Research Report

Joint and Conditional Maximum Likelihood Estimation for the Rasch Model for Binary Responses

Shelby J. Haberman

Research & Development

May 2004
RR-04-20

Joint and Conditional Maximum Likelihood Estimation for the Rasch Model for Binary Responses

Shelby J. Haberman

ETS, Princeton, NJ

May 2004

Abstract

The usefulness of joint and conditional maximum-likelihood is considered for the Rasch model under realistic testing conditions in which the number of examinees is very large and the number is items is relatively large. Conditions for consistency and asymptotic normality are explored, effects of model error are investigated, measures of prediction are estimated, and generalized residuals are developed.

Key words: Logarithmic penalty, entropy, consistency, normal approximation

Acknowledgements

The authors would like to thank Matthias von Davier, Paul Holland, Sandip Sinharay, and Hariharan Swaminathan for their helpful comments.

Introduction

The Rasch model (Rasch, 1960) remains a commonly used model for analysis of item responses despite competition from more general two-parameter and three-parameter logistic (2PL and 3PL) models (Hambleton et al., 1991, chap. 2). The Rasch model has the attraction of relative simplicity, and parameter estimation is feasible without use of a parametric model for latent ability. Nonetheless, many rather basic problems affect its use in psychometrics. Significant problems arise in estimation of parameters, both in a computational sense and in terms of asymptotic theory, in realistic cases in which the number of items is relatively large and the number of examinees is very large. The effects of lack of fit need examination, as does the fundamental problem of assessment of the size of model error and of formally testing for lack of fit. Methods for residual analysis are also required that can be justified in terms of large-sample theory.

This report examines these issues for the commonly used joint and conditional maximum likelihood approaches (Hambleton et al., 1991, chap. 3). To simplify matters, only binary responses are considered. In later reports, versions of the Rasch model for polytomous responses will be examined. To illustrate results, data from the October 2, 2002, SAT® I Math and Verbal examinations are used. In these tests, the number of items ranges from 60 in the Math test to 78 in the Verbal test, and 446,607 individuals are examined. For simplicity, responses are taken as correct or other (incorrect or omitted). This approach is not entirely satisfactory to the extent that scoring distinguishes between incorrect and omitted responses except in the case of constructed responses. On the other hand, the very large number of observations permits an examination of the extent of model misfit that is not readily accomplished with fewer data, and the tests under study have the advantage of careful construction based on long experience. Later reports are expected to examine variations of the Rasch model in which incorrect responses are distinguished from omitted responses. Later reports are also expected to consider common extensions of the Rasch model in which item discrimination is not constant and to consider marginal estimation procedures in which the ability distribution is assumed to satisfy a parametric model. The choice of methods in this report reflects a desire to examine cases in which log-linear models are applied to directly observed data and in which the responses are as

simple as possible. As evident in this report, even the relatively simple situation under study in this report leads to substantial methodological problems.

Section 1 examines joint maximum-likelihood estimation (JMLE) for binary responses (Andersen, 1972; Fischer, 1981; Haberman, 1977b). Known results are reviewed and considered for their implications for the SAT data. It is established that, even if the model is true, JMLE will not lead to fully satisfactory approximate confidence intervals for item difficulties and that the normal approximation for the distribution of ability estimates will not be fully satisfactory. These results may not be evident from the author's previous work, although no actual contradiction is involved (Haberman, 1977b). It is also shown that some problems will arise in use of normal approximations for estimates of the expected log penalty function (Gilula & Haberman, 1994; Gilula & Haberman, 1995). Behavior of JMLE is established in cases in which model error is present, and possible use of generalized residuals (Haberman, 1978) is considered. Results for generalized residuals are found to be somewhat discouraging.

Section 2 examines conditional maximum-likelihood estimation (CMLE) for binary responses (Andersen, 1972; Andersen, 1973a; Andersen, 1973b; Fischer, 1981). The basic properties of conditional maximum-likelihood estimates are reviewed, and computation with the Newton-Rapshon algorithm is described. It is shown that convolutions can be used to yield a version of the Newton-Raphson algorithm that is computationally efficient (Liou, 1994). It is also shown that appropriate starting values are readily obtained, at least for large numbers of items, by use of joint estimation. Normal approximations for estimates of item difficulty are established whether the model is true and/or whether the number of items increases, and satisfactory estimates are developed for the expected log penalty function. Satisfactory generalized residuals are determined, and a simple generalization of the Rasch model is developed that permits assessment of lack of fit for sample sizes and numbers of items encountered with the SAT I Math and Verbal examinations.

Section 3 considers some implications of CMLE to latent-structure Rasch models (Cressie & Holland, 1983; Tjur, 1982). The difference between the log-linear model corresponding to the CMLE approach and the Rasch model is considered. It is shown that many latent-structure models can yield the same observed joint distribution of item

responses, and it is shown that the joint distribution can typically be produced by use of a latent-class model with approximates the number of latent classes equal to about half the number of items.

Section 4 summarizes the implications of the research for psychometric practice and discusses some further areas of possible development.

1 Joint Maximum-likelihood estimation

To describe joint maximum-likelihood estimation, let examinees i from 1 to $n \geq 2$ provide responses Y_{ij} equal to 1 or 0 to items j from 1 to $q \geq 2$. Normally Y_{ij} is 1 for a correct response of subject i to item j, and Y_{ij} is 0 otherwise. Assume that associated with examinee i is a real ability parameter θ_i. Let the θ_i be independent and identically distributed random variables with distribution function D, so that $D(x)$ is the probability that $\theta_i \leq x$ for each real x. For simplicity, assume that, for some bounded real interval Θ, θ_i is in Θ for each examinee i. Assume that the nq responses Y_{ij}, $1 \leq i \leq n$, $1 \leq j \leq q$, are conditionally independent given the n ability parameters θ_i, $1 \leq i \leq n$. Assume that the conditional distribution of response Y_{ij} for a particular examinee i and response j given the n ability parameters θ_h, $1 \leq h \leq n$, depends only on the ability parameter θ_i for examinee i. Let the logistic distribution function lgt be defined on the real line so that

$$\mathrm{lgt}(x) = [1 + \exp(-x)]^{-1}$$

for all real x. For each item j, let β_j be a fixed parameter that measures the difficulty of item j. Let $\boldsymbol{\beta}$ denote the q-dimensional vector with coordinates β_j for $1 \leq j \leq q$. Given that $\theta_i = \theta$ for a real number θ, let the conditional probability $P(Y_{ij} = 1|\theta_i = \theta)$ that $Y_{ij} = 1$ be

$$\mathrm{lgt}(\theta - \beta_j),$$

so that the conditional log odds that $Y_{ij} = 1$ given that $\theta_i = \theta$ is

$$\log\left[\frac{P(Y_{ij} = 1|\theta_i = \theta)}{P(Y_{ij} = 0|\theta_i = \theta)}\right] = \theta - \beta_j.$$

The random variable

$$p_{ij} = \mathrm{lgt}(\theta_i - \beta_j)$$

may be employed to characterize the conditional distribution of Y_{ij} given θ_i. To ensure identifiability of the item difficulty parameters under the estimation procedures in this report, it is convenient to assume that $\beta_1 = 0$. In addition, assume that the number of observations n exceeds the number of items q. In large-sample results in which q increases, also assume that the empirical distribution of the β_j, $1 \le j \le q$, converges weakly to the distribution of a bounded random variable β^*. Without loss of generality, β^* can be defined to be independent of the θ_i and Y_{ij}.

Under the proposed model, a marginal log likelihood function can be constructed with little difficulty, at least if computational considerations are ignored. Let $\mathbf{Y}_i$ denote the q-dimensional vector with coordinates Y_{ij}, $1 \le j \le q$, so that each $\mathbf{Y}_i$ is in the set Γ of q-dimensional vectors with coordinates 0 or 1. Let $\mathbf{c}$ be in Γ, let $Y_{i+} = \sum_{i=1}^{q} Y_{ij}$ be the number of items correctly answered by examinee i, and let $k = \sum_{j=1}^{q} c_j$. Let

$$\mathbf{x}^T\mathbf{y} = \sum_{j=1}^{q} x_j y_j$$

for q-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$. Let

$$\Psi(\boldsymbol{\beta}, \theta) = \frac{1}{\prod_{j=1}^{q}[1 + \exp(\theta - \beta_j)]}$$

for real θ. Given customary notation for a Lebesgue-Stieltjes integral, the probability that $\mathbf{Y}_i = \mathbf{c}$ is the expected value

$$\begin{aligned} p_J(\mathbf{c}) &= E\left(\exp\{\sum_{j=1}^{q}[c_j \log p_{1j} + (1 - c_j)\log(1 - p_{ij})]\}\right) \\ &= E\left(\frac{\exp(k\theta_1 - \boldsymbol{\beta}^T\mathbf{c})}{\prod_{j=1}^{q}[1 + \exp(\theta_1 - \beta_j)]}\right) \\ &= \exp(-\boldsymbol{\beta}^T\mathbf{c}) \int_{-\infty}^{\infty} e^{k\theta}\Psi(\boldsymbol{\beta}, \theta) dD(\theta) \end{aligned}$$

