Exploiting Auxiliary Information About Items in the Estimation of Rasch Item Difficulty Parameters

Robert J. Mislevy
Educational Testing Service

Standard procedures for estimating the item parameters in IRT models make no use of auxiliary information about test items, such as their format, their content, or the skills they require for solution. This paper describes a framework for exploiting this information, thereby enhancing the precision and stability of item parameter estimates and providing diagnostic information about items' operating characteristics. The principles are illustrated in a context for which a relatively simple approximation is available: empirical Bayesian estimation of Rasch item difficulty parameters.
Index terms: Bayesian estimation, Collateral information, Empirical Bayesian estimation, Exchangeability, Hierarchical models, Item response theory, Linear logistic test model, Rasch model item parameters.

Two active lines of research in item response theory (IRT) incorporate additional information into the process of parameter estimation, augmenting the information conveyed by item responses alone. One line, motivated by statistical considerations, uses Bayesian procedures to obtain more accurate estimates of item and examinee parameters. Enhanced stability and lower mean squared errors can be achieved by assuming exchangeability over item parameters of a given type (e.g., difficulty parameters), effectively shrinking estimates toward their mean in inverse proportion to the degree of information directly available about the parameters (Mislevy, 1986; Swaminathan & Gifford, 1982, 1985). A second line, motivated by psychological considerations, incorporates theories about specific skills or subtasks required to answer an item correctly. Scheiblechner (1972) and Fischer's (1973) linear logistic test model (LLTM) is a prime example; Rasch-model item difficulty parameters are cast as linear combinations of more basic parameters that reflect the contributions of psychologically salient features of each item.

This paper represents a confluence of these two lines of research. The idea is to embed the LLTM in a Bayesian framework, maintaining the notion that item features may indeed reveal something about item parameters (but admitting that they may not reveal everything). Final item parameter estimates are a compromise between LLTM estimates, where items with identical features would have identical estimates, and unrestricted maximum likelihood estimates (MLEs).

In order to focus on concepts rather than numerical procedures, a context is required for which a relatively simple approximation is available. The Rasch one-parameter IRT model for dichotomous items

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 3, September 1988, pp. 281–296

0146-6216/88/030281-16$2.05

is assumed; a linear regression model with normal, homoscedastic residuals is posited for item parameters given their salient features; and, with what is commonly called an empirical Bayesian approximation, final item parameter estimates are calculated with MLEs of the regression model treated as known. The result is a simplified version of Smith's (1973) linear model with response-surface prior distributions. The procedures are illustrated with data from a fractions test for junior high school students. Precision gains and diagnostic uses of the approach are discussed.

Background

The Rasch Model

Let x_{ij} denote the response of examinee i to item j, taking the value 1 if correct and 0 if not. The Rasch model (Rasch, 1960/1980) gives the probability of a correct response as

$$P_j(\theta_i) = P(x_{ij} = 1|\theta_i,\beta_j) = \frac{\exp(\theta_i-\beta_j)}{1+\exp(\theta_i-\beta_j)} \quad , \tag{1}$$

where β_j characterizes the difficulty of item j, and θ_i characterizes the ability of examinee i. Under the usual assumption of local independence, the probability of a vector pattern $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})'$ of responses to n items is

$$P(\mathbf{x}_i|\theta_i,\boldsymbol{\beta}) = \prod_j P_j(\theta_i)^{x_{ij}} Q_j(\theta_i)^{1-x_{ij}} \quad , \tag{2}$$

where $Q_j(\theta) = 1 - P_j(\theta)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)'$. Assuming the independence of responses over examinees, the probability of the data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)'$ of N examinees is the product of expressions like Equation 2:

$$P(\mathbf{X}|\boldsymbol{\theta},\boldsymbol{\beta}) = \prod_i P(\mathbf{x}_i|\theta_i,\boldsymbol{\beta}) \quad . \tag{3}$$

Once $\mathbf{X}$ has been observed, Equation 3 is interpreted as a likelihood function, and provides a basis for estimating parameters. The literature offers a number of alternative procedures for doing so, including

1. Joint maximum likelihood (JML), which finds values of $\boldsymbol{\beta}$ and each θ that, taken together, maximize Equation 3 (Wright & Panchapakesan, 1969);
2. Conditional maximum likelihood (CML), which finds the maximizing value of $\boldsymbol{\beta}$ given examinees' total scores (Andersen, 1973); and
3. Marginal maximum likelihood (MML), which finds the maximizing value of $\boldsymbol{\beta}$ after averaging over a distribution of examinee parameters (Bock & Aitkin, 1981; Thissen, 1982).

These solutions provide similar estimates of $\boldsymbol{\beta}$ when neither the number of items nor the number of examinees is small; under appropriate assumptions they are asymptotically equivalent, consistent, and multivariate normal (for details see Haberman, 1977, on CML and JML, and de Leeuw & Verhelst, 1986, on CML and MML).

The normal approximation to MML will be used below. The MML likelihood function is obtained from Equation 3 by averaging over (or "marginalizing with respect to") the examinee distribution:

$$L_M(\boldsymbol{\beta}|\mathbf{X}) = \prod_i \int P(\mathbf{x}_i|\theta,\boldsymbol{\beta}) p(\theta)\, \partial\theta \quad , \tag{4}$$

where $p(\theta)$, the density function for examinee parameters, may be specified a priori (as in Bock & Aitkin, 1981; Thissen, 1982) or estimated from the data (as in Cressie & Holland, 1983). When the numbers of both items and examinees are large, the likelihood function is approximately a product over items of

independent normal distributions:

$$L_M(\boldsymbol{\beta}|\mathbf{X}) \propto \prod_j \exp\left[\frac{-(\beta_j-\hat{\beta}_j)^2}{2\hat{\sigma}_j^2}\right] \quad , \tag{5}$$

where $\hat{\boldsymbol{\beta}}$ are MML estimates and $\hat{\sigma}_j$ are their estimated standard errors. (Large N is sufficient for multivariate normality, but large n is also necessary for independence.)

