Exploiting Auxiliary Information About Examinees in the Estimation of Item Parameters

Robert J. Mislevy
Educational Testing Service

The precision of item parameter estimates can be increased by taking advantage of dependencies between the latent proficiency variable and auxiliary examinee variables such as age, courses taken, and years of schooling. Gains roughly equivalent to two to six

additional item responses can be expected in typical educational and psychological applications. Empirical Bayesian computational procedures are presented and illustrated with data from the National Assessment of Educational Progress survey.

A pervasive problem in item response theory (IRT) is the difficulty of simultaneously estimating large numbers of parameters from limited data. Even large samples of examinees may not eliminate the problem when each examinee responds to only a few items, as in educational assessment and adaptive testing. Certain improvements are obtained by using hierarchical models in the manner proposed by Lindley and Smith (1972); treating examinee parameters as a sample from a common population enhances the stability and precision of both item parameter and examinee parameter estimates. This approach has been applied to IRT by a number of researchers in recent years, including Bock and Aitkin (1981), Leonard and Novick (1985), Rigdon and Tsutakawa (1982), and Swaminathan and Gifford (1982).

For the most part, these authors considered all examinees to be members of a single, undifferentiated population. This framework gives rise to such beliefs as "If the parameters of most examinees seem to lie between -3 and $+3$, then the parameter of an examinee who answered both of two difficult math items correctly is probably somewhere between 1.5 and 3.5 —even though his/her maximum likelihood estimate is $+\infty$." Additional stability and precision may be achieved if auxiliary information is available about examinees, such as educational background or status on demographic variables. This would allow more useful inferences such as "The parameter of an examinee who answered both of two difficult math items correctly *and* studied calculus in college is probably between 2.7 and 3.7 ."

This paper addresses the use of auxiliary information about examinees in estimating item parameters. The following section reviews item parameter estimation when examinee parameters are known. Attention then turns to unknown proficiencies under the assumptions of (1) an undifferentiated population and (2) a population differentiated with respect to auxiliary variables. Following this are sections that discuss anticipated gains in precision, outline computational procedures, and illustrate the approach with responses to 10 items from the 1983/84 National Assessment of Educational Progress (NAEP) reading survey.

APPLIED PSYCHOLOGICAL MEASUREMENT

Vol. 11, No. 1, March 1987, pp. 81–91

© Copyright 1987 Applied Psychological Measurement Inc.

0146-6216/87/010081-11\$1.80

The Role of Auxiliary Information

The relevance of auxiliary examinee variables to item parameter estimation is not immediately obvious, as they play no role in the basic model for item responses. Letting $\mathbf{x}_i = (x_{i1}, \dots, x_{in})$ represent the responses of examinee i to n test items and y_i represent values of auxiliary variables such as educational and demographic status, the standard IRT assumption of local independence states that

$$p(\mathbf{x}_i | \theta_i, y_i, \boldsymbol{\beta}) = \prod_j^n p(x_{ij} | \theta_i, \beta_j) \quad , \quad (1)$$

where θ_i is the examinee parameter, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_n)$ are possibly vector-valued item parameters, and the form of $p(x_{ij} | \theta_i, \beta_j)$ is specified a priori through the item response model. It follows that y_i would indeed be irrelevant to item parameter estimation if θ_i were known. The likelihood to be maximized with respect to $\boldsymbol{\beta}$, given the data matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ of responses from N examinees with proficiencies $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)$ and auxiliary variables $\mathbf{Y} = (y_1, \dots, y_N)$, would be simply

$$L = \prod_i^N p(\mathbf{x}_i | \theta_i, \boldsymbol{\beta}) \quad . \quad (2)$$

The maximum likelihood estimate (MLE) $\hat{\boldsymbol{\beta}}$ would satisfy the likelihood equation

$$\boldsymbol{\theta} = \sum_i \partial \ell_i(\theta_i) / \partial \boldsymbol{\beta} \quad , \quad (3)$$

where

$$\ell_i(\theta) = \log p(\mathbf{x}_i | \theta, \boldsymbol{\beta}) \quad , \quad (4)$$

and the covariance matrix of estimation error variances for $\hat{\boldsymbol{\beta}}$, or $\text{Var}_{\theta, \mathbf{x}}(\hat{\boldsymbol{\beta}})$, could be approximated by the inverse of the observed information matrix $\mathbf{I}_{\theta, \mathbf{x}}$:

$$\mathbf{I}_{\theta, \mathbf{x}} = \sum_i \left[\frac{\partial \ell_i(\theta_i)}{\partial \boldsymbol{\beta}} \right] \left[\frac{\partial \ell_i(\theta_i)}{\partial \boldsymbol{\beta}'} \right] \bigg|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} \quad . \quad (5)$$

