mid x_j)$ for the unadjusted and adjusted methods are in Tables 5 to 7. The adjustment makes the expected reduction in w_j and increase in $\sigma(T_j \mid x_j)$.

Table 5
Means, Standard Deviations, and Correlations
of w_j for Half Tests

		Unadjusted					Adjusted				
	Obj.	Mean		SD			Mean		SD		
Test	No.	I	II	I	II	r	I	II	I	II	r
WA											
	12	.76	.82	.23	.15	.17	.72	.77	.23	.16	.23
RV											
	24	.67	.63	.14	.06	.34	.41	.34	.21	.06	.39
RC											
	32	.71	.73	.09	.14	-.02	.59	.62	.09	.13	.00
LM											
	43	.71	.68	.21	.23	.11	.63	.60	.20	.23	.13
	44	.74	.69	.23	.23	.15	.67	.62	.23	.22	.21
	45	.65	.68	.26	.26	.35	.57	.60	.27	.26	.42
LE											
	52	.72	.69	.21	.25	.26	.66	.63	.22	.26	.34
MC											
	59	.75	.69	.18	.12	.07	.61	.52	.20	.10	.06
	62	.65	.70	.17	.13	.07	.48	.53	.20	.13	.08
MA											
	73	.67	.71	.20	.16	.29	.55	.58	.18	.14	.40
	76	.67	.71	.19	.16	.18	.60	.64	.18	.16	.22

Table 6
Means, Standard Deviations, and Correlations
of $\tilde{T}_j$ for Half Tests

		Unadjusted					Adjusted				
	Obj.	Mean		SD			Mean		SD		
Test	No.	I	II	I	II	r	I	II	I	II	r
WA											
	12	.53	.53	.26	.24	.82	.53	.53	.27	.24	.82
RV											
	24	.50	.47	.28	.29	.88	.50	.47	.29	.30	.86
RC											
	32	.60	.60	.28	.30	.86	.60	.60	.28	.31	.85
LM											
	43	.58	.59	.28	.30	.82	.58	.59	.29	.30	.82
	44	.54	.53	.27	.29	.78	.54	.52	.27	.29	.78
	45	.64	.63	.27	.26	.78	.64	.63	.27	.27	.78
LE											
	52	.69	.70	.30	.31	.80	.69	.70	.31	.32	.79
MC											
	59	.60	.55	.19	.23	.63	.60	.55	.19	.23	.61
	62	.59	.58	.25	.23	.71	.59	.58	.26	.24	.70
MA											
	73	.56	.55	.21	.21	.77	.56	.54	.22	.21	.77
	76	.69	.68	.23	.24	.71	.70	.68	.23	.24	.70

Table 7
Means, Standard Deviations, and Correlations
of $\sigma(T_j \mid x_j)$ for Half Tests

		Unadjusted					Adjusted				
	Obj.	Mean		SD			Mean		SD		
Test	No.	I	II	I	II	r	I	II	I	II	r
WA											
	12	.07	.07	.04	.03	.29	.08	.08	.05	.03	.34
RV											
	24	.07	.08	.03	.02	.44	.09	.10	.04	.03	.46
RC											
	32	.08	.07	.03	.03	.33	.09	.09	.03	.03	.36
LM											
	43	.08	.09	.04	.05	.41	.10	.10	.05	.05	.45
	44	.07	.08	.04	.04	.33	.08	.10	.04	.04	.37
	45	.09	.08	.05	.04	.57	.10	.09	.05	.05	.56
LE											
	52	.08	.08	.06	.06	.54	.09	.09	.06	.07	.57
MC											
	59	.08	.10	.02	.02	.13	.10	.12	.03	.03	.18
	62	.10	.09	.04	.03	.33	.11	.11	.04	.03	.36
MA											
	73	.08	.08	.02	.02	.30	.10	.10	.02	.02	.35
	76	.09	.09	.04	.04	.47	.10	.10	.05	.04	.51

The percents of overlapping credibility intervals and the predicted posterior standard deviations and observed standard errors are in Table 8. The unadjusted procedure produces percents overlap from 65 to 75, which is generally lower than the range expected from the normal approximation; the observed standard errors of the OPI values are greater than the predicted posterior standard deviations by .02 to .04.

The adjusted procedure produces percents overlap from 72 to 79, which is in the range expected from the normal approximation; the observed standard errors are greater than the predicted posterior standard deviations by .00 to .03.

Table 8
Proportion of Examinees with Overlapping Credibility Intervals,
Observed OPI Standard Errors, and
Predicted Posterior Standard Deviations

	Obj.	Proportion Overlap		Unadjusted		Adjusted	
Test	No.	Unadjusted	Adjusted	SE	$\sigma(T_j \mid x_j)$	SE	$\sigma(T_j \mid x_j)$
WA							
	12	.74	.78	.11	.08	.11	.08
RV							
	24	.75	.74	.10	.08	.11	.10
RC							
	32	.65	.73	.11	.08	.11	.10
LM							
	43	.74	.76	.12	.10	.13	.11
	44	.69	.73	.13	.09	.13	.10
	45	.70	.72	.12	.10	.13	.11
LE							
	52	.70	.73	.14	.10	.14	.11
MC							
	59	.65	.75	.13	.09	.14	.11
	62	.67	.75	.13	.10	.14	.12
MA							
	73	.73	.79	.10	.08	.10	.10
	76	.72	.75	.13	.10	.13	.11

Conclusions

The unadjusted OPI procedure overestimates the amount of independent information provided by the prior distribution and underestimates the length of the 67% credibility interval. However, it should be noted that the absolute magnitude of this underestimation is small and would be reflected in the addition on the average of roughly 1 or 2 print positions (about 1 or 2 millimeters) at the ends of each credibility interval. Such a lengthening of the interval would seem unlikely to have much of an effect on a teacher's judgment about a student's knowledge of an objective.

The adjusted OPI procedure produces credibility intervals that are in line with expectations based on the normal approximation. Use of the adjusted procedure would produce credibility intervals that would be on the average less than one print position too short.

References

CTB/McGraw-Hill (1985). *California achievement tests, Form E*. Monterey, CA: Author.

CTB/McGraw-Hill (1986). *California achievement tests, Forms E and F, Technical bulletin 2.* Monterey, CA: Author.

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

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

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

Yen, W. M. (1982). *Use of three-parameter item response theory in the development of CTBS, Form U and TCS.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21,* 93-111.