Bayesian Estimation of Item Parameters

The simultaneous estimation of many parameters can often be improved when it is reasonable to consider subsets of parameters as exchangeable members of corresponding populations (Efron & Morris, 1975; Lindley & Smith, 1972). The subjective notion that parameters are "in some sense similar" implies a correlational structure on prior beliefs, which can be formalized by modeling the parameters as if they were a random sample from a population whose parameters are themselves imperfectly known. Data related directly to each individual parameter also convey information about the higher-level population parameters; the population structure in turn provides information about the individual parameters.

In typical applications, resulting estimates of individual parameters are drawn toward the center of their distribution in inverse proportion to the amount of information available about them directly. An intuitive justification of shrinkage is that unrestricted ML estimates contain sampling errors; hence it would be expected that the more extreme estimates reflect, in part, large sampling errors in that direction. This reasoning is consistent with the fact that the expected variance of ML estimates in such cases generally exceeds the variance of the true parameters.

Swaminathan and Gifford (1982) applied this idea to the Rasch model by assuming exchangeability over examinees and over items. In a Bayesian extension of JML, they provided estimation equations for the joint mode of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ in the posterior distribution

$$p(\boldsymbol{\theta},\boldsymbol{\beta}|\mathbf{X}) \propto P(\mathbf{X}|\boldsymbol{\theta},\boldsymbol{\beta})p(\boldsymbol{\theta})p(\boldsymbol{\beta}) \quad , \tag{6}$$

where $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\beta})$ are marginalizations over respective normal distributions, the parameters of which are estimated in part from the data. As expected, Swaminathan and Gifford's simulations showed the Bayesian estimates to be closer to their overall mean than unrestricted MLEs, and to have smaller mean squared error.

A similar extension of MML is described in Mislevy (1986). Marginalizing over $\boldsymbol{\theta}$ but not over the mean (μ) and standard deviation (ϕ) of identical normal priors for the βs, he gave estimation equations for the joint mode of $\boldsymbol{\beta}$, μ, and ϕ^2 in the posterior distribution

$$P(\boldsymbol{\beta},\mu,\phi^2) \propto L_M(\boldsymbol{\beta}|\mathbf{X}) \times \prod_j p(\beta_j|\mu,\phi^2) \times p(\mu,\phi^2) \quad . \tag{7}$$

As with Swaminathan and Gifford's procedure, this approach also yields estimates of βs that are closer to their estimated mean than those of the corresponding maximum likelihood procedure.

The Linear Logistic Test Model

In addition to positing a Rasch model for item responses, as in Equations 1 through 3, the LLTM assumes a linear model for the item parameters:

$$\beta_j = \sum_{k=1}^{K} q_{kj}\eta_k = \mathbf{q}_j'\boldsymbol{\eta} \quad , \tag{8}$$

or, in matrix notation,

$$\beta = Q'\eta \quad . \tag{9}$$

The basic parameters of the LLTM are η_k ($k = 1, \ldots, K$). They reflect the additive contributions to item difficulty of selected item features. The vector q_j contains coefficients relating item j to basic parameters. In Fischer's (1973) calculus example, q indicated the number and the type of operations an examinee must carry out in order to solve a differentiation item. In Mitchell's (1983) analysis of Paragraph Comprehension subtests from the Armed Services Vocational Aptitude Battery (ASVAB), q conveyed semantic and lexicographic features of a question and an associated reading passage. The reader is referred to Fischer and Formann (1982) for additional applications of the LLTM.[1]

Estimates of LLTM basic parameters can be obtained by suitable modification of JML, CML, and MML algorithms for the unconstrained Rasch model. Differences in -2 log likelihood values between the two models can be compared with the chi-square distribution with $n - K$ degrees of freedom, to test the significance of the constraints of the LLTM under the assumption that the Rasch model is true.

Fischer and Formann (1982) noted that the initial hope of explaining all reliable variation of item difficulties in terms of basic parameters has not been fulfilled; rigorous tests of fit almost always reject the LLTM. This finding is consistent with what test developers have known for decades: Two items written to test the same skill will differ in difficulty as a function of idiosyncratic features such as visual format and word choice.

Typically, however, a meaningful amount of variation can be explained. The proportion of variance of unconstrained estimates accounted for was 76% in Fischer's calculus test, and ranged from 66% to 96% in Mitchell's Paragraph Comprehension tests. Even though LLTM estimates $\bar{\beta} = Q'\hat{\eta}$ are not wholly acceptable as estimates of β, their ability to relate item performance to cognitive theory has proven useful in applications such as assessing treatment effects and modeling item bias. To the extent that the LLTM does fit, it helps to clarify exactly what makes items difficult. To the extent that it does not fit, departures indicate items that are unexpectedly difficult or easy, given the features that usually determine difficulty. Poor item construction or alternative response strategies can be detected in this way.

A Combined Model

Rationale

The assumption of exchangeability in the Bayesian estimation procedures, described above, typically leads to item parameter estimates that are more stable and have lower mean squared errors. Strictly speaking, however, assuming exchangeability over all parameters of a given type, and consequently shrinking them all to the same center, is justified only if there is no prior information to distinguish among them. This is rarely the case in item parameter estimation. In vocabulary tests, for example, it is known which words are frequently used and which ones are not; the familiar words are expected to be easier. In Fischer's calculus test, an item demanding several differentiation rules would be expected to be more difficult than one demanding only a subset of the same rules.

As Fischer and Formann (1982) pointed out, a few salient features cannot generally be expected to completely explain item parameters. However, many of a researcher's prior beliefs can be expressed in terms of such features. In particular, a model combining key aspects of the LLTM and the exchangeability

---

[1]In order to isolate the linear determinacy of the LLTM, Fischer wrote $\beta = Q'\eta + c\mathbf{1}$, where $\mathbf{1}' = (1, \ldots, 1)$ and c is an arbitrary constant. This is subsumed in the form used here by incorporating $\mathbf{1}$ into Q and c into η. The indeterminacy may be resolved by enforcing a constraint such as $\Sigma\beta = 0$ or $E(\theta) = 0$.

concept of Bayesian estimation might consider as exchangeable only parameters of items with the same pedagogically or psychologically relevant features. Shrinkage would then be observed toward the center of the subset to which an item belongs—as estimated from items of that type and possibly from other items as well, if they shared some features with it. This shrinkage could quite possibly be in the opposite direction from the center of the item set as a whole.