The square roots of the diagonal of $\text{Var}_{\theta, \mathbf{x}}(\hat{\boldsymbol{\beta}})$ are standard errors for the corresponding elements of $\hat{\boldsymbol{\beta}}$. The subscript $(\theta, \mathbf{x})$ indicates variance or information with respect to repeated samples in which observations consist of values of θ and $\mathbf{x}$. Note that local independence implies $\text{Var}_{\theta, \mathbf{x}}(\hat{\boldsymbol{\beta}}) = \text{Var}_{\theta, \mathbf{x}}(\hat{\boldsymbol{\beta}})$.

But these results are conditioned on θ , and θ is not known in practice. The problem that must actually be solved is to maximize the expectation of Equation 2, or the marginal likelihood

$$L_M = \prod_i \int p(\mathbf{x}_i | \theta, \boldsymbol{\beta}) \partial F_i(\theta) \quad , \quad (6)$$

where $F_i(\theta)$ is the distribution of the unknown proficiency of examinee i (de Leeuw & Verhelst, 1984). This is an "incomplete data" problem, in the terminology of Dempster, Laird, and Rubin (1977); uncertainty about θ distinguishes it from the "complete data" problem of maximizing Equation 2. Assuming that the required integrals exist, the likelihood equations under marginal maximum likelihood (MML) of $\boldsymbol{\beta}$ are obtained as

$$\boldsymbol{\theta} = \sum_i p^{-1}(\mathbf{x}_i) \int [\partial \ell_i(\theta) / \partial \boldsymbol{\beta}] \partial F_i(\theta) \quad , \quad (7)$$

where

$$p(\mathbf{x}_i) = \int \ell_i(\theta) \partial F_i(\theta) \quad . \quad (8)$$

A number of different choices for the distribution functions F_i appear in the literature on item parameter estimation. Positing unique and unconstrained F_i s for each examinee, then maximizing Equation

6 with respect to β and each individual θ_i , gives Birnbaum's (1968) joint maximum likelihood solution. It uses only information from x_i to reduce uncertainty about θ_i . Positing a common F for all examinees is tantamount to assuming that θ s are independent and identically distributed. Bock and Aitkin (1981) gave a MML solution for β when F is the normal distribution; Mislevy (1986) assumed a parametric form for F and estimated its parameters α jointly with β ; de Leeuw and Verhelst (1984) obtained a nonparametric estimate of F jointly with β . These solutions use two sources of information to reduce uncertainty about θ_i , namely x_i and membership in an undifferentiated population.

A third alternative, falling between unique, unconstrained F 's and identical F 's, is to posit distributions that depend on auxiliary variables: that is, $F_i(\theta) = F_{y_i}(\theta)$. Examinees with identical y values are considered a random sample from a population indexed by that particular value of y , and these conditional distributions are allowed to vary with y . The variance of item parameter estimates thus obtained, or $\text{Var}_{x,y}(\hat{\beta})$, is again approximated by the negative inverse of an appropriate observed information matrix, namely

$$I_{x,y} = \sum_i p^{-1}(x_i) \int \left\{ \left[\frac{\partial \ell_i(\theta)}{\partial \beta} \right] \left[\frac{\partial \ell_i(\theta)}{\partial \beta'} \right] \right\} \bigg|_{\beta=\hat{\beta}} dF_{y_i}(\theta) \quad (9)$$

How Much Can Be Gained?