(Cressie & Holland, 1983). For $\mathbf{p}_J$ equal to the array of $p_J(\mathbf{c})$ for $\mathbf{c}$ in Γ, the marginal log likelihood function is then

$$\ell(\mathbf{p}_J) = \sum_{i=1}^{n} \log p_J(\mathbf{Y}_i).$$

For an alternative expression, let

$$\mathbf{Y}_+ = \sum_{i=1}^{n} \mathbf{Y}_i.$$

For integers k, $0 \le k \le q$, let $\Gamma(k)$ be the subset of $\mathbf{c}$ in Γ such that $\sum_{j=1}^{q} c_j = k$. Under the constraint that

$$p_J(\mathbf{c}) = \exp(-\boldsymbol{\beta}^T \mathbf{c}) \int_{-\infty}^{\infty} \exp(k\theta) \Psi(\boldsymbol{\beta}, \theta) dD(\theta)$$

for $\mathbf{c} \in \Gamma(k)$ and $0 \le k \le q$ for some $\boldsymbol{\beta}$ with $\beta_1 = 0$ and some distribution function D,

$$\ell(\mathbf{p}_J) = -\boldsymbol{\beta}^T \mathbf{Y}_+ + \sum_{i=1}^{n} \log \int_{-\infty}^{\infty} \exp(Y_{i+}\theta) \Psi(\boldsymbol{\beta}, \theta) dD(\theta). \tag{1}$$

Let ℓ_M be the maximum of $\ell(\mathbf{p}_J)$ subject to the constraints used in (1). Then $\hat{\boldsymbol{\beta}}_J$ and $\hat{D}_J$ are joint unrestricted marginal maximum-likelihood estimates of $\boldsymbol{\beta}$ and D if

$$\ell(\hat{\mathbf{p}}_J) = \ell_M,$$

the first coordinate $\hat{\beta}_{1J}$ of $\hat{\boldsymbol{\beta}}$ is 0, and $\hat{p}_J$ is defined so that

$$\hat{p}_J(\mathbf{c}) = \exp(-\hat{\boldsymbol{\beta}}_J^T \mathbf{c}) \int_{-\infty}^{\infty} e^{k\theta} \Psi(\hat{\boldsymbol{\beta}}_J, \theta) d\hat{D}_J(\theta)$$

for $\mathbf{c}$ in $\Gamma(k)$ for $0 \le k \le q$.

Because determination of possible values of $\hat{\boldsymbol{\beta}}_J$ and $\hat{D}_J$ is far from trivial, alternative approaches are commonly pursued. In this section, JMLE is considered. In Section 2, CMLE is developed. This report does not consider approaches in which parametric models are used for the distribution function D. It is planned that such approaches will be discussed in later reports.

In JMLE, estimation is performed with the θ_i regarded as fixed parameters, a practice with a long history of controversy in many areas of statistics (Kiefer & Wolfowitz, 1956). The joint log likelihood function

$$\ell_J(\mathbf{p}) = \sum_{i=1}^{n} \sum_{j=1}^{q} [Y_{ij} \log p_{ij} + (1 - Y_{ij}) \log(1 - p_{ij})]$$

is maximized subject to the constraints that the array $\mathbf{p}$ of

$$p_{ij} = \mathrm{lgt}(\mu_{ij}),\ 1 \le i \le n, 1 \le j \le q, \tag{2}$$

satisfies

$$\mu_{ij} = \theta_i - \beta_j,\ 1 \le i \le n, 1 \le j \le q, \tag{3}$$

the θ_i are real for $1 \leq i \leq n$, the β_j are real for $1 \leq j \leq q$, and $\beta_1 = 0$. Under these constraints,

$$\ell_J(\mathbf{p}) = -\boldsymbol{\beta}^T \mathbf{Y}_+ + \sum_{i=1}^{n} \log[\exp(Y_{i+}\theta_i)\Psi(\boldsymbol{\beta}, \theta_i)]. \tag{4}$$

For any distribution function D and any q-dimensional vector $\boldsymbol{\beta}$ with $\beta_1 = 0$, the maximum of $\exp(Y_{i+}\theta)\Psi(\boldsymbol{\beta}, \theta)$ over θ is at least the integral

$$\int_{-\infty}^{\infty} \exp(Y_{i+}\theta)\Psi(\boldsymbol{\beta}, \theta) dD(\theta),$$

with equality only if the maximum of $\exp(Y_{i+}\theta)\Psi(\boldsymbol{\beta}, \theta)$ is achieved by some θ, $D(x) = 0$ for $x < \theta$, and $D(x) = 1$ for $x \geq \theta$ (Haberman, 1977b). Let ℓ_{JM} be the maximum value of $\ell_J(\mathbf{p})$ subject to the model constraints (2), (3), and $\beta_1 = 0$. Then the maximum ℓ_{JM} is at least ℓ_M. It is readily verified that $\ell_{JM} > \ell_M$ unless each Y_{i+} has the same value. Thus joint maximum-likelihood differs somewhat from marginal maximum likelihood in terms of the function to be maximized.

If it exists, then the joint maximum-likelihood estimate $\hat{\mathbf{p}}$ of $\mathbf{p}$ is the array of joint maximum-likelihood estimates $\hat{p}_{ij}$ of p_{ij}, $1 \leq i \leq n$, $1 \leq j \leq q$, that satisfies

$$\hat{p}_{ij} = \text{lgt}(\hat{\mu}_{ij}), \tag{5}$$

$$\hat{\mu}_{ij} = \hat{\theta}_i - \hat{\beta}_j,\ 1 \leq i \leq n, 1 \leq j \leq q, \tag{6}$$

$\hat{\theta}_i$ is real for $1 \leq i \leq n$, $\hat{\beta}_j$ is real for $1 \leq j \leq q$, $\hat{\beta}_1 = 0$, and

$$\ell_J(\hat{\mathbf{p}}) = \ell_{JM}.$$

Let $\hat{\boldsymbol{\beta}}$ be the q-dimensional vector with coordinates $\hat{\beta}_j$. If

$$Y_{+j} = \sum_{i=1}^{n} Y_{ij},$$

then the maximum-likelihood equations

$$\hat{p}_{i+} = \sum_{j=1}^{q} \hat{p}_{ij} = Y_{i+} \tag{7}$$

and

$$\hat{p}_{+j} = \sum_{i=1}^{n} \hat{p}_{ij} = Y_{+j} \tag{8}$$

are satisfied (Haberman, 1977b). If real θ'_i and β'_j exist such that $\beta'_1 = 0$,

$$\mu'_{ij} = \theta'_i - \beta'_j,$$

$$p'_{ij} = \text{lgt}(\mu'_{ij}),$$

$$p'_{i+} = \sum_{j=1}^{q} p'_{ij} = Y_{i+},$$

and

$$p'_{+j} = \sum_{i=1}^{n} p'_{ij} = Y_{+j},$$

then $\theta'_i = \hat{\theta}_i$ and $\beta'_j = \hat{\beta}_j$, so that $\mu'_{ij} = \hat{\mu}_{ij}$ and $p'_{ij} = \hat{p}_{ij}$. This result implies that, if they exist, joint maximum-likelihood estimates are uniquely defined. The case of $1 \leq h < i \leq n$, $\beta'_j = \hat{\beta}_j$, $\theta'_g = \hat{\theta}_g$ for g not h or i, $\theta'_h = \hat{\theta}_i$, and $\theta'_i = \hat{\theta}_h$ implies that $\hat{\theta}_i = \hat{\theta}_h$ whenever $Y_{i+} = Y_{h+}$.

If any Y_{i+} is 0 or q, so that examinee i answers no item correctly or answers all items correctly, then joint maximum-likelihood estimates cannot exist, for each $\hat{p}_{ij}$ must be positive and less than 1, so that $0 < \hat{p}_{i+} < q$ and $\hat{p}_{i+}$ cannot satisfy the maximum-likelihood equation $\hat{p}_{i+} = Y_{i+}$. This matter is not a purely academic issue. For the SAT Math data under study, one examinee had no correct response, and 646 examinees answered all items correctly. For the SAT Verbal exam, two examinees answered no item correctly, and 29 examinees answered all items correctly. The issue of existence of joint maximum-likelihood estimates will be discussed more thoroughly after consideration of collapsed tables.

1.1 Computations and Collapsed Tables

Computation of joint maximum-likelihood estimates is greatly simplified by use of a collapsed table based on the counts f_{kj}, $0 \leq k \leq q$, $1 \leq j \leq q$. Here f_{kj} is the number of examinees i, $1 \leq i \leq n$, of examinees i, $1 \leq i \leq n$, with total number correct $Y_{i+} = k$ with a correct response to item j ($Y_{ij} = 1$). For integers k from 0 to q, let n_k be the number of examinees i with total number correct $Y_{i+} = k$. Let K be the set of integers k such that $n_k > 0$, and let N_K be the number of elements of K, so that $N_K \leq q + 1$. If joint maximum-likelihood estimates exist, then $n_0 = n_q = 0$, so that $N_K \leq q - 1$. Consider a table with entries f_{kj}, k in K, $1 \leq j \leq q$. As shown in this section, computation of joint

maximum-likelihood estimates of θ_i and β_j is then based on a logit model in which the observations f_{kj}, k in K, $1 \leq j \leq q$, are regarded as independent binomial random variables with respective sample sizes n_k and probabilities

$$p_{kjC} = \text{lgt}(\theta_{kC} - \beta_{jC})$$

for some real θ_{kC}, k in K, and β_{jC}, $1 \leq j \leq q$, such that $\beta_{jC} = 0$. If joint maximum-likelihood estimates exist, then the maximum-likelihood estimate of β_{jC} exists and equals the joint maximum-likelihood estimate $\hat{\beta}_j$, and the maximum-likelihood estimate of θ_{kC}, k in K, exists and equals the common value of $\hat{\theta}_i$ for $Y_{i+} = k$. The gain here is that computation for an n by q array of binary responses is reduced to computation for an N_K by q array of binomial responses. Because the number n of items is typically much larger than is the number q of items, the computational savings are very large. The collapsed table is also useful in verification that joint maximum-likelihood estimates exist at all. In cases of nonexistence, the collapsed table remains important in computations.