The General Form of the Model

Let the known (possibly vector-valued) quantity $\mathbf{q}_j$ represent auxiliary information about item j; let $p(\boldsymbol{\beta}|\mathbf{q})$ be the density function representing the distribution of $\boldsymbol{\beta}$ parameters for items with the same (generic) value of $\mathbf{q}$. [The possibility that $p(\boldsymbol{\beta}|\mathbf{q})$ may depend on unknown parameters is introduced below.] The posterior distribution of $\boldsymbol{\beta}$, given the data $\mathbf{X}$ and the auxiliary information $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_n)$, is obtained as

$$p(\boldsymbol{\beta}|\mathbf{X},\mathbf{Q}) \propto L_M(\boldsymbol{\beta}|\mathbf{X})p(\boldsymbol{\beta}|\mathbf{Q}) = \prod_i \int P(\mathbf{x}_i|\theta,\boldsymbol{\beta})p(\theta)\,\partial\theta \times \prod_j p(\boldsymbol{\beta}_j|\mathbf{q}_j) \quad . \tag{10}$$

An implementation of Equation 10 inspired by the LLTM begins with the assumption of a linear regression model for $p(\boldsymbol{\beta}|\mathbf{q})$—a response-surface prior, as introduced by Smith (1973) in the context of linear models. With $\mathbf{Q}$ and $\boldsymbol{\eta}$ defined exactly as in the LLTM, prior beliefs about item parameters can be approximated as $\text{MVN}(\mathbf{Q}'\boldsymbol{\eta},\phi^2\mathbf{I})$. Considering $\boldsymbol{\eta}$ and ϕ^2 as additional unknown parameters, the marginal posterior is obtained as

$$P(\boldsymbol{\beta},\boldsymbol{\eta},\phi^2|\mathbf{X},\mathbf{Q}) \propto L_M \times \phi^{-m}\prod_j \exp\left[\frac{-(\beta_j - q_j'\boldsymbol{\eta})^2}{2\phi^2}\right] p(\boldsymbol{\eta},\phi^2) \quad . \tag{11}$$

As in the LLTM, a linear model based on salient features gives the central tendency of items with the same features $\mathbf{q}_j$, namely $\overline{\beta}_j \equiv \mathbf{q}_j'\boldsymbol{\eta}$. Unlike the LLTM, however, variation of true parameters around these central values is anticipated.

Computational procedures for computing the posterior mode of $\boldsymbol{\beta}$, or of $\boldsymbol{\beta}$, μ, and ϕ^2 jointly, are readily obtained by generalizing the algorithms given in Mislevy (1986). The resulting solutions can be applied in the two- and three-parameter logistic models as well as for the Rasch model. The technical details of this solution are not central to the present paper, however.

A Computing Approximation for the Rasch Model

Empirical Bayesian (EB) estimation of Rasch item parameters assumes normal linear regression on salient item features. Two simplifications are applied to the exact posterior distribution given in Equation 11. First, the marginal likelihood function of $\boldsymbol{\beta}$ is replaced by the normal approximation given in Equation 5. Second, MLEs of the population parameters $\boldsymbol{\eta}$ and ϕ^2 are treated as known, after they have been estimated from MLEs $\hat{\boldsymbol{\beta}}_j$ with their standard errors $\hat{\sigma}_j$ treated as known. (It is this use of point estimates of population parameters that is commonly associated with the term "empirical Bayes.") The resulting approximation takes the following form:

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{X},\mathbf{Q}) &\propto L_M(\boldsymbol{\beta}|\mathbf{X}) \times p(\boldsymbol{\beta}|\mathbf{Q}) \\ &\propto L_M(\boldsymbol{\beta}|\mathbf{X}) \times \int\int \prod_j p(\beta_j|\mathbf{q}_j,\boldsymbol{\eta},\phi^2)p(\boldsymbol{\eta},\phi^2)\,\partial\boldsymbol{\eta}\,\partial\phi^2 \\ &\dot{\propto} \prod_j \exp\left[\frac{-(\beta_j - \hat{\beta}_j)^2}{2\hat{\sigma}_j^2}\right] \times \prod_j \exp\left[\frac{-(\beta_j - q_j'\hat{\boldsymbol{\eta}})^2}{2\hat{\phi}^2}\right] \quad . \end{aligned} \tag{12}$$

From this combination of a likelihood and prior that are both proportional to independent normal densities, independent normal posteriors follow (Box & Tiao, 1973, p. 74):

$$p(\boldsymbol{\beta}|\mathbb{X},\mathbb{Q}) \propto \prod_j \exp\left[\frac{-(\beta_j-\tilde{\beta}_j)^2}{2\tilde{\sigma}_j^2}\right] \quad , \tag{13}$$

where the means and variances are given by well-known formulas:

$$\tilde{\beta}_j = \frac{\hat{\sigma}_j^{-2}\hat{\beta}_j + \hat{\phi}^{-2}\mathbb{q}_j'\hat{\eta}}{\hat{\sigma}_j^{-2}+\hat{\phi}^{-2}} \tag{14}$$

and

$$\tilde{\sigma}_j^2 = (\hat{\sigma}_j^{-2}+\hat{\phi}^{-2})^{-1} \quad . \tag{15}$$

Computation thus proceeds in three steps, as described below.

Step 1: Unrestricted Maximum Likelihood Estimates of Item Parameters

Rasch item parameter estimates $\hat{\beta}_j$ and corresponding standard errors $\hat{\sigma}_j$ can be obtained with any of a number of widely available computer programs. Numerical values and small-sample properties of JML, CML, and MML estimates certainly differ, but any are sufficient for the present purposes. For long tests and many examinees, all support the approximation of the marginal likelihood as a product of independent normal distributions, with means given by MLEs and standard deviations given by the associated standard errors.