How does the use of auxiliary information affect the precision of estimates of β ? Upper and lower bounds follow from general results on estimation in the presence of missing data. From Orchard and Woodbury (1972),

$$\text{Var}_x(\hat{\beta}) \geq \text{Var}_{x,y}(\hat{\beta}) \geq \text{Var}_{\theta,x,y}(\hat{\beta}) = \text{Var}_{\theta,x}(\hat{\beta}) \quad (10)$$

where $A \geq B$ means that the matrix difference $A - B$ is positive semidefinite. Thus estimates using auxiliary information will be at least as precise as those based only on response vectors and membership in an undifferentiated population, but they will never be more precise than those that would be obtained if θ values were available as well. Note that this result pertains to true variances. For specific datasets, estimated variances and standard errors for particular items, as calculated with Equation 9, can sometimes be higher for estimates that use y than those for estimates that do not.

To go beyond boundary results, some simple approximations can provide an indication of the gains that might be expected in typical educational and psychological applications. Because uncertainty about θ values can be reduced either by acquiring auxiliary information or additional item responses, the contribution of auxiliary variables can be gauged in units of additional item responses. In the special case of dichotomous items, the amount of information about a particular θ conveyed by item responses alone can be expressed as

$$I(\theta) = \sum_j \frac{P'_j(\theta)^2}{P_j(\theta)[1 - P_j(\theta)]} \quad (11)$$

where $P_j(\theta) = p(x_j = 1|\theta)$ and $P'_j(\theta) = \partial P_j(\theta)/\partial \theta$. For examinees with finite MLEs, Bayes' theorem applied with a diffuse prior leads to the approximation $p(\theta|x_i) \sim N(\hat{\theta}, \sigma_x^2)$ with $\sigma_x^2 = I^{-1}$. This follows by first rescaling the likelihood so that it integrates to 1, then using its mode and curvature at the mode in a normal approximation. I indicates precision; σ_x^2 indicates uncertainty.

Consider now the one-parameter logistic (Rasch) model, under which

$$P_j(\theta) \equiv p(x = 1|\theta, b_j) = 1/[1 + \exp(b_j - \theta)] \quad (12)$$

The contribution of item j to information about θ is $P_j(\theta)[1 - P_j(\theta)]$, and the total information from n identical items for which $b_j = \theta$ is simply $.25n$. To assess the additional impact of differentiating the

population through auxiliary variables, consider numerical results from a regression model with homoscedastic residuals. Suppose y values account for $(100 \times R^2)\%$ of the variance in a population with total variance 1.0, so that $F_y(\theta) \sim N(\mu_y, \sigma_e^2)$ with $\sigma_e^2 = 1 - R^2$. If the normalized likelihood induced by item responses is approximately $N(\hat{\theta}, \sigma_x^2)$, then knowledge about θ , from both x_i and y_i is obtained as follows:

$$p(\theta|x_i, y_i) \sim N\left[\frac{\hat{\theta}\sigma_x^{-2} + \mu_y\sigma_e^{-2}}{\sigma_x^{-2} + \sigma_e^{-2}}, (\sigma_e^{-2} + \sigma_x^{-2})^{-1}\right] \quad (13)$$

(Box & Tiao, 1973, p. 74). The key result is that posterior precision is simply the sum of precision from the two sources.

Table 1 shows, for various values of R^2 and test length n , increases in precision about θ gained from the use of y . *Absolute* increases do not vary with test length. When $R^2 = .3$, for example, the contribution from population membership and y is 1.43, as compared to the contribution of 1.00 from population membership alone. This translates to a gain equivalent to that of about two additional items. *Relative* increases do depend on test length. Two items' worth of increase translates to a 50% gain in the setting of a 4-item test, but only about 2% gain for a 128-item test. In the realistic setting of a short (16-item) test and auxiliary variables accounting for about 25% of population variance, gains in precision of 10% to 20% might be expected.

The numerical values shown in Table 1 are specific to the choice of model, the population variance, and the assumption that items are maximally informative for examinees. Any choices, however, would reveal similar basic patterns. In particular, in the range of R^2 typical in educational and psychological measurement, the proportional gains that accrue from the use of auxiliary information are most dramatic for short tests, and fade to insignificance as test length increases.

Models and Methods

This section presents two IRT models that differentiate examinees by means of auxiliary variables, and suggests computing approximations based on Bock and Aitkin's (1981) MML (empirical Bayesian) procedures. It can be shown that as long as item parameters β are not confounded with the parameters of $p(\theta|y)$, likelihood inferences about β are the same whether examinees were sampled at random from the population at large, or were sampled with stratification on y to produce the observed y distribution. The following discussion will therefore be limited to the latter, simpler, case.