To verify the relationship between joint maximum-likelihood estimation and the logit model for the f_{kj}, some preliminary results concerning summations of the counts f_{kj} are needed. Let the indicator variable I_{ik} be 1 for $Y_{i+} = k$ and 0 otherwise. Then

$$f_{kj} = \sum_{i=1}^{n} Y_{ij} I_{ik}.$$

It follows that the row summation

$$\begin{aligned} f_{k+} &= \sum_{j=1}^{q} f_{kj} \\ &= \sum_{j=1}^{q} \sum_{i=1}^{n} Y_{ij} I_{ik} \\ &= \sum_{i=1}^{n} I_{ik} \sum_{j=1}^{q} Y_{ij} \\ &= k \sum_{i=1}^{n} I_{ik} \\ &= k n_k, \end{aligned}$$

for $0 \leq k \leq q$, and, for $1 \leq j \leq q$, the column summation

$$\begin{aligned} f_{+j} &= \sum_{k=0}^{q} f_{kj} \\ &= \sum_{k=0}^{q} \sum_{i=1}^{n} Y_{ij} I_{ik} \\ &= \sum_{i=1}^{n} \sum_{k=0}^{q} I_{ik} \\ &= Y_{+j}, \end{aligned}$$

the number of examinees i with response j correct ($Y_{ij} = 1$). Because $f_{kj} = 0$ if $n_k = 0$,

$$f_{+j} = \sum_{k \in K} f_{kj}.$$

For the logit model for the f_{kj}, k in K, $1 \leq j \leq q$, the log likelihood function

$$\ell_{JC}(\mathbf{p}_C) = \sum_{k \in K} \sum_{j=1}^{q} [f_{kj} \log p_{kjC} + (n_k - f_{kj} \log(1 - p_{kjC})]$$

for arrays $\mathbf{p}_C$ with elements p_{kjC}, k in K, $1 \leq j \leq q$, such that

$$p_{kjC} = \text{lgt}(\mu_{kjC}),$$

$$\mu_{kjC} = \theta_{kC} - \beta_{jC},$$

θ_{kC} and β_{jC} are real, and $\beta_{1C} = 0$. Let ℓ_{JCM} be the supremum of ℓ_{JC}. If $\ell_{JC}(\hat{\mathbf{p}}_C) = \ell_{JCM}$ for

$$\hat{p}_{kjC} = \text{lgt}(\hat{\mu}_{kjC}) \tag{9}$$

and

$$\hat{\mu}_{kjC} = \hat{\theta}_{kC} - \hat{\beta}_{jC}, \tag{10}$$

if $\hat{\theta}_{kC}$ and $\hat{\beta}_{jC}$ are real, and if $\hat{\beta}_{1C} = 0$, then the likelihood equations

$$\sum_{k \in K} n_k \hat{p}_{kjC} = f_{+j} = Y_{+j},\ 1 \leq j \leq q, \tag{11}$$

and

$$n_k \sum_{j=1}^{q} \hat{p}_{kjC} = f_{k+} = k n_k,\ k \in K, \tag{12}$$

must hold (Haberman, 1978, pp. 198, 294–295), so that

$$\sum_{j=1}^{q} \hat{p}_{kjC} = k. \tag{13}$$

If real θ'_{kC}, k in K, and β'_{jC}, $1 \leq j \leq q$, exist such that $\beta'_{1C} = 0$,

$$\mu'_{kjC} = \theta'_{kC} - \beta'_{jC},$$

$$p'_{kjC} = \text{lgt}(\mu'_{kjC}),$$

$$\sum_{j=1}^{q} p'_{kjC} = k,$$

and

$$\sum_{k \in K} n_k p'_{kjC} = Y_{+j},$$

then $p'_{kjC} = \hat{p}_{kjC}$, $\theta'_{kC} = \hat{\theta}_{kC}$, $\beta'_{jC} = \hat{\beta}_{jC}$, and $\mu'_{kjC} = \hat{\mu}_{kjC}$, so that the maximum-likelihood estimate of $\mathbf{p}_C$ is uniquely defined.

The relationship of the $\hat{\theta}_{kC}$ and $\hat{\beta}_{jC}$ to joint maximum-likelihood estimates is straightforward. Suppose that the joint maximum-likelihood estimates $\hat{\theta}_i$, $\hat{\beta}_j$, $\hat{\mu}_{ij}$, and $\hat{p}_{ij}$ exist. Because $\hat{\theta}_i = \hat{\theta}_h$ for $Y_{i+} = Y_{h+}$, the equations (9), (10), (11), and (13) that define $\hat{p}_{kjC}$, $\hat{\mu}_{kjC}$, $\hat{\theta}_{kC}$, and $\hat{\beta}_{jC}$ are satisfied and $\hat{\beta}_{1C} = 0$ if $\hat{\theta}_{kC}$ is the common value of $\hat{\theta}_i$ for each examinee i with $Y_{i+} = k$, $\hat{\beta}_{jC} = \hat{\beta}_j$ for each item j, $\hat{\mu}_{kjC}$ is the common value of $\hat{\mu}_{ij}$ for $Y_{i+} = k$, and $\hat{p}_{kjC}$ is the common value of $\hat{p}_{ij}$ for $Y_{i+} = k$. Thus $\hat{\theta}_{kC}$, $\hat{\beta}_{jC}$, $\hat{\mu}_{kjC}$, and $\hat{p}_{kjC}$ are maximum-likelihood estimates for θ_{kC}, β_{jC}, μ_{kjC}, and p_{kjC}, respectively. Conversely, if $\hat{\theta}_{kC}$, $\hat{\beta}_{jC}$, $\hat{\mu}_{kjC}$, and $\hat{p}_{kjC}$ are maximum-likelihood estimates for θ_{kC}, β_{jC}, μ_{kjC}, and p_{kjC}, respectively, so that (9), (10), (11), and (13) hold and $\hat{\beta}_{1C} = 0$, then (5), (6), (7), and (8) are satisfied and $\hat{\beta}_1 = 0$ if $\hat{\theta}_i = \hat{\theta}_{kC}$ for examinees i with $Y_{i+} = k$ and $\hat{\beta}_j = \hat{\beta}_{jC}$ for items j, so that the $\hat{\beta}_j$, $\hat{\theta}_i$, $\hat{\mu}_{ij}$, and $\hat{p}_{ij}$ are joint maximum-likelihood estimates of β_j, θ_i, μ_{ij}, and p_{ij}, respectively. Thus joint maximum-likelihood estimates are readily found by maximization of ℓ_{JC}.

1.2 Existence of Joint Maximum-likelihood Estimates

As already noted, joint maximum-likelihood estimates do not exist if n_0 or n_q is 0, so that some examinee answers all items correctly or all items incorrectly. A necessary and

sufficient condition for existence is provided by the following theorem (Haberman, 1977b; Fischer, 1981).

Theorem 1 *Joint maximum-likelihood estimates exist if, and only if, there is no real a and b such that the following conditions hold:*

1. *a and b are not integers,*
2. *$0 < a < q$ and $0 < b < n$,*
3. *$f_{kj} = 0$ for $k < a$ and $Y_{+j} < b$,*
4. *$f_{kj} = n_k$ for $k > a$ and $Y_{+j} > b$,*
5. *k and j exist such that $n_k > 0$ and either $k < a$ and $Y_{+j} < b$ or $k > a$ and $Y_{+j} > b$.*

Several basic cases should be noted. If $n_0 > 0$, then the case $a = 0.5$ and $b = q - 0.5$ shows that joint maximum-likelihood estimates do not exist, for $Y_{+j} \leq n - n_0 < b$ and $f_{0j} = 0$ for $1 \leq j \leq q$. If $n_q > 0$, then the case $a = q - 0.5$ and $b = 0.5$ shows that joint maximum-likelihood estimates do not exist, for $Y_{+j} \geq n_q > b$ and $f_{qj} = n_q$ for $1 \leq j \leq q$. Similar arguments show that joint maximum-likelihood estimates do not exist if Y_{+j} is 0 or n for some j from 1 to q. If $n_0 = n_q = 0$ and if to each j corresponds a k such that $0 < f_{kj} < n_k$ and $1 \leq k \leq q - 1$, then joint maximum-likelihood estimates exist.