Step 2: Point Estimates of the Regression Parameters

The regression structure for item parameters and the normal approximation for the marginal likelihood lead to the following system of regression equations:

$$\hat{\beta}_j = \beta_j + e_j \quad , \tag{16}$$

where

$$(e_1, \ldots, e_n) \sim \text{MVN}[\mathbb{0},\text{diag}(\sigma_1^2, \ldots, \sigma_n^2)] \quad , \tag{17}$$

and

$$\beta_j = \mathbb{q}_j'\eta + f_j \quad , \tag{18}$$

where

$$(f_1, \ldots, f_n) \sim \text{MVN}(\mathbb{0},\phi^2\mathbb{I}) \quad . \tag{19}$$

Taken together, they imply

$$\hat{\beta}_j = \mathbb{q}_j'\eta + h_j \quad , \tag{20}$$

where

$$(h_1, \ldots, h_n) \sim \text{MVN}[\mathbb{0},\text{diag}(\sigma_1^2+\phi^2, \ldots, \sigma_n^2+\phi^2)] \quad . \tag{21}$$

MLEs for η and ϕ^2 can be obtained simultaneously by applying Dempster, Laird, and Rubin's (1977) EM algorithm. A special case of Braun and Jones' (1985) implementation was employed for the examples that appear in the following section. Using provisional estimates $\hat{\eta}$ and $\hat{\phi}^2$, the E-step computes conditional expectations of the unknown item parameters:

$$\tilde{\beta}_j = \text{E}(\beta_j|\hat{\beta}_j,\hat{\sigma}_j,\hat{\eta},\hat{\phi}^2) = \frac{\hat{\phi}^{-2}\overline{\beta}_j + \hat{\sigma}_j^{-2}\hat{\beta}_j}{\hat{\phi}^{-2}+\hat{\sigma}_j^{-2}} \quad , \tag{22}$$

where $\bar{\beta}_j = \mathbf{q}_j'\hat{\boldsymbol{\eta}}$ is the (provisional) modeled mean for all items with the same features as item j. The M-step uses these results to produce improved estimates:

$$\hat{\boldsymbol{\eta}} = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\tilde{\boldsymbol{\beta}} \tag{23}$$

and

$$n\hat{\phi}^2 = \tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\eta}}'\mathbf{Q}\mathbf{Q}'\hat{\boldsymbol{\eta}} \quad . \tag{24}$$

Cycles of this type are repeated until convergence is attained. Because the distribution of the hypothetical "complete data" $(\hat{\boldsymbol{\beta}},\boldsymbol{\beta})$, with parameters ϕ^2 and $\boldsymbol{\eta}$, belongs to the exponential family if σ is assumed known, convergence to a unique maximum is assured (Dempster et al., 1977).

Step 3: Final Estimates of Item Parameters

The posterior means and variances for the βs that follow from the simplifying assumptions can be calculated as in Equations 14 and 15. The EB estimate $\tilde{\beta}_j$ is thus a weighted average of the ML estimate $\hat{\beta}_j$ and the regression estimate $\bar{\beta}_j$. The relative weights are the precisions of the two estimates being combined, implying that

1. Poorly-estimated $\hat{\beta}$s shrink toward their predicted means more strongly then well-estimated $\hat{\beta}$s;
2. If all $\hat{\beta}$s are well estimated in comparison with the estimated variation around their modeled means, little shrinkage occurs and $\tilde{\boldsymbol{\beta}}$ approaches $\hat{\boldsymbol{\beta}}$; and
3. If all $\hat{\beta}$s are poorly estimated in comparison with the expected variation around their modeled means, much shrinkage occurs and $\tilde{\boldsymbol{\beta}}$ approaches $\bar{\boldsymbol{\beta}}$.

Posterior precision, or $\tilde{\sigma}_j^{-2} = \hat{\sigma}_j^{-2} + \hat{\phi}^{-2}$, is the sum of precision about β_j conveyed directly through the likelihood function and that conveyed indirectly through knowledge about item features. By exploiting auxiliary information, then, the precision of item parameter estimates can be increased without testing additional examinees.

EB estimates are distinguished most significantly from "true" Bayesian estimates by their failure to account for uncertainty associated with $\boldsymbol{\eta}$ and ϕ^2. The *nature* of the consequent differences is to overstate the apparent precision of the final EB item parameter estimates, while affecting their values only minimally. The posterior variances tend to be too small, and the distributions should be more platykurtic, like a t-distribution rather than the normal. The *magnitude* of these effects diminishes as $\boldsymbol{\eta}$ and ϕ^2 are better determined by the data. Larger N generally leads to greater precision, but test length n and the matrix of cross-products $\mathbf{Q}'\mathbf{Q}$ are also important. These influences affect the precision of regression parameters and residual variance in much the same manner as in standard regression analyses.

Numerical Example

This section reports on the application of EB estimation procedures to the 20-item Fractions subtest of the California Achievement Test (CAT), Level 3, Form A (Tiegs & Clark, 1970). The data were Rasch item difficulty estimates and standard errors, estimated from the responses of 150 sixth-grade students with the JML routine in Wright, Mead, and Bell's (1980) BICAL computer program. These values appear in Table 1, along with a specification of salient features of each item. These features, based on the CAT table of item specifications, are as follows:

1. Addition (ADD). The student must solve an addition problem involving one or more fractions and/or mixed numbers.
2. Subtraction (SUB). The student must solve a subtraction problem involving one or more fractions and/or mixed numbers.