Table 1
Gains in Posterior Precision for θ from Auxiliary Information

R^2	Contribution to Posterior Precision	Contribution in Item Units	Percentage Gain in Posterior Precision from Auxiliary Variables					
			n=4	n=8	n=16	n=32	n=64	n=128
0.00	1.00	4.00	—	—	—	—	—	—
0.10	1.11	4.44	8.89	7.41	5.56	3.70	2.22	1.23
0.20	1.25	5.00	20.00	16.67	12.50	8.33	5.00	2.78
0.30	1.43	5.71	34.29	28.57	21.43	14.29	8.57	4.76
0.40	1.67	6.67	53.33	44.44	33.33	22.22	13.33	7.41
0.50	2.00	8.00	80.00	66.67	50.00	33.33	20.00	11.11
0.60	2.50	10.00	120.00	100.00	75.00	50.00	30.00	16.67
0.70	3.33	13.33	186.67	155.56	116.67	77.78	46.47	25.93
0.80	5.00	20.00	320.00	266.67	200.00	133.33	80.00	44.44
0.90	10.00	40.00	720.00	600.00	450.00	300.00	180.00	100.00

Mixtures of Finite Distributions

Mislevy (1984) described a nonparametric approximation of a continuous density function of a latent variable in terms of a distribution with mass at a finite number of prespecified points. The proficiency of each examinee, or θ_i , is thus assumed to take one of only Q known values. The latent trait problem is thereby replaced by an analogous "latent class" problem that is easier to solve. A single population was addressed in that presentation, and item parameters were assumed known. Extensions to the simultaneous estimation of item parameters, and to multiple subpopulations indexed by an auxiliary variable y , will now be considered. This approach provides considerable flexibility in the distributions $F_i(\theta) = F_{y_i}(\theta)$. It lends itself well to discrete auxiliary variables with relatively few values.

It proves convenient to write such an auxiliary variable as a vector of 0-1 indicators. Define $y_i = (y_{i1}, \dots, y_{iK})$ by letting $y_{ik} = 1$ if examinee i is associated with the k th of K exhaustive and mutually exclusive subpopulations, and 0 otherwise. The probability of observing response pattern $\mathbf{x}_i$ from an examinee selected at random from a specified subpopulation is given by

$$p(\mathbf{x}_i | y_i, \beta) = \prod_k \left[\int p(x_i | \theta, \beta) dF_k(\theta) \right]^{y_{ik}}, \quad (14)$$

where F_k is the distribution in subpopulation k , that is, $F_i(\theta) = F_k(\theta)$ when $y_{ik} = 1$. This probability can be approximated by a finite distribution as

$$p(\mathbf{x}_i | y_i, \beta) \approx \prod_k \left[\sum_q p(\mathbf{x}_i | \Theta_q, \beta) W_{qk} \right]^{y_{ik}}, \quad (15)$$

where $\Theta_1, \dots, \Theta_Q$ is a grid of points specified a priori, and W_{qk} is the weight or density at point q in subpopulation k .

Let $(\mathbf{X}, \mathbf{Y})$ be the data matrix observed from a sample of $N = \sum_k N_k$ examinees selected as random subsamples stratified on y . The probability of $\mathbf{X}$ given $\mathbf{Y}$ is proportional to

$$L = \prod_i \prod_k \left[\sum_q p(\mathbf{x}_i | \Theta_q, \beta) W_{qk} \right]^{y_{ik}}, \quad (16)$$

the logarithm of which is

$$\ell = \log L = \sum_i \sum_k y_{ik} \log \sum_q p(\mathbf{x}_i | \Theta_q, \beta) W_{qk}. \quad (17)$$

Relative maxima with respect to β and W can be obtained by means of the EM algorithm, under the special case of missing indicators for a multinomial distribution (Dempster et al., 1977, Section 4.3). Using provisional estimates β^0 and W^0 , the expectation step computes the following quantities:

1. The expected number of examinees with proficiency Θ_q from a sample of size N_k from subpopulation k , conditional on $\mathbf{X}$, $\mathbf{Y}$, β^0 , and W^0 :

$$N_{qk}^1 = \sum_i y_{ik} p_k^0(\Theta_q | \mathbf{x}_i), \quad (18)$$

where

$$p_k^0(\Theta_q | \mathbf{x}_i) = \frac{p(\mathbf{x}_i | \Theta_q, \beta = \beta^0) W_{qk}^0}{\sum_r p(\mathbf{x}_i | \Theta_r, \beta = \beta^0) W_{rk}^0}, \quad (19)$$

an application of Bayes' theorem, gives the posterior probability that the proficiency of examinee i is Θ_q , given provisional parameter estimates β^0 and W^0 .