1.3 Extended Joint Maximum-likelihood Estimates

The definition of joint maximum-likelihood estimates may be extended to yield a unique estimate $\hat{\mathbf{p}}$ of the probability array $\mathbf{p}$ without any conditions (Haberman, 1974, pp. 402–404). The resulting estimate, termed the extended joint maximum-likelihood estimate of $\mathbf{p}$, is uniquely defined by the conditions that

$$\ell_J(\hat{\mathbf{p}}) = \ell_{JM},$$

$\hat{\mathbf{p}}$ has elements $\hat{p}_{ij}$ for $1 \leq i \leq n$ and $1 \leq j \leq q$, and $\theta_{i\nu}$, $1 \leq i \leq n$, and real $\beta_{j\nu}$, $1 \leq j \leq q$, exist for $\nu \geq 1$ such that $\beta_{1\nu} = 0$ and $\text{lgt}(\theta_{i\nu} - \beta_{j\nu})$ approaches $\hat{p}_{ij}$, $1 \leq i \leq n$, $1 \leq j \leq q$, as ν approaches ∞. The estimates $\hat{p}_{ij}$ satisfy (7) and (8), just as in the case of conventional

joint maximum-likelihood estimates. If the conventional joint maximum-likelihood estimate $\hat{\mathbf{p}}$ of $\mathbf{p}$ is defined, then the definition implies that $\hat{\mathbf{p}}$ is also the extended joint maximum-likelihood estimate of $\mathbf{p}$.

On the other hand, if $\theta'_{i\nu}$, $1 \leq i \leq n$, and $\beta'_{j\nu}$, $1 \leq j \leq q$, are real for $\nu \geq 0$, $\beta'_{1\nu} = 0$, $\text{lgt}(\theta'_{i\nu} - \beta'_{j\nu})$ approaches p'_{ij} for $1 \leq i \leq n$, $1 \leq j \leq q$, as ν approaches ∞, $p'_{i+} = Y_{i+}$, and $p'_{+j} = Y_{+j}$, then $p'_{ij} = \hat{p}_{ij}$.

The definition of $\hat{\mathbf{p}}$ implies that $0 \leq \hat{p}_{ij} \leq 1$, so that $\hat{p}_{ij} = 0$ if $Y_{i+} = 0$ or $Y_{+j} = 0$ and $\hat{p}_{ij} = 1$ if $Y_{i+} = q$ or $Y_{+j} = n$.

In terms of the collapsed table of counts f_{kj}, k in K, $1 \leq j \leq q$, extended maximum-likelihood estimates may also be defined. There is a unique $\hat{\mathbf{p}}_C$ such that

$$\ell_{JC}(\hat{\mathbf{p}}_C) = \ell_{JCM},$$

$\hat{\mathbf{p}}_C$ has elements $\hat{p}_{kjC}$ for k in K and $1 \leq j \leq q$, real $\theta_{kC\nu}$, k in K, and $\beta_{jC\nu}$, $1 \leq j \leq q$, are defined for $\nu \geq 1$ so that $\beta_{1C\nu} = 0$ and $\text{lgt}(\theta_{kC\nu} - \beta_{jC\nu})$ approaches $\hat{p}_{kjC}$ as ν approaches ∞. The equations (11) and (13) hold. If it exists, the conventional maximum-likelihood estimate of $\mathbf{p}_C$ is also the extended maximum-likelihood estimate.

If real $\theta'_{kC\nu}$, k in K, and $\beta'_{jC\nu}$, $1 \leq j \leq q$, exist for $\nu \geq 1$ such that $\beta'_{1C\nu} = 0$ and $\text{lgt}(\theta'_{kC\nu} - \beta'_{jC\nu})$ approaches p'_{kjC} for k in K, $1 \leq j \leq q$, as ν approaches ∞, if

$$\sum_{j=1}^{q} p'_{kjC} = k,$$

and if

$$\sum_{k \in K} n_k p'_{kjC} = f_{+j},$$

then $p'_{kjC} = \hat{p}_{kjC}$.

The relationship between estimates is quite straightforward. Nearly the same arguments used for conventional joint maximum-likelihood estimates show that $\hat{p}_{ij} = \hat{p}_{kjC}$ if $Y_{i+} = k$. It follows that $\hat{p}_{0jC} = 0$ if $n_0 > 0$ and $\hat{p}_{qjC} = 1$ if $n_q > 0$.

Extended maximum-likelihood is a bit more complicated when parameters θ_i, θ_{kC}, β_j, β_{jC}, μ_{ij}, and μ_{kjC} are considered. For μ_{ij} and μ_{kjC}, a reasonable definition is available if

infinite values are permitted. One has

$$\hat{\mu}_{ij} = \hat{\mu}_{kjC} = \begin{cases} \infty, & \hat{p}_{kjC} = 1, \\ \log\left(\frac{\hat{p}_{kjC}}{1-\hat{p}_{kjC}}\right), & 0 < \hat{p}_{kjC} < 1, \\ -\infty, & \hat{p}_{kjC} = 0, \end{cases}$$

for $Y_{i+} = k$. Because $\theta_i = \mu_{i1}$ and $\theta_{kC} = \mu_{k1C}$, it also follows that

$$\hat{\theta}_i = \hat{\theta}_{kC} = \begin{cases} \infty, & \hat{p}_{k1C} = 1, \\ \log\left(\frac{\hat{p}_{k1C}}{1-\hat{p}_{k1C}}\right), & 0 < \hat{p}_{k1C} < 1, \\ -\infty, & \hat{p}_{k1C} = 0 \end{cases}$$

if $Y_{i+} = k$. Conventions for $\hat{\beta}_j = \hat{\beta}_{jC}$ are somewhat more complicated. If conventional joint maximum-likelihood estimates exist, then

$$\hat{\beta}_j = \hat{\beta}_{jC} = \hat{\mu}_{ij} - \hat{\mu}_{i1} = \hat{\mu}_{kjC} - \hat{\mu}_{1jC}$$

for $Y_{i+} = k$ and k in K. This formula is applicable to extended joint maximum-likelihood estimation as long as some integer k in K exists such that $\hat{p}_{k1C}$ and $\hat{p}_{kjC}$ are not both 0 or both 1. For any examinee i with $Y_{i+} = k$ and any k in K such that $\hat{p}_{k1C}$ and $\hat{p}_{kjC}$ are not both 0 or both 1, the difference $\hat{\mu}_{kjC} - \hat{\mu}_{k1C}$ has the same value, and this value may be assigned to $\hat{\beta}_j = \hat{\beta}_{jC}$. This practice may lead to an estimate of ∞ or $-\infty$. If no k in K exists such that $\hat{p}_{kjC}$ and $\hat{p}_{k1C}$ are not both 0 or both 1, then there is no obvious basis for definition of $\hat{\beta}_j = \hat{\beta}_{jC}$. In this instance, the convention is adopted that $\hat{\beta}_j = \hat{\beta}_{jC} = 0$.

1.4 Consistency

Even if the Rasch model is valid, if the number q of items is constant, the β_j are constant, and n approaches ∞, then the $\hat{\theta}_i$ are not consistent estimates of the θ_i, and the $\hat{\beta}_j$ are not consistent estimates of the β_j (Andersen, 1973a, pp. 66–69). Indeed, the probability approaches 1 that ordinary joint maximum-likelihood estimates do not even exist (Haberman, 1977b). Even if extended joint maximum-likelihood estimates are employed, $\hat{\theta}_i$ remains inconsistent for θ_i, and $\hat{\beta}_j$ remains inconsistent for β_j. As shown in this section, $\hat{\beta}_j$ converges almost surely to a limit β_{jM} that differs from β_j by a term of order q^{-1}.

Consistency results are known to be quite different if the Rasch model holds and the number q of items increases (Haberman, 1977b). If $q^{-1} \log n$ approaches 0, then the probability approaches 1 that each $\hat{\theta}_i$, $\hat{\beta}_j$, and $\hat{\mu}_{ij}$ is finite, and

$$\max_{1 \le i \le n} |\hat{\theta}_i - \theta_i|,$$

$$\max_{1 \le j \le q} |\hat{\beta}_j - \beta_j|,$$

and

$$\max_{1 \le i \le n} \max_{1 \le j \le q} |\hat{\mu}_{ij} - \mu_{ij}|$$

all converge in probability to 0. Given that the function lgt has a bounded derivative, it also true that $\max_i \max_j |\hat{p}_{ij} - p_{ij}|$ converges in probability to 0. These results are not fully satisfactory for the SAT I data under study. For the Math exam, $q = 60$, $n = 446,607$, and $q^{-1} \log n = 0.217$. For the Verbal exam, $q = 78$ and n is still 446,607, so that $q^{-1} \log n$ is 0.167. Neither value of $q^{-1} \log n$ is very small, so that the asymptotic results are not inconsistent with the already observed problem that not all $\hat{\theta}_i$ are finite in either the Math or the Verbal exam.

In this report, the consistency results previously derived are extended to new situations. It is shown that the ability estimate $\hat{\beta}_j$ for item j converges in probability to the corresponding ability β_j for item j whenever the number q of items approaches ∞. Indeed $\max_{1 \le j \le q} |\hat{\beta}_j - \beta_j|$ converges in probability to 0. For any specific examinee i, $\hat{\theta}_i - \theta_i$ converges in probability to 0. In addition, weak convergence results for the distribution of θ_i follow. For example, the empirical distribution function of the $\hat{\theta}_i$ converges to the common distribution function D of θ_i at any continuity point of D. Consistency results are also demonstrated for estimates of expected logarithmic penalty functions associated with the Rasch model.

To begin, consider the case of a fixed number q of items. The following theorems summarize the basic consistency problems for estimation of the examinee ability θ_i, the logit μ_{ij} for examinee i and item j, and the probability p_{ij} of a correct response for examinee i and item j.