Table 1
Rasch Item Parameter Estimates ($\hat{b}$), Standard Errors ($\hat{\sigma}$), and Item Specifications for All Items

Item	$\hat{b}$	$\hat{\sigma}$	Item Specifications					
			ADD	SUB	MUL	DIV	CD	RED
1	−3.73	.31	1	0	0	0	0	0
2	−2.02	.20	1	0	0	0	0	0
3	1.45	.28	1	0	0	0	1	0
4	1.16	.26	1	0	0	0	1	0
5	1.63	.31	1	0	0	0	1	1
6	−2.42	.21	0	1	0	0	0	0
7	−3.23	.27	0	1	0	0	0	0
8	−1.05	.18	0	1	0	0	0	0
9	1.28	.27	0	1	0	0	1	0
10	.30	.21	0	1	0	0	0	1
11	−.41	.18	0	0	1	0	0	0
12	−.80	.18	0	0	1	0	0	0
13	2.22	.38	0	0	1	0	0	1
14	1.72	.31	0	0	1	0	0	1
15	1.41	.28	0	0	1	0	0	1
16	−1.35	.18	0	0	0	1	0	0
17	.26	.21	0	0	0	1	0	0
18	1.28	.27	0	0	0	1	0	1
19	1.41	.28	0	0	0	1	0	1
20	1.05	.25	0	0	0	1	0	1

3. Multiplication (MUL). The student must solve a multiplication problem involving one or more fractions and/or mixed numbers.
4. Division (DIV). The student must solve a division problem involving one or more fractions and/or mixed numbers.
5. Common denominators (CD). The student must find a common denominator for two fractions with unlike denominators.
6. Reduction (RED). The student must reduce a fraction or mixed number to lowest terms.

A sequence of three models was fit to these data:

Model 1: EB item parameter estimates were obtained under an assumption of global exchangeability. That is, all items were shrunk toward their common mean. The resulting estimates approximate the results of Swaminathan and Gifford's (1982) procedures.

Model 2: EB estimates were obtained under the assumption of exchangeability among items with the same features, based on Table 1.

Model 3: EB estimates were again obtained, after modifying the model along lines suggested by an examination of the estimates and residuals from Model 2.

Model 1: 20 Items, Global Exchangeability

Most applications of EB estimation involve shrinkage to the common center of the parameter set. This is accomplished in the present framework by using a vector of *1*s for Q. The results of such an

analysis for the CAT Fractions test are presented in Table 2 and Figure 1. The grand mean toward which all estimates are shrunk is 0.0 (the result of the scaling convention used in BICAL); the estimated standard deviation $\hat{\phi}$ of the βs, with $\hat{\sigma}$ treated as σ, is 1.71. This compares with a standard deviation of 1.74 for

Table 2
Item-Level Results From Models 1 and 2

Model and Items	$\hat{\beta}$	$\hat{\sigma}$	$\bar{\beta}$	$\hat{\phi}$	$\tilde{\beta}$	$\tilde{\sigma}$	Shrinkage	Standardized Difference
Model 1								
1	−3.73	.31	.00	1.71	−3.61	.31	.03	−2.14
2	−2.02	.20	.00	1.71	−1.99	.20	.01	−1.17
3	1.45	.28	.00	1.71	1.41	.28	.03	.84
4	1.16	.26	.00	1.71	1.13	.26	.02	.67
5	1.63	.31	.00	1.71	1.58	.31	.03	.94
6	−2.42	.21	.00	1.71	−2.38	.21	.01	−1.40
7	−3.23	.27	.00	1.71	−3.15	.27	.02	−1.86
8	−1.05	.18	.00	1.71	−1.04	.18	.01	−.61
9	1.28	.27	.00	1.71	1.25	.27	.02	.74
10	.30	.21	.00	1.71	.30	.21	.01	.17
11	−.41	.18	.00	1.71	−.41	.18	.01	−.24
12	−.80	.18	.00	1.71	−.79	.18	.01	−.46
13	2.22	.38	.00	1.71	2.12	.37	.05	1.27
14	1.72	.31	.00	1.71	1.67	.31	.03	.99
15	1.41	.28	.00	1.71	1.37	.28	.03	.81
16	−1.35	.18	.00	1.71	−1.34	.18	.01	−.78
17	.26	.21	.00	1.71	.26	.21	.01	.15
18	1.28	.27	.00	1.71	1.25	.27	.02	.74
19	1.41	.28	.00	1.71	1.37	.28	.03	.81
20	1.05	.25	.00	1.71	1.03	.25	.02	.61
Model 2								
1	−3.73	.31	−2.75	.58	−3.61	.27	.22	−1.49
2	−2.02	.20	−2.75	.58	−2.10	.19	.11	1.19
3	1.45	.28	.75	.58	1.32	.25	.19	1.09
4	1.16	.26	.75	.58	1.09	.24	.17	.64
5	1.63	.31	2.65	.58	1.86	.27	.22	−1.55
6	−2.42	.21	−2.08	.58	−2.38	.20	.12	−.55
7	−3.23	.27	−2.08	.58	−3.02	.24	.18	−1.80
8	−1.05	.18	−2.08	.58	−1.14	.17	.09	1.69
9	1.28	.27	1.42	.58	1.30	.24	.18	−.22
10	.30	.21	−.18	.58	.24	.20	.12	.77
11	−.41	.18	−.34	.58	−.40	.17	.09	−.11
12	−.80	.18	−.34	.58	−.76	.17	.09	−.75
13	2.22	.38	1.56	.58	2.02	.32	.30	.96
14	1.72	.31	1.56	.58	1.68	.27	.22	.25
15	1.41	.28	1.56	.58	1.44	.25	.19	−.23
16	−1.35	.18	−.61	.58	−1.29	.17	.09	−1.21
17	.26	.21	−.61	.58	.16	.20	.12	1.42
18	1.28	.27	1.29	.58	1.28	.24	.18	−.01
19	1.41	.28	1.29	.58	1.39	.25	.19	.19
20	1.05	.25	1.29	.58	1.09	.23	.16	−.37

the $\hat{\beta}$s, reflecting the expectation that a set of MLEs will be more dispersed than the set of parameters they estimate. Accordingly, under the assumption of exchangeability over all items, the EB estimates shrank toward their common mean.