2. The expected number of correct responses to item j from examinees in subpopulation k with proficiency Θ_q , given a random sample of size N_k (again given β^0 and W^0):

$$R_{jqk} = \sum_i y_{ik} x_{ij} p_k^0(\Theta_q | x_i) \quad (20)$$

The maximization step computes what would be MLEs of β and W if N^1 and R^1 were observed quantities rather than conditional expectations. For W ,

$$W_{qk}^1 = N_{qk}^1 / N_k \quad (21)$$

For β , conditional expectations of likelihood equations are solved:

$$0 = \sum_q \frac{R_{jq+}^1 N_{q+}^1 P_j(\Theta_q)}{P_j(\Theta_q) [1 - P_j(\Theta_q)]} \left[\frac{\partial P_j(\Theta_q)}{\partial \beta} \right] \quad (22)$$

where $R_{jq+}^1 = \sum_k R_{jqk}^1$ and N_{q+}^1 is similarly defined. Under the two-parameter logistic model, for example, Equation 22 takes the following forms:

$$a_j: 0 = \sum_q [R_{jq+}^1 - N_{q+}^1 P_j(\Theta_q)] (\Theta_q - b_j) \quad (23)$$

$$b_j: 0 = \sum_q [R_{jq+}^1 - N_{q+}^1 P_j(\Theta_q)] a_j \quad (24)$$

In principle, the linear indeterminacy in the one-, two-, and three-parameter logistic and normal IRT models presents no impediment to the EM algorithm, which readily converges to one of the infinitely many solutions on a ridge. Numerical stability and the quality of the finite characterization of F are enhanced, however, by controlling the scaling of the solution at this point. One convenient way of doing this is to standardize the weighted average distribution. The points Θ_q have been referred to as specified a priori; given the linear indeterminacy, only their relative spacing may be thought of as prespecified. After an EM cycle, then, the points may be rescaled as follows:

$$\Theta_q = (\Theta_q - \bar{\Theta}) / s \quad (25)$$

where

$$\bar{\Theta} = N^{-1} \sum_k N_k \sum_q \Theta_q W_{qk}^1 \quad (26)$$

and

$$s = N^{-1} \sum_k N_k \sum_q (\Theta_q - \bar{\Theta})^2 W_{qk}^1 \quad (27)$$

Item parameters are adjusted accordingly. Under the two- and three-parameter models, $\hat{b}_j$ is replaced by $(\hat{b}_j - \bar{\Theta})/s$ and a_j is replaced by sa_j . Under one-parameter models, rescaling takes place only with respect to $\bar{\Theta}$.

Iterations continue until convergence is achieved. Iteration from several starting values helps to verify whether a given solution is indeed a global maximum. The observed information matrix for the item parameter estimates can then be approximated using Equation 9. Employing Louis's (1982) simplifications for "missing multinomial indicators" problems, the expression

$$I_{x,y}(\hat{\beta}) \approx \sum_i \sum_k y_{ik} \sum_q \left[\frac{\partial \ell_i(\Theta_q)}{\partial \beta} \right] \left[\frac{\partial \ell_i(\Theta_q)}{\partial \beta'} \right] \hat{p}_k(\Theta_q | x_i) \quad (28)$$

is obtained, where $\hat{p}_k(\Theta_q | x_i)$ is evaluated at the final estimates at $\hat{\beta}$ and $\hat{W}$.

A Linear Model

The unrestricted mixture solution described above becomes unwieldy as the number of potential

values of the auxiliary variable increases. The more structured alternative of a linear model for $p(\theta|y)$ is suitable when y is vector-valued or is continuous rather than discrete. Assuming residuals to be homoscedastic and normal, then

$$\theta \sim N(y'\alpha, \sigma^2) \quad , \quad (29)$$

where auxiliary variables are coded so that the K columns of

$$Y = (y_1, \dots, y_K)' \quad , \quad (30)$$

which are basis vectors for the K elements of the regression parameter α , are linearly independent. They may include values on measured variables such as previous test scores, and dummy regression variables that encode selected contrasts among categorical auxiliary variables.