Theorem 2 *Let the number q of items be fixed, and let the number n of examinees approach ∞. Then the probability that joint maximum-likelihood estimates exist approaches 0.*

Proof. Let P_k, $0 \leq k \leq q$, be the unconditional probability $P(Y_{i+} = k)$ that $Y_{i+} = k$. Then each P_k is positive, so that the probability $P_0 + P_q$ is positive that examinee i has either no correct responses ($Y_{i+} = 0$) or all correct responses ($Y_{i+} = q$). Joint maximum-likelihood estimates only can exist if $0 < Y_{i+} < q$ for each examinee i from 1 to n. The probability that $0 < Y_{i+} < q$ for $1 \leq i \leq n$ is $[1 - (P_0 + P_k)]^n$. As n approaches 0, this probability approaches 0.

Theorem 3 *Under the conditions of Theorem 2, for any integer $i \geq 1$, $\hat{\theta}_i - \theta_i$ does not converge in probability to 0. For each integer $i \geq 1$ and any integer j, $1 \leq j \leq q$, $\hat{\mu}_{ij} - \mu_{ij}$ does not converge in probability to 0.*

Proof. If $Y_{i+} = 0$, then $\hat{\theta}_i = -\infty$. If $Y_{i+} = q$, then $\hat{\theta}_i = \infty$. Because the probabilities P_0 and P_q defined in the proof of Theorem 2 are positive and constant and because $\hat{\theta}_i = -\infty$ with probability at least P_0 and $\hat{\theta}_i = \infty$ with probability at least P_q, it follows that $\hat{\theta}_i - \theta_i$ does not converge in probability to 0. Virtually the same argument applies to $\hat{\mu}_{ij}$, so that $\hat{\mu}_{ij} - \mu_{ij}$ does not converge in probability to 0. It then follows that $\hat{p}_{ij}$ is 0 with probability at least P_0, and $\hat{p}_{ij} = 1$ with probability at least P_q, so that $\hat{p}_{ij} - p_{ij}$ does not converge in probability to 0.

To demonstrate inconsistency of the $\hat{\beta}_j$ in the case of q fixed is relatively complicated in the case of extended joint maximum-likelihood estimates. Results depend on the expectation $E(f_{kj})$ of the count f_{kj} for $0 \leq k \leq q$ and $1 \leq j \leq q$. To find $E(f_{kj})$, consider the conditional expectation $m_{kjC} = m_{kj}(\boldsymbol{\beta})$ of Y_{ij} given $Y_{i+} = k$. As is well-known, this expectation is a function of the vector $\boldsymbol{\beta}$ of item difficulties $\beta_{j'}$, $1 \leq j' \leq q$, and not of the examinee ability θ_i. To find $m_{kj}(\boldsymbol{\beta})$, let

$$s_k(\boldsymbol{\beta}) = \sum_{\mathbf{c} \in \Gamma(k)} \exp(-\boldsymbol{\beta}^T \mathbf{c})$$

be the symmetric function of order k for the q variables $\exp(-\beta_j c_j)$, $1 \leq j \leq q$ (Fischer, 1981). Let

$$s_{kj}(\boldsymbol{\beta}) = \sum_{\mathbf{c} \in \Gamma(k)} c_j \exp(-\boldsymbol{\beta}^T \mathbf{c}),$$

so that $-s_{kj}(\boldsymbol{\beta})$ is the partial derivative of $s_k(\boldsymbol{\beta})$ with respect to β_j. Let $\mathbf{Y}_i$ denote the q-dimensional vector of Y_{ij}, $1 \leq j \leq q$. Then the conditional probability that $\mathbf{Y}_1 = \mathbf{c}$ in $\Gamma(k)$ given $Y_{1+} = k$ is

$$p_{JC}(\mathbf{c}) = \exp(-\boldsymbol{\beta}^T \mathbf{c})/s_k(\boldsymbol{\beta}),$$

and

$$m_{kj}(\boldsymbol{\beta}) = \sum_{\mathbf{c} \in \Gamma(k)} c_j p_{JC}(\mathbf{c}) = \frac{s_{kj}(\boldsymbol{\beta})}{s_k(\boldsymbol{\beta})}.$$

For $k = 0$, $m_{kj}(\boldsymbol{\beta})$ is 0. For $k = q$, $m_{kj}(\boldsymbol{\beta})$ is 1. Otherwise, $m_{kj}(\boldsymbol{\beta}) > 0$. Given the probability $P_k = P(Y_{1+} = k)$ from the proof of Theorem 2 and the conditional probability m_{kjC}, it follows that $E(n_k) = nP_k$ and

$$E(f_{kj}) = E(n_k)m_{kjC} = nP_k m_{kjC}.$$

The strong law of large numbers implies that f_{kj}/n converges almost surely to $P_k m_{kjC} > 0$.

Given the conditional probabilities m_{kjC}, $0 \leq k \leq q$, and $1 \leq j \leq q$, the probabilities P_k, $0 \leq k \leq q$, and the unconditional probabilities $p_j^Y = P(Y_{1j} = 1)$, $1 \leq j \leq q$, the basic limiting properties of $\hat{\theta}_{kC}$ and $\hat{\beta}_j$ can be summarized as in the following theorem.

Theorem 4 *Under the conditions of Theorem 2, $\hat{\theta}_{kC}$ converges almost surely to θ_{kM}, $0 \leq k \leq q$, $\hat{\theta}_{0C}$ in K, and $\hat{\beta}_j$ converges almost surely to β_{jM}, $1 \leq j \leq q$, where $\theta_{0M} = -\infty$, $\theta_{qM} = \infty$, real θ_{kM}, $1 \leq k \leq q-1$, real β_{jM}, $1 \leq j \leq q$, and real p_{kjM}, $0 \leq k \leq q$, $1 \leq j \leq q$, are uniquely determined by the conditions that $p_{0jM} = 0$, $p_{qjM} = 1$, $\beta_{jM} = 0$,*

$$p_{kjM} = \mathrm{lgt}(\theta_{kM} - \beta_{jM}),\ 1 \leq k \leq q-1, 1 \leq j \leq q, \tag{14}$$

$$\sum_{k=0}^{q} P_k p_{kjM} = \sum_{k=0}^{q} P_k m_{kjC} = p_j^Y,\ 1 \leq j \leq q, \tag{15}$$

and

$$\sum_{j=1}^{q} p_{kjM} = \sum_{j=1}^{q} m_{kjC} = k,\ 0 \leq k \leq q. \tag{16}$$

Proof. Existence and uniqueness of θ_{kM}, β_{jM}, and p_{kjM} follows from standard results for log-linear models (Haberman, 1974, chap. 8). Results on almost sure convergence follow from general results on concave likelihood functions (Haberman, 1989).

To interpret the limit parameters θ_{kM}, β_{jM}, and p_{kjM}, logarithmic penalty functions may be employed (Gilula & Haberman, 1994; Gilula & Haberman, 1995). Let

$$H(x, y) = -x \log(y) - (1 - x) \log(1 - y)$$

for x and y between 0 and 1, where $0 \log 0 = 0$. Consider probability prediction of the responses $\mathbf{Y}_1$ from the sums Y_{1+} under the incorrect model that, conditional on $Y_{1+} = k$, the Y_{ij}, $1 \leq j \leq q$, are independently distributed with probability

$$\pi_{kj} = \text{lgt}(\theta'_k - \beta'_j)$$

that $Y_{ij} = 1$ for unknown real parameters θ'_k, $0 \leq k \leq q$, and β'_j, $1 \leq j \leq q$, $\beta'_j = 0$. The expected logarithmic penalty per item is

$$q^{-1} \sum_{k=0}^{q} \sum_{j=1}^{q} P_k H(m_{kjC}, \pi_{kj}).$$

If $0 \log 0 = 0$, then the minimum expected penalty per item is

$$H_J = q^{-1} \sum_{k=0}^{q} \sum_{j=1}^{q} P_k H(m_{kjC}, p_{kjM}).$$

The expected penalty per observation approaches H_J if θ'_k approaches θ_{kM}, $0 \leq k \leq q$, and β'_j approaches β_{jM}, $1 \leq j \leq q$. Theorem 4 implies that the estimated expected log penalty function per item

$$\hat{H}_J = -\frac{1}{nq} \ell_{JCM}$$

converges almost surely to H_J.

Theorem 4 implies that inconsistency of $\hat{\beta}_j$ is observed when β_j and β_{jM} differ. This situation is typically but not necessarily the case. For instance, $\beta_j = \beta_{jM}$ if all β_j are 0. On the other hand, it is rather unusual to have $\beta_j = \beta_{jM}$ for all items j. In the simplest case, $q = 2$, so that

$$m_{1j}(\boldsymbol{\beta}) = \begin{cases} \frac{\exp(-\beta_1)}{1+\exp(-\beta_1)}, & j = 1, \\ \frac{1}{1+\exp(-\beta_1)}, & j = 2, \end{cases}$$

$$p_{1jM} = \text{lgt}(\theta_{1M} - \beta_{jM}) = m_{1j}(\boldsymbol{\beta}),$$

and

$$p_{11M} + p_{12M} = 1.$$

It follows that $\theta_{1M} = \beta_1$ and $\beta_{1M} = 2\beta_1$ (Andersen, 1973a). If more than two items are present, then no simple expressions are normally available for β_{jM}.