They did not shrink very much, however. If shrinkage for item j is defined as $(\hat{\beta}_j - \tilde{\beta}_j)/(\hat{\beta}_j - q_j'\eta)$, then the average shrinkage was only about 2%. The reason is that the estimated variance of β, about 2.92, is very large compared to the estimation error variance of the individual item parameters, about .06 on the average. Information from the likelihood function from a sample size of 150 is sufficient to overwhelm the information about inter-item similarities, when the items are as dissimilar in difficulty as those in the Fractions test.

Model 2: 20 Items, Exchangeability Given Salient Features

A second model posited exchangeability for items with the same CAT specifications. The Q matrix in this case consisted of the columns of feature indicators given in Table 1. Estimates of η and ϕ are given in Table 3; item-level results are listed in Table 2 and illustrated in Figure 2.

The values of the regression parameters η shown in Table 3 are reasonably consistent with expectations. The values for ADD, SUB, MUL, and DIV can be interpreted as values to which items exhibiting only that feature will be shrunk. ADD and SUB show lower (easier) values than MUL and DIV. The values for CD and RED are both positive, indicating additional difficulty for an item if this subskill is demanded in order to carry out the basic operation. The modeled mean for straight ADD items, for example, is -2.75; the mean for ADD items that also require reduction is $-2.75 + 1.90$, or $-.85$. Such ADD items are nearly as difficult as straight DIV items.

Figure 1
Maximum Likelihood, Regression, and Empirical Bayesian Item Parameter Estimates: Model 1

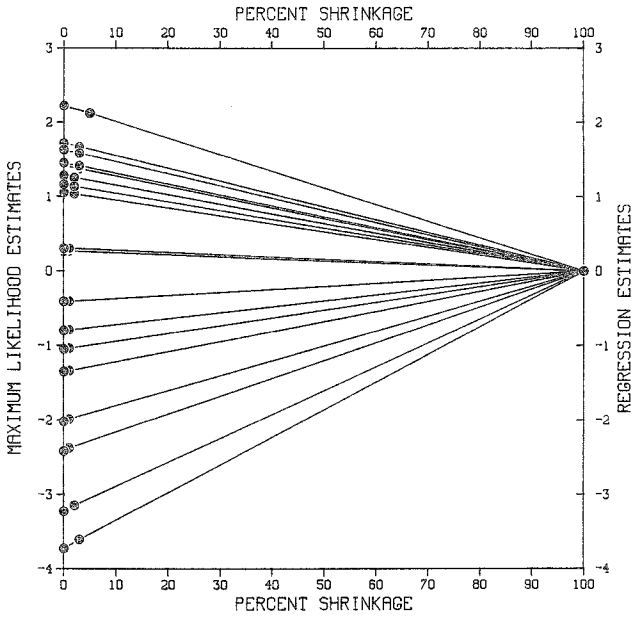


Table 3
Estimates of Regression Parameters
Under Models 2 and 3

Model and Effect (η)	Estimate
Model 2	
1. Addition	−2.75
2. Subtraction	−2.08
3. Multiplication	−.34
4. Division	−.61
5. Common denominators	3.50
6. Reduction	1.90
Standard deviation (ϕ)	.58
Model 3	
1. Addition	−1.90
2. Subtraction	−1.28
3. Multiplication	−.32
4. Division	−.25
5. Common denominators	3.10
6. Reduction	1.71
7. Whole numbers only	−1.41
Standard deviation (ϕ)	.23

The residual standard deviation $\hat{\phi}$ under Model 2 is .58, much lower than the comparable value of 1.71 in Model 1 and closer to the typical standard error of about .3. EB item parameter estimates in Table 2 thus exhibit greater shrinkage—9% to 30%. Now that items within the smaller subsets over which exchangeability is assumed are in fact more similar, the structure contributes more information with which to improve item parameter estimates. Average posterior precision increases by roughly 25%, an amount equivalent to that attainable by testing about 40 more examinees.

Note that estimates now shrink toward the appropriate mean of several predicted means rather than to a single overall mean. One item whose EB estimate moves *away* from the overall mean is item 8, the most difficult of three straight SUB items. Even though it was easier than average to begin with, the imposed exchangeability structure indicates that it would be expected to be easy based on the tasks it presents; in this particular dataset it may have been a bit more difficult than expected.

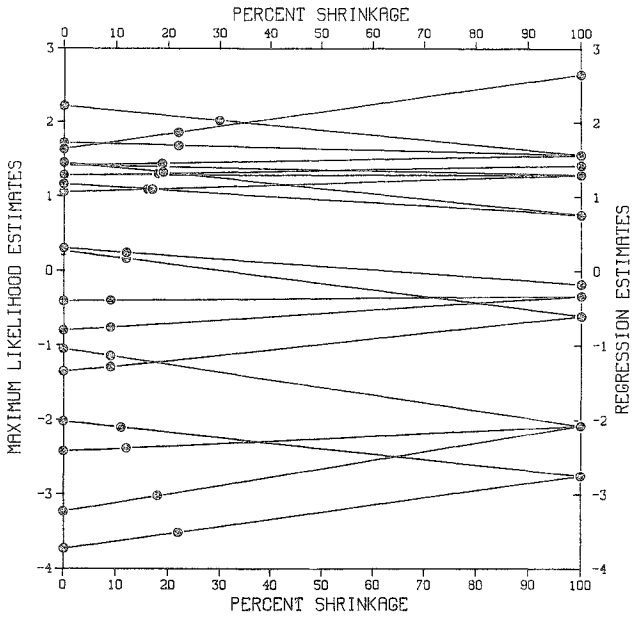
The last column in Table 2, labeled "standardized difference", gives the distance of a ML estimate from its predicted center, in standard deviation units:

$$\text{standardized difference} = \frac{\hat{\beta}_j - \bar{\beta}_j}{(\hat{\phi}^2 + \hat{\sigma}_j^2)^{1/2}} \quad . \tag{25}$$

By highlighting items that are unexpectedly far from their predicted means, these values can be useful for model modification. In conjunction with plots like that in Figure 2, they can reveal systematic departures from a user's expectations, which, upon reflection, lead the user to modify the model.