Maximum likelihood solutions for α and σ^2 in the special case of structured means for the cells of a multi-way design have been given by Mislevy (1985) under the assumption that item parameters are known, and by Zwarts and Veldhuesen (1985) under the assumption that $p(x|\theta)$ is the Rasch model with unknown item parameters to be estimated jointly. These solutions are readily extended to the case of a general IRT model with unknown item parameters. This section describes an approximation over a grid of prespecified points so that computation is similar to the nonparametric solution described above. Attention is focused for convenience upon the one-, two-, and three-parameter logistic and normal IRT models.

The linear indeterminacies of these models are again conveniently resolved by restrictions on the population parameters. First, σ^2 may be fixed at unity without loss of generality to set the unit size of the scale. (For one-parameter models, a slope parameter that is common over items is then estimated.) Second, the origin may be set by centering the elements of each column of Y at zero. All effects are thus cast as deviations around a grand mean of zero. This restriction, in conjunction with the independence of the basis vectors, completes the resolution of the scale.

The marginal likelihood for a sample of size N is written as

$$L = \prod_i \int p(x_i|\theta, \beta) \phi(\theta - y_i'\alpha) d\theta \quad , \quad (31)$$

where ϕ represents the standard normal density function. Approximation over a finite grid of points is accomplished by

$$L^* = \prod_i \sum_q p(x_i|\Theta_q, \beta) W_q(\alpha) \quad , \quad (32)$$

where

$$W_q(\alpha) = \frac{\exp[-(\Theta_q - y_i'\alpha)^2/2]}{\sum_r \exp[-(\Theta_r - y_i'\alpha)^2/2]} \quad . \quad (33)$$

The weights W play the same role as those in the preceding solution. The difference is that they are no longer estimated without restriction, but are modeled as functions of the effect parameters α .

MML estimation can again proceed in EM cycles that solve the likelihood equations. Let β^0 and α^0 be provisional estimates. Again the E-step computes expected counts of examinees and correct responses at each point:

$$N_q^1 = \sum_i P(\Theta_q|x_i, \beta^0, \alpha^0) \quad (34)$$

and

$$R_{jq}^1 = \sum_i x_{ij} P(\Theta_q|x_i, \beta^0, \alpha^0) \quad (35)$$

where

$$P(\Theta_q | \mathbf{x}_i, \beta^0, \alpha^0) = \frac{p(\mathbf{x}_i | \Theta_q, \beta^0) W_{qi}(\alpha^0)}{\sum_r p(\mathbf{x}_i | \Theta_r, \beta^0) W_{ri}(\alpha^0)} \quad (36)$$

It also computes the conditional expected value of each examinee's proficiency:

$$\theta_i^1 = \sum_q \Theta_q P(\Theta_q | \mathbf{x}_i, \beta^0, \alpha^0) \quad (37)$$

The M-step pseudo-likelihood equations for item parameters can be written as in Equation 22. The equations for α simplify to

$$\alpha^1 = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\boldsymbol{\theta}^1, \quad (38)$$

where $\boldsymbol{\theta}^1 = (\theta_1^1, \dots, \theta_N^1)$. The posterior information matrix for β can again be approximated using Equation 28.

Numerical Example

This section illustrates the procedures described above. The data are responses to 10 items from the 1983/84 NAEP (National Assessment of Educational Progress, 1986) survey of reading. The 1983/84 survey addressed performance levels in three subpopulations of students: Age 9/Grade 4, Age 13/Grade 8, and Age 17/Grade 11. Under the NAEP item-sampling design, 5 of the items were administered to a sample from the first of these subpopulations, and all 10 items were administered to samples from the latter two. When only examinees who responded to all the items they were presented are considered, the resulting sample sizes are 900, 1,087, and 1,159 examinees, respectively. Their item proportions-correct are shown in Table 2. As might be expected, proportion correct increases with age.