The expected logarithmic penalty per item H_J is at least as large as the conditional entropy measure per item

$$H_M = -q^{-1} \sum_{k=1}^{q-1} P_k \sum_{j=1}^{q} H(m_{kjC}, m_{kjC})$$

that corresponds to the conditional entropy per item of Y_{1B} given Y_{i+} for a random variable B uniformly distributed on the integers 1 to q and independent of the Y_{ij}, $1 \leq j \leq q$. One has $H_J = H_M$ if, and only if, $p_{kjM} = m_{kjC}$. The entropy per item H_M has an estimate

$$\hat{H}_M = -\frac{1}{nq} \sum_{k=1}^{q-1} n_k H(f_{kj}/n_k, f_{kj}/n_k)$$

that converges almost surely to H_M. In the definition of $\hat{H}_M$, the convention is followed that $0/0 = 0$.

The magnitude of the difference between β_j and β_{jM} is of order q^{-1}. For a formal statement and proof of this claim, consider the following theorem in which the number of items is allowed to increase.

Theorem 5 *A real number $\tau > 0$ exists such that $|\beta_{jM} - \beta_j| < \tau/q$ for all $q \geq 1$ and all items j, $1 \leq j \leq q$.*

Proof. To verify this claim, consider the difference between m_{kjC} and $p_{kjC} = \text{lgt}(\theta_{kC} - \beta_j)$, $1 \leq k \leq q-1$, where p_{kjC} is uniquely defined by the condition that

$$\sum_{j=1}^{q} p_{kjC} = k$$

(Haberman, 1974, chap. 10). Let U_{kj} be independent Bernoulli observations with probability p_{kjC} that $U_{kj} = 1$. The conditional probability m_{kjC} that $Y_{ij} = 1$ given that $Y_{i+} = k$ is then the conditional probability that $U_{kj} = 1$ given that

$$U_{k+} = \sum_{j=1}^{q} = k.$$

This latter probability is then

$$P(U_{kj} = 1)P(U_{k+} - U_{kj} = k - 1)/P(U_{k+} = k).$$

Let

$$\sigma_{kjC}^2 = p_{kjC}(1 - p_{kjC})$$

be the variance of U_{kj}. If k, q, and n are selected so that

$$\sigma_{k+C}^2 = \sum_{j=1}^{q} \sigma_{kjC}^2$$

approaches ∞, then

$$(U_{k+} - k)/\sigma_{k+C}$$

and

$$(U_{k+} - U_{kj} - 1 + p_{kjC}))/(\sigma_{k+C}^2 - \sigma_{kjC}^2)^{1/2}$$

converge in distribution to a standard normal random variable (Cramér, 1946, pp. 217–218). A refinement of this result permits approximation of m_{kjC}. To derive the desired approximations requires some simple modifications of results on Edgeworth expansions for lattice distributions (Esseen, 1945). Terms are used based on the normal density function and on its first three derivatives. Let

$$\psi_k = \frac{1}{\sigma_{k+C}^2} \sum_j (2p_{kjC} - 1)\sigma_{kjC}^2,$$

so that $-\psi_k/\sigma_{k+C}$ is the skewness coefficient of U_{k+}. It then follows that

$$\sigma_{k+C}^4[m_{kjC} - p_{kjC} - (2\sigma_{k+C}^2)^{-1}\sigma_{kjC}^2(2p_{kjC} - 1 - \psi_k)],\ 1 \leq j \leq q$$

is uniformly bounded. This result indicates that $m_{kjC} - p_{kjC}$ is of order $1/\sigma_{k+C}^2$.

To show that $\beta_{jM} - \beta_j$ is of order q^{-1} requires use of fixed point theorems (Loomis & Sternberg, 1968, pp. 228–234). In applications in this paper, the maximum norm is used, so that a q-dimensional vector $\mathbf{x}$ with coordinate x_j for $1 \leq j \leq q$ has maximum norm

$$|\mathbf{x}| = \max_{1 \leq j \leq q} |x_j|.$$

Consider solution of (15) for $2 \leq j \leq q$ subject to the constraints that $\beta_1 = 0$ and (14) and (16) hold. For each q-dimensional vector $\mathbf{x}$ with coordinates x_j, $1 \leq j \leq q$, there is a unique real value $g_k(\mathbf{x})$, $1 \leq k \leq q-1$, such that

$$\sum_{j=1}^{q} \mathrm{lgt}(g_k(\mathbf{x}) - x_j) = k.$$

Let w be the real function with value

$$w(y) = \mathrm{lgt}(y)[1 - \mathrm{lgt}(y)]$$

for real y, and let

$$w_{kj}(\mathbf{x}) = w(g_k(\mathbf{x}) - x_j).$$

The function g_k is infinitely differentiable, and the gradient ∇g_k of g_k has coordinate j equal to

$$g_{kj}(\mathbf{x}) = -\frac{w_{kj}(\mathbf{x})}{\sum_{h=1}^{q} w_{kh}(\mathbf{x})}.$$

Let $\mathbf{F}(\mathbf{x}, \mathbf{y})$ be defined for q-dimensional vectors $\mathbf{x}$ and $q-1$ by q arrays $\mathbf{y}$ with coordinates y_{kj} for k from 1 to $q-1$ and j from 1 to q so that $\mathbf{F}(\mathbf{x}, \mathbf{y})$ has coordinate j equal to

$$F_j(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{q-1} [k^{-1} y_{k+} \mathrm{lgt}(g_k(\mathbf{x}) - x_j) - y_{kj}],$$

where

$$y_{k+} = \sum_{j=1}^{q} y_{kj}.$$

Then

$$\mathbf{F}(\boldsymbol{\beta}, \mathbf{z}) = \mathbf{0}$$

for $z_{kj} = P_k p_{kjC}$, and

$$\mathbf{F}(\boldsymbol{\beta}_M, \mathbf{y}') = 0$$

for $\boldsymbol{\beta}_M$ with coordinates β_{jM} for $1 \leq j \leq q$ and for $\mathbf{y}'$ with coordinates $y'_{kj} = P_k p_{kjM}$. The function $\mathbf{F}$ is infinitely differentiable. It is linear in the second argument. To evaluate the partial differential with respect to the first argument, let δ be the Kronecker delta function. Then the partial gradient of F_j with respect to the first argument has coordinate j' equal to

$$F_{jj'}(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{q-1} k^{-1} y_{k+} w_{kj}(\mathbf{x})[g_{kj'}(\mathbf{x}) - \delta_{jj'}],$$

so that the sum

$$\sum_{j'=1}^{q} F_{jj'}(\mathbf{x}, \mathbf{y}) = 0.$$

Let $\nabla \mathbf{F}(\mathbf{x}.\mathbf{y})$ be the q by q matrix with elements $F_{jj'}(\mathbf{x}, \mathbf{y})$ for integers j and j' from 1 to q.

Use of fixed point theorems yields the desired conclusion, although some technical complications emerge in that the most efficient parameterization for these results uses parameters that sum to 0. The basic observation is that $\mathbf{F}(\boldsymbol{\gamma})$ and $\mathbf{F}(\boldsymbol{\beta})$ are the same if

$$\gamma_j = \beta_j - q^{-1} \sum_{h=2}^{q} \beta_h.$$

Similarly, $m_{kj}(\boldsymbol{\beta}) = m_{kj}(\boldsymbol{\gamma})$. In like manner, $\boldsymbol{\gamma}_M$ is defined with coordinates

$$\gamma_{jM} = \beta_{jM} - q^{-1} \sum_{h=2}^{q} \beta_{hM}.$$

One has

$$\beta_j = \gamma_j - \gamma_1$$

and

$$\beta_{jM} = \gamma_{jM} - \gamma_{1M}.$$

Let Ω be the smallest value of

$$\sum_{k=1}^{q-1} P_k w_{kj}(\boldsymbol{\beta}) g_{kj'}(\boldsymbol{\beta})$$

for j and j' integers between 1 and q. Let $\mathbf{1}$ be the q-dimensional vector with all coordinates equal to 1. One seeks $\mathbf{x}$ given $\mathbf{y}$ such that

$$\mathbf{G}(\mathbf{x}, \mathbf{y}) = \mathbf{F}(\mathbf{x}, \mathbf{y}) - \Omega \mathbf{1}\mathbf{1}^T \mathbf{x} = \mathbf{0}.$$

The use of Ω leads to $\boldsymbol{\gamma}$ as the unique solution of $\mathbf{G}(\mathbf{x}, \mathbf{z}) = \mathbf{0}$ and leads to $\boldsymbol{\gamma}_M$ as the unique solution of $\mathbf{G}(\mathbf{x}, \mathbf{y}') = \mathbf{0}$. Let

$$\nabla \mathbf{G}(\mathbf{x}, \mathbf{y}) = \nabla \mathbf{F}(\mathbf{x}, \mathbf{y}) - \Omega \mathbf{1}\mathbf{1}^T,$$

and let

$$\mathbf{D}(\mathbf{x}, \mathbf{y}) = \mathbf{x} - [\nabla \mathbf{G}(\boldsymbol{\gamma}, \mathbf{y})]^{-1} \mathbf{G}(\mathbf{x}, \mathbf{y}).$$

Consider an algorithm in which an initial value is $\boldsymbol{\gamma}_0 = \boldsymbol{\gamma}$ and subsequent values $\boldsymbol{\gamma}_t$ are defined so that

$$\boldsymbol{\gamma}_{t+1} = \mathbf{D}(\boldsymbol{\gamma}_t, \mathbf{z}).$$

The result on fixed points that is required is that

$$|\boldsymbol{\gamma}_t - \boldsymbol{\gamma}_M| < C^t \delta/(1-C)$$

if C and δ are positive constants such that $C < 1$ and

$$|\mathbf{D}(\mathbf{x}, \mathbf{z}) - \mathbf{D}(\boldsymbol{\gamma}, \mathbf{z})| < \delta |\mathbf{x} - \boldsymbol{\gamma}|$$

whenever $|\mathbf{x} - \boldsymbol{\gamma}| < C/(1-\delta)$. Use of standard albeit tedious arguments from calculus shows that the upper limit of $q|\boldsymbol{\gamma}_M - \boldsymbol{\gamma}|$ is finite. It then follows that the upper limit of $q|\boldsymbol{\beta}_M - \boldsymbol{\beta}|$ is finite. The conclusion of the theorem then follows.