Consider as an example the three straight SUB items, 6, 7, and 8. As mentioned above, item 8 is more difficult than modeled, to an extent that ranks it among the largest residuals in absolute value. The largest absolute residual, and in the opposite direction, is the item in the same subset, namely item 7. This item is considerably easier than modeled. An inspection of item content offers an explanation: Item 7 asks for the solution of "⅙ − ⅙," which can be obtained without any knowledge of fractions at all.

Figure 2
Maximum Likelihood, Regression, and Empirical Bayesian Item Parameter Estimates: Model 2



Despite its usefulness in ranking examinees, this item may not be tapping the skills the test is ostensibly attempting to measure. Further investigation reveals a similar phenomenon among straight DIV items, where item 16 asks for the solution of "⅘ ÷ ⅘." An atypically large negative residual (easier than expected) for this item is balanced by an atypically large positive residual for another item (17) with the same features.

Further examination of items with large residuals reveals two items that are noticeably easier than expected for the same reason: Although formally they are fractions items, both item 1 (straight ADD) and item 6 (straight SUB) require only whole number operations with a fraction carried along. Because it fails to distinguish these items from straight ADD or SUB items that combine two actual fractions, Model 2 overpredicts the difficulty of items 1 and 6.

A final anomaly appears in Figure 2, for item 5. Item 5 is one of the more difficult items to begin with, but the regression model yields a prediction that is much higher than even the highest ML estimate observed. This is the only item requiring both the CD and RED skills, and the higher prediction follows from the additivity of the model. This unappealing result suggests an interaction of sorts; while two additional subskills are required, it appears likely that examinees who possess the CD skill (the more difficult of the two) also possess the RED skill. Thus, incremental difficulty over straight ADD when both skills are present is not much more than that expected from the CD subskill alone.

Model 3: 18 Items, Exchangeability Given Salient Features

The final model illustrated here modifies Model 2 in three ways:

1. Items 7 and 16, which could be solved by means of properties of operations alone, were eliminated from further consideration.

2. A column was added to the Q matrix reflecting a new salient feature: WN, or whole numbers only, applying to items 1 and 6 (which require only operations on whole numbers while a fraction is carried along).
3. To reflect the interaction of CD and RED observed for item 5, its q value for RED was changed from 1 to 0. That is, the difficulty parameters of ADD items requiring CD and RED were now considered exchangeable with those of items requiring CD (the more difficult skill) alone.

The item data for Model 3 are shown in Table 4. The results of the analysis are shown in Table 3 (regression parameter estimates), Table 5 (item-level results), and Figure 3 (a plot of ML, EB, and regression estimates). The revisions from Model 2 reduced the residual standard deviation substantially, from .58 to .23. This is about the same degree of precision as is available from the likelihood, so that EB estimates are roughly a 50/50 compromise between ML and regression estimates. Taking the approximate posterior variances at face value (recall that they are probably underestimated), it can be concluded that use of auxiliary information about items yields an increase in precision equivalent to doubling the size of the sample of examinees.

The average magnitude of standardized residuals is about the same as that from Model 2 because the denominator with which they are calculated decreased when the estimate of ϕ^2 decreased. Neither these residuals nor Figure 3 exhibit readily interpretable patterns of departures from the model.

As with any model-fitting procedure, the analysis that led to Model 3 capitalizes to some degree upon idiosyncratic features of the data at hand. Resulting estimates of precision are overly optimistic for this reason, in addition to the expedients employed by the estimation procedure. Any serious attempt to model item difficulties in the fractions domain would obviously require more data and more thought than were needed simply to illustrate computational procedures.

Table 4
Rasch Item Parameter Estimates ($\hat{b}$), Standard Errors ($\hat{\sigma}$), and Item Specifications for the Reduced Set of Items

			Item Specifications						
Item	$\hat{b}$	$\hat{\sigma}$	ADD	SUB	MUL	DIV	CD	RED	WN
1	−3.73	.31	1	0	0	0	0	0	1
2	−2.02	.20	1	0	0	0	0	0	0
3	1.45	.28	1	0	0	0	1	0	0
4	1.16	.26	1	0	0	0	1	0	0
5	1.63	.31	1	0	0	0	1	0	0
6	−2.42	.21	0	1	0	0	0	0	1
(7)									
8	−1.05	.18	0	1	0	0	0	0	0
9	1.28	.27	0	1	0	0	1	0	0
10	.30	.21	0	1	0	0	0	1	0
11	−.41	.18	0	0	1	0	0	0	0
12	−.80	.18	0	0	1	0	0	0	0
13	2.22	.38	0	0	1	0	0	1	0
14	1.72	.31	0	0	1	0	0	1	0
15	1.41	.28	0	0	1	0	0	1	0
(16)									
17	.26	.21	0	0	0	1	0	0	0
18	1.28	.27	0	0	0	1	0	1	0
19	1.41	.28	0	0	0	1	0	1	0
20	1.05	.25	0	0	0	1	0	1	0

Table 5
Item-Level Results From Model 3

Items	$\hat{\beta}$	$\hat{\sigma}$	$\bar{\beta}$	$\hat{\phi}$	$\tilde{\beta}$	$\tilde{\sigma}$	Shrinkage	Standardized Difference
1	−3.73	.31	−3.30	.23	−3.45	.18	.65	−1.11
2	−2.02	.20	−1.89	.23	−1.96	.15	.44	−.43
3	1.45	.28	1.21	.23	1.30	.18	.61	.67
4	1.16	.26	1.21	.23	1.19	.17	.57	−.14
5	1.63	.31	1.21	.23	1.35	.18	.65	1.10
6	−2.42	.21	−2.70	.23	−2.55	.15	.47	.89
(7)								
8	−1.05	.18	−1.28	.23	−1.14	.14	.39	.80
9	1.28	.27	1.82	.23	1.60	.17	.59	−1.53
10	.30	.21	−.32	.23	.36	.15	.47	−.42
11	−.41	.18	−.32	.23	−.38	.14	.39	−.30
12	−.80	.18	1.39	.23	−.61	.14	.39	−1.65
13	2.22	.38	1.39	.23	1.60	.19	.74	1.89
14	1.72	.31	1.39	.23	1.50	.18	.65	.87
15	1.41	.28	1.39	.23	1.39	.18	.61	.07
(16)								
17	.26	.21	−.25	.23	.02	.15	.47	1.66
18	1.28	.27	1.46	.23	1.38	.17	.59	−.50
19	1.41	.28	1.46	.23	1.44	.18	.61	−.13
20	1.05	.25	1.46	.23	1.27	.17	.55	−1.21