Item parameters for the three-parameter logistic model were estimated from these data in two ways. The first used Mislevy and Bock's (1982) BILOG computer program to obtain estimates that are essentially MML, except with the imposition of mild prior distributions to ensure finite values. The distribution of examinee proficiency was estimated as a histogram over 10 points, without differentiation by subpopulation membership. The second method used a modification of BILOG to obtain estimates that were similar in all respects, except that examinees were distinguished by subpopulations in the manner of the mixture solution described above. The unit size and origin of the scale were set in both runs by standardizing the sample as a whole.

Table 2
Proportion Correct for Each Item
by Age/Grade Group

Item	Age/Grade			Total
	9/4	13/8	17/11	
1	---	.78	.85	.82
2	---	.82	.93	.88
3	---	.66	.83	.75
4	---	.58	.67	.63
5	---	.82	.92	.87
6	.27	.34	.54	.40
7	.61	.64	.66	.64
8	.11	.11	.26	.17
9	.61	.87	.94	.83
10	.42	.79	.91	.74

The differentiated-population solution yielded subpopulation means of $-.932$, $.071$, and $.654$; corresponding standard deviations were $.712$, $.727$, and $.823$. Subpopulation membership thus accounted for roughly 40% of the variation among examinees.

Item parameter estimates are shown in Table 3. Estimates of a and c parameters vary little in the two solutions, with differences generally less than twice the standard error of the undifferentiated solution. The same is true for b estimates of the last five items, which were administered to all examinees. More

Table 3
Precision Gain in Item Parameter Estimates

Item	Undifferentiated Population		Differentiated Population		Percentage Gain in Precision
	$\hat{a}$	S.E.	$\hat{a}$	S.E.	
1	0.511	0.049	0.588	0.055	-20.628
2	2.328	0.245	2.585	0.281	-23.981
3	1.419	0.146	1.704	0.139	10.326
4	0.820	0.069	0.842	0.059	36.771
5	1.432	0.131	1.676	0.145	-18.378
6	1.056	0.082	1.234	0.078	10.519
7	0.338	0.033	0.322	0.131	13.319
8	2.186	0.147	2.939	0.134	20.344
9	1.529	0.126	1.442	0.094	79.674
10	1.402	0.127	1.605	0.118	15.836
Average					12.380

Item	$\hat{b}$	S.E.	$\hat{b}$	S.E.	Percent
1	-1.149	0.165	-0.842	0.132	56.250
2	-0.614	0.056	-0.383	0.049	30.612
3	-0.139	0.056	0.051	0.038	117.175
4	0.109	0.059	0.210	0.053	23.923
5	-0.756	0.082	-0.482	0.062	74.922
6	1.033	0.061	1.074	0.053	32.467
7	0.047	0.119	0.257	0.129	-14.903
8	1.716	0.075	1.671	0.062	46.332
9	-0.746	0.069	-0.820	0.064	16.235
10	-0.361	0.055	-0.365	0.046	42.957
Average					42.597

Item	$\hat{c}$	S.E.	$\hat{c}$	S.E.	Precision
1	0.220	0.044	0.213	0.044	.000
2	0.232	0.041	0.223	0.043	-9.086
3	0.238	0.029	0.240	0.027	15.364
4	0.143	0.027	0.120	0.026	7.840
5	0.193	0.046	0.205	0.044	9.298
6	0.180	0.013	0.205	0.012	17.361
7	0.161	0.026	0.203	0.025	8.160
8	0.126	0.008	0.123	0.007	30.612
9	0.237	0.037	0.208	0.038	-5.194
10	0.181	0.028	0.180	0.027	7.545
Average					8.190

noticeable differences appear only in the b estimates of the first five items, which were administered only to the two older subpopulations.

For these five items the b estimates from the differentiated solution exceed those of the undifferentiated solution by about three standard errors on the average, always in the direction of making the item seem more difficult. The explanation lies in the exploitation of subpopulation membership. Loosely speaking, low examinee proficiency and high item difficulty can both lead to lower probabilities of correct response; knowing that the examinees who are responding incorrectly to these items belong to the higher-proficiency subpopulations supports the belief that higher item difficulty is the cause.