Given the definitions of θ_{kM} and θ_{kC} and the properties of g_k, the proof of Theorem 5 implies that $q|\theta_{kM} - \theta_{kC}|$ and $q|p_{kjM} - p_{kjC}|$ are bounded if q approaches ∞ and k/q converges to a positive constant less than 1. More precise expressions for these differences can be obtained but are not especially attractive.

The arguments used in Theorem 5 also can be applied to a variety of expected logarithmic penalties. To begin, consider the conditional entropy H_B of Y'_B given Y_{1+} and B for B uniformly distributed on the integers 1 to q and Y'_j Bernoulli random variables for $1 \leq j \leq q$ such that $P(Y'_j = 1 | Y_{1+} = k) = p_{kjC}$. Then

$$H_B = -\sum_{k=1}^{q-1} P_k \sum_{j=1}^{q} H(p_{kjC}, p_{kjC})$$

and H_M differ by a term of order q^{-1}. A variety of entropy measures are closely linked. The difference $H_J - H_M$ is of order q^{-2}, so that $H_J - H_B$ is of order q^{-1}. With a similar argument based on the normal approximation for the distribution of Y_{1+} given θ_1, it follows that $H_J - H_\theta$ is of order q^{-1} if

$$H_\theta = -q^{-1} \sum_{j=1}^{q} E(H(p_{1j}, p_{1j}))$$

is the conditional entropy per item of $\mathbf{Y}_1$ given θ_1. In turn, if $p^* = \text{lgt}(\theta_1 - \beta^*)$, then H_θ converges to

$$H^* = -E(H(p^*, p^*)),$$

the conditional entropy of a random variable Y^* with values 0 and 1 such that, conditional on θ_1 and β^*, the probability is $\text{lgt}(\theta_1 - \beta^*)$ that Y^* is 1.

Because Y_{1+} has no more than $q+1$ possible values, the conditional entropy per item

$$H_{+\theta} = -q^{-1} \sum_{k=0}^{q} E((P(Y_{1+} = k|\theta_1) \log P(Y_{1+} = k|\theta_i))$$

of Y_{1+} given θ_i and the unconditional entropy per item

$$H_+ = -q^{-1} \sum_{k=0}^{q} P_k \log P_k$$

of Y_{1+} cannot exceed $q^{-1} \log(q+1)$. It follows that the conditional entropy per item

$$\begin{aligned} H_C &= -q^{-1} \sum_{k=0}^{q} P_k \sum_{\mathbf{c} \in \Gamma(k)} p_{JC}(\mathbf{c}) \log p_{JC}(\mathbf{c}) \\ &= H_\theta - H_{+\theta} \end{aligned}$$

of $\mathbf{Y}_1$ given Y_{1+} and θ_1 differs from H_θ by a term of order $q^{-1} \log q$. The conditional distribution of $\mathbf{Y}_1$ given Y_{1+} and θ_1 is assumed independent of θ_1, so that H_C is also the conditional entropy per item of $\mathbf{Y}_1$ given Y_{1+}. It follows that H_C differs from H_B, H_J, and H_M by terms of order $q^{-1} \log q$. In turn, the unconditional entropy per item

$$\begin{aligned} H_U &= -q^{-1} \sum_{k=0}^{q} \sum_{\mathbf{c} \in \Gamma(k)} p_J(\mathbf{c}) \log p_J(\mathbf{c}) \\ &= H_C + H_+ \end{aligned}$$

of $\mathbf{Y}_1$ differs from H_θ, H_C, H_J, H_B, and H_M by terms of order $q^{-1} \log q$. Thus H_U, H_θ, H_C, H_J, H_B, and H_M all approach H^* as the number of items q becomes large.

Given that the bias $\beta_{jM} - \beta_j$ is reduced as q increases, there is the suggestion that the inconsistency of the joint maximum likelihood estimators for the Rasch model can be removed if the asymptotic framework is changed so that both the sample size n and the number of items q both approach infinity (Haberman, 1977b). The following result is available.

Theorem 6 *Let q approach ∞, so that n approaches ∞. Then $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_M|$ and $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|$ both converge in probability to 0.*

Proof. Consider solution of

$$\mathbf{G}(\mathbf{x}, \mathbf{f}) = \mathbf{0}$$

for $\mathbf{f}$ with coordinates f_{kj}. The previous argument in the proof of Theorem 5 based on fixed-point theorems is easily modified. The normal approximations for the sums

$$\sum_{k=1}^{q-1} (f_{kj} - n_k m_{kjC})$$

and large-deviation arguments may be used to demonstrate that the probability approaches 1 that $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_M|$ and $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|$ both converge in probability to 0.

Under the conditions of Theorem 6, it also follows that $\hat{\theta}_{kM} - \theta_{kC}$ and $\hat{p}_{kjM} - p_{kjC}$ converge in probability to 0 if k/q converges to a positive constant less than 1. In turn, it follows that, for any specific individual i, $\hat{\theta}_i$ converges in probability to θ_i. Thus for any real $\delta > 0$, the fraction of examinees $i \leq n$ with $|\hat{\theta}_i - \theta_i| > \delta$ converges in probability to 0. This result permits estimation of the distribution of the random variable θ_i. Let $\hat{D}$ be the empirical distribution function of the $\hat{\theta}_i$, so that $\hat{D}(x)$ is the fraction of the $\hat{\theta}_i$ that do not exceed the real number x. If D is continuous at x, then $|\hat{D}(x) - D(x)|$ converges in probability to 0. The argument is essentially the same as one used to study convergence in distribution of sums of two random variables, one of which converges in probability to 0 (Rao, 1973, pp. 122–123). Simple modification of the proof of the Helly-Bray theorem implies that if h is a continuous or piecewise-continuous bounded function on the extended real line and h is continuous at θ_1 with probability 1, then

$$\hat{E}(h(\theta)) = n^{-1} \sum_{i=1}^{n} h(\hat{\theta}_i)$$

converges in probability to $E(h(\theta_1))$ (Rao, 1973, pp. 117–118). If the common distribution function D of the θ_i is continuous, as is the case for θ_1 a continuous random variable, then similar arguments show that

$$|\hat{D} - D| = \sup_x |\hat{D}(x) - D(x)|$$

converges in probability to 0.

The difference $\hat{H}_J - H_U$ then converges in probability to 0, so that the various conditional entropies under study can be estimated. The difference $\hat{H}_M - H_M$ can only be expected to converge in probability to 0 if q^2/n approaches 0. For the SAT I data under study, $\hat{H}_J$ is not likely to be a very accurate estimate of the unconditional entropy H_U, for $q^{-1} \log q$ is 0.068 for the Math test and 0.056 for the Verbal test. The observed values of $\hat{H}_J$ is 0.450 for the Math exam and 0.501 for the Verbal exam. Thus the bias issue is potentially a major problem.

1.5 Normal Approximations

The bias issues already noted in the discussion of consistency have an unusual effect on normal approximations. It is relatively easy to find a normal approximation for the item difficulty estimate $\hat{\beta}_j$, but this approximation is not really satisfactory because the asymptotic mean is β_{jM} rather than β_j. A normal approximation for the ability estimate $\hat{\theta}_i$ is available with relatively little difficulty for q large, but there are problems in practice with the accuracy achieved.