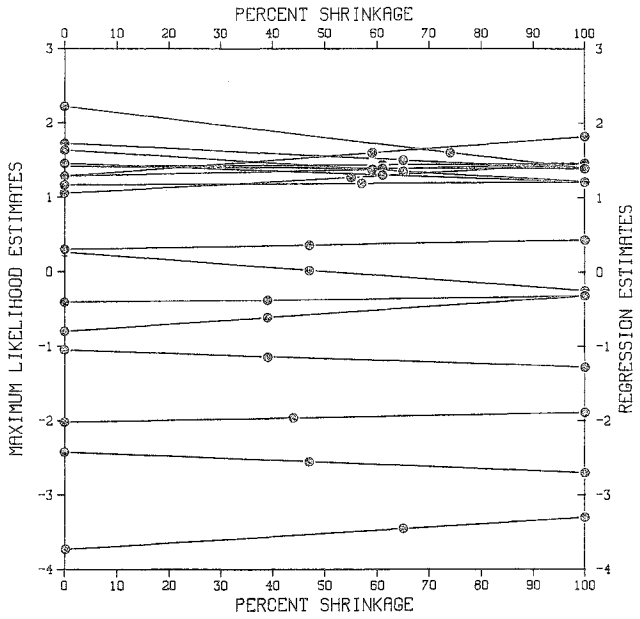
Discussion

The potential benefits of using auxiliary information about items in item parameter estimation are increased precision and diagnostic capabilities. In the numerical example above, auxiliary information contributed as much information about item parameters as did the likelihood function. Conditional on the veracity of the assumed exchangeability structure, then, precision was increased by an amount equal to that attainable by doubling the number of examinees. Diagnostic checks revealed two items that might not be measuring the skills intended, because their solution did not require actual manipulation of the fractions they contained.

The plausibility of the exchangeability structure can also be verified with diagnostic checks. Two additional safeguards also mitigate the effects of specification errors at this stage. First, if the structure is badly in error and items assumed exchangeable turn out not to be very similar, shrinkage will be minimal (as in Model 1 of the example). Of course, minimal shrinkage does not necessarily signal misspecification or lack of exchangeability; all other things being equal, shrinkage decreases as N increases. Second, increasing N leads to consistent item parameter estimates even if the exchangeability structure is flawed.

The simplified computing approximation used in this paper works best for the Rasch model, where it is needed least; even fairly small samples give reasonably good item parameter estimates there. The same ideas can be applied more profitably to IRT models with more parameters, each less well-determined by data (e.g., the three-parameter logistic model, and models for multiple-category item responses). The

Figure 3
Maximum Likelihood, Regression, and Empirical Bayesian Item Parameter Estimates: Model 3



computational procedures for the general model are then required, because it may not be possible to obtain finite unrestricted ML estimates and their standard errors. No explicit averaging of ML and regression estimates can be accomplished in those cases, and Bayesian estimates must be obtained directly from item responses.

References

Andersen, E. B. (1973). *Conditional inference and models for measuring*. Copenhagen: Danish Institute for Mental Health.

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.

Braun, H. I., & Jones, D. H. (1985). *Use of empirical Bayes methods in the study of the validity of academic predictors of graduate school performance* (GRE Board Professional Report No. 79-13p; ETS Research Report 84-34). Princeton NJ: Educational Testing Service.

Cressie, N., & Holland, P. W. (1983). Characterizing the manifest probabilities of latent trait models. *Psychometrika, 48,* 129–141.

de Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11,* 183–196.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, 39,* 1–38.

Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association, 70,* 311–319.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Fischer, G. H., & Formann, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Psychological Measurement, 6,* 397–416.

Haberman, S. (1977). Maximum likelihood estimates in exponential response models. *Annals of Statistics, 5,* 815–841.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B, 34,* 1–41.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–196.

Mitchell, K. J. (1983). *Cognitive processing determinants of item difficulty on the verbal subtests of the Armed Services Vocational Aptitude Battery* (Technical Report 598). Alexandria VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. [Expanded edition, University of Chicago Press, 1980.]

Scheiblechner, H. (1972). Das Lernen und Losen komplexer Denkaufgaben. *Zeitschrift für Experimentalle und Angewandte Psychologie, 19,* 476–506.

Smith, A. F. M. (1973). Bayes estimates in one-way and two-way models. *Biometrika, 60,* 319–329.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7,* 175–191.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50,* 349–364.

Thissen, D. (1982). Marginal maximum likelihood estimation in the one-parameter logistic model. *Psychometrika, 47,* 175–186.

Tiegs, E., & Clark, W. (1970). *The California Achievement Tests: 1970 edition.* Monterey CA: McGraw-Hill.

Wright, B. D., Mead, R. J., & Bell, S. R. (1980). *BICAL: Calibrating items with the Rasch model* (Research Memorandum 23C). Chicago: University of Chicago, Department of Education, Statistical Laboratory.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23–48.

Acknowledgments

This work was supported by Contract No. N00014-85-K-0683, project designation NR 150-539, from Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research. The author thanks Charles Lewis and Peter Pashley for their comments and suggestions, Henry Braun and Bruce Kaplan for their assistance in applying the EM estimation procedure described in the example, and Maxine Kingston for the figures.

Author's Address

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