Also shown in Table 3 are percentage gains in efficiency, or ratios of squared standard errors. Most of the values are positive; they average around 10% for a and c parameters, and around 40% for b s. While gains for c s are similar for both subsets of items, interesting differences appear for a and b gains, again explicable in terms of the specific contribution of the auxiliary information. For b s, for the reason noted above, information gain is more dramatic for those items not presented to the young examinees—about 60% on the average, as opposed to 30%. For a s, however, the finding is reversed—near zero as opposed to 30%. This is because examinees above and below an item's b parameter contribute more information about its a value than examinees near its b , and the undifferentiated solution, which tends to overestimate the likelihood that examinees were of low ability, proceeds as if the sample contained more low-ability examinees.

Discussion

This paper has outlined procedures for incorporating auxiliary information about examinees into the IRT framework. Enhancing the precision of item parameter estimates was the primary focus. The increase in information about item parameters in typical educational and psychological settings can be expected to lie in the range of two to six items. The numerical example suggests that the increase will vary as a function of item parameter type and data-gathering design.

The expected increase is modest, to be sure, but in many applications it is free in the sense that it is already available for use. Because its incremental value decreases for longer tests, auxiliary information would be most useful in settings where relatively few responses are solicited from each examinee. This would include two applications of great current interest, namely educational assessment and adaptive testing. In assessment, data that are sparse at the level of individuals—say, 5 or 10 items in a given scale—yield more efficient estimates of population parameters for a given total number of item responses. In adaptive testing, new items are calibrated using joint response patterns with previously calibrated items while the number of old items is held to minimally acceptable levels—as few as, say, 15.

A side issue in the present paper, but a fundamentally important result, is that when examinees are indeed a random sample from a well-defined population, the estimated population distributions and effect parameters are consistent within the limits of precision afforded by the numerical approximations. (See Mislevy, 1984, 1985, on population estimation when item parameters are known.) This stands in contrast to the asymptotically biased results obtained by using the distribution of $\hat{\theta}$ to approximate the distribution of θ . In fact, the discrepancy between the two distributions is largest in exactly those cases in which the present procedures offer the most benefit for item parameter estimation, namely short tests.

Finally, it is implicit in the foregoing discussion that auxiliary information about examinees can lead to improved estimates of individual proficiencies. Whether estimates that are improved in the sense of minimum mean squared error are unequivocally “better” for all applications is not clear, however. The present author has avoided advocating the use of auxiliary information in obtaining examinee scores when tests are used as contests (i.e., when important placement or selection decisions are made for individual examinees), because it would seem that in these situations the tester ought to gather enough data directly

dependent upon proficiency (i.e., item responses) to make satisfactorily precise decisions on that strength alone. In adaptive testing, for example, the use of auxiliary information to improve item parameter estimation would be recommended, but not its use to estimate scores that will be used to compare individual examinees.

References

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading MA: Addison-Wesley.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, 46, 443-459.
- Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading MA: Addison-Wesley.
- de Leeuw, J., & Verhelst, N. (1984). *Maximum likelihood estimation in generalized Rasch models* (Department of Data Theory Research Report RR-84-11). Leiden, The Netherlands: University of Leiden.
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from complete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39, 1-38.
- Leonard, T., & Novick, M. R. (1985). *Bayesian inference and diagnostics for the three-parameter logistic model* (ONR Technical Report 85-5). Iowa City IA: CADA Research Group, University of Iowa.
- Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34, 1-41.
- Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 44, 226-233.
- Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, 49, 359-381.
- Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, 80, 993-997.
- Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, 51, 177-195.
- Mislevy, R. J., & Bock, R. D. (1982). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville IN: Scientific Software.
- National Assessment of Educational Progress. (1986). *1983-84 public-use data tapes version 3.1 user's guide*. Princeton NJ: Educational Testing Service.
- Orchard, T., & Woodbury, M. A. (1972). A missing information principle: Theory and applications. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 697-715.
- Rigdon, S. E., & Tsutakawa, R. K. (1982). Parameter estimation in latent trait models. *Psychometrika*, 48, 567-574.
- Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, 9, 23-80.
- Zwarts, M., & Veldhuesen, N. (1985). *Maximum marginal likelihood estimation in latent trait models*. Paper presented at the European meeting of the Psychometric Society, Cambridge, England.

Acknowledgments

This work was supported by Contract No. N00014-85-K-0683, project designation NR 150-539, from the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research. The author is grateful to Kathleen Sheehan and Martha Stocking for their comments and suggestions.

Author's Address

Send requests for reprints or further information to Robert J. Mislevy, Educational Testing Service, Princeton NJ 08541, U.S.A.