If q is constant and n becomes large, then a normal approximation is available for $\hat{\boldsymbol{\beta}}$ but not for the $\hat{\theta}_i$. The normal approximation for $\hat{\boldsymbol{\beta}}$ is somewhat different than the conventional approximation expected in a logit model, for the asymptotic mean is $\boldsymbol{\beta}_M$ rather than $\boldsymbol{\beta}$, and the asymptotic covariance matrix is a relatively complicated expression. To define the required asymptotic covariance matrix requires a somewhat lengthy series of intermediate definitions. Let Y_{ij}^+ be the adjusted random variable with value $Y_{ij} - p_{kjM}$ for $Y_{i+} = k$. Let $\mathbf{V}^+$ be the covariance matrix of the q-dimensional vector $\mathbf{Y}_i^+$ with coordinates Y_{ij}^+ for $1 \leq j \leq q$. Let $V_{jj'}^+$ be row j and column j' of $\mathbf{V}^+$. Let

$$\sigma_{kjM}^2 = p_{kjM}(1 - p_{kjM})$$

be the variance of a Bernoulli random variable that is 1 with probability p_{kjM}, let

$$\sigma_{+jM}^2 = \sum_{k=1}^{q-1} P_k \sigma_{kjM}^2,$$

$$\sigma^2_{k+M} = \sum_{j=1}^{q} \sigma^2_{kjM},$$

and let

$$W_{jj'} = \sigma^2_{+jM}\delta_{jj'} - \sum_{k=1}^{q-1} P_k \sigma^2_{kjM}\sigma^2_{kj'M}/\sigma^2_{k+M}.$$

Let $\mathbf{W}$ be the q by q matrix with row j and column j' equal to $W_{jj'}$. Note that

$$\sum_{j'=1}^{q} W_{jj'} = \sum_{j=1}^{q} V^+_{jj'} = 0,$$

and $\mathbf{W}$ and $\mathbf{V}^+$ are symmetric and positive semi-definite. Let $\mathbf{W}^-$ be the Moore-Penrose inverse of $\mathbf{W}$, so that

$$\mathbf{W}\mathbf{W}^- = \mathbf{W}^-\mathbf{W} = \mathbf{I} - q^{-1}\mathbf{1}\mathbf{1}^T,$$

$$\mathbf{W}\mathbf{W}^-\mathbf{W} = \mathbf{W},$$

and

$$\mathbf{W}^-\mathbf{W}\mathbf{W}^- = \mathbf{W}^-,$$

where $\mathbf{I}$ is the q by q identity matrix (Stewart, 1973, p. 326). Let $\mathbf{K}$ be the q by q matrix with row i and column j equal to

$$K_{ij} = \begin{cases} 0, & i = 1, \\ 0, & i \neq j, \\ 1, & i = j > 1, \\ -1, & j = 1 < i, \end{cases}$$

so that $\hat{\boldsymbol{\beta}} = \mathbf{K}\hat{\boldsymbol{\gamma}}$. Let $\mathbf{K}^T$ be the transpose of $\mathbf{K}$.

Given the definitions of $\mathbf{V}^+$, $\mathbf{W}$, and $\mathbf{K}$, the following result can be derived.

Theorem 7 *Under the conditions of Theorem 2, $n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_M)$ converges in distribution to a multivariate normal random vector with mean equal to the q-dimensional zero vector $\mathbf{0}$ and covariance matrix $\mathbf{W}^-\mathbf{V}^+\mathbf{W}^-$, and $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_M)$ converges in distribution to a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{K}\mathbf{W}^-\mathbf{V}^+\mathbf{W}^-\mathbf{K}^T$.*

Proof. The normal approximation is derived by conventional arguments based on the function $\mathbf{F}$ developed in Section 1.4. Once again, fixed point theorems are employed. Details are omitted.

It should be noted that Theorem 7 differs from customary results for maximum likelihood both in terms of the asymptotic mean and in terms of the asymptotic covariance matrix. Were the customary results to hold, $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ would converge in distribution to a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{K}\mathbf{W}^{-}\mathbf{K}^{T}$ (Haberman, 1978, pp. 339-340).

If the number q of items increases, then normal approximations remain available, but a few changes in results are needed due to the changing dimension of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ and due to the behavior of $\mathbf{V}$ for large q. The normal approximations are somewhat unsatisfactory in many cases due to unconventional asymptotic mean. Let

$$\sigma_{+j}^2 = E(w(\theta_1 - \beta_j))$$

be the variance of $Y_{1j} - E(Y_{1j}|\theta_1)$. Let r be an integer constant greater than 1. For $q \geq r$, let $\hat{\boldsymbol{\beta}}_r$ be the r-dimensional vector with coordinates $\hat{\beta}_j$ for $1 \leq j \leq r$, let $\boldsymbol{\beta}_{rM}$ be the r-dimensional vector with coordinates β_{jM} for $1 \leq j \leq r$, and let $\boldsymbol{\beta}_r$ be the r-dimensional vector with coordinates β_j for $1 \leq j \leq r$. Let $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_r)$ be the r by r matrix with row j and column j' equal to

$$\frac{\delta_{jj'}}{\sigma_{+j}} - \frac{\delta_{j1} + \delta_{j'1} - 1}{\sigma_{+1}}.$$

Then the following theorem may be proven.

Theorem 8 *If q approaches ∞, then $n^{1/2}(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_{rM})$ converges in distribution to a multivariate normal random vector with mean the r-dimensional zero vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_r)$.*

Proof. Let $v_{kjj'C} = v_{kjj'}(\boldsymbol{\beta})$ be the conditional covariance of Y_{ij} and $Y_{ij'}$ given $Y_{i+} = k$, so that

$$v_{kjj'}(\boldsymbol{\beta}) = \frac{s_{kjj'}(\boldsymbol{\beta})}{s_k(\boldsymbol{\beta})} - m_{kj}(\boldsymbol{\beta})m_{kj'}(\boldsymbol{\beta}),$$

where

$$s_{kjj'}(\boldsymbol{\beta}) = \sum_{\mathbf{c}\in\Gamma(k)} c_j c_{j'} \exp(-\boldsymbol{\beta}^T \mathbf{c}).$$

Obviously $v_{kjjC} > 0$ for $0 < k < q$. As shown in the appendix, $v_{kjj'C}$ is negative if $j \neq j'$ and k is neither 0 nor q. As in the case of m_{kjC}, if k, q, and n are selected so that σ^2_{k+C} approaches ∞, then

$$\sigma^4_{k+C}\left[v_{kjjC} - \sigma^2_{kjC} + \frac{(2p_{kjC}-1)\sigma^2_{kjC}(2p_{kjC}-1-\psi_k)}{2\sigma^2_{k+C}}\right],\ 1 \leq j \leq q,$$

and

$$\sigma^4_{k+C}\left[v_{jj'kC} + \frac{\sigma^2_{kjC}\sigma^2_{kj'C}}{\sigma^2_{k+C}}\right],\ 2 \leq j < j' \leq q,$$

are uniformly bounded.

Use of the maximum norm shows that $n^{1/2}|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_M - \mathbf{Z}|$ converges in probability to 0 if

$$\mathbf{Z} = \mathbf{K}\mathbf{M},$$

$$\mathbf{M} = \mathbf{W}^{-}\mathbf{T},$$

and $\mathbf{T}$ is the q-dimensional vector with coordinates

$$T_j = n^{-1}\sum_{k=1}^{q-1}(f_{kj} - n_k m_{kjC}).$$

Indeed, use of large-deviation theory shows that $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_M - \mathbf{Z}|$ is of order $n^{-1}\log q$ (Haberman, 1977b). Because the sum of the coordinates of $\mathbf{T}$ is 0, if Ω' is the smallest value of $\sum_{k=1}^{q-1} P_k \sigma^2_{kjM}\sigma^2_{kj'M}/\sigma^2_{k+M}$, then

$$\mathbf{M} = (\mathbf{W} + \Omega'\mathbf{1}\mathbf{1}^T)^{-1}\mathbf{T}.$$

Elementary albeit somewhat tedious calculations show that the variance of $n^{1/2}(M_j - T_j/\sigma^2_{+j})$ approaches 0. Thus

$$n^{1/2}(\hat{\beta}_j - \beta_{jM} - T_j/\sigma^2_{+j} + T_1/\sigma^2_{+1})$$

converges in probability to 0, so that the conclusion of the theorem follows.

Theorem 8 is somewhat similar but not identical to the standard normal approximation for maximum-likelihood estimates. The asymptotic mean of $\hat{\boldsymbol{\beta}}_r$ would be expected to be

$\boldsymbol{\beta}_r$, and the asymptotic covariance matrix would be the limit of the submatrix of $\mathbf{K}\mathbf{W}^{-}\mathbf{K}^T$ formed from the first r rows and r columns. A slight change in the calculation of the variance of $n^{1/2}(M_j - T_j/\sigma_{+j}^2)$ can be used to verify that the limit of the submatrix is $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_r)$. Thus the major issue is the difference between $\boldsymbol{\beta}_M$ and $\boldsymbol{\beta}$.

In practice, the asymptotic normality result is somewhat unsatisfactory. Clearly $\hat{\beta}_j$ is intended to estimate β_j rather than β_{jM}. If n/q^2 approaches 0, then standard results hold, so that $n^{1/2}(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_r)$ converges in distribution to a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_r)$. This result is a little stronger than is found in the literature (Haberman, 1977b). Nonetheless, the asymptotic approximation is not very helpful for an SAT I exam with q of 60 or 78 and $n = 446,607$. Obviously, n/q^2 is too large. As a practical matter, the results indicate that ordinary asymptotic confidence intervals for β_j cannot be derived by use of the normal approximation for $\hat{\beta}_j$. Nonetheless, it should be emphasized that some estimation gain is achieved if the sample size n is large. Let $0 < \alpha < 1$ and let z be defined so that the probability is α and the absolute value of a standard normal deviate exceeds z. Let

$$\sigma^2(\hat{\beta}_j) = \frac{1}{\sigma_{+j}} + \frac{1}{\sigma_{+1}}.$$

For $j > 1$, the probability approaches $1 - \alpha$ that

$$\beta_{jM} - \beta_j - n^{-1/2} z\sigma(\hat{\beta}_j) \leq \hat{\beta}_j - \beta_j \leq \beta_{jM} - \beta_j + n^{-1/2} z\sigma(\hat{\beta}_j).$$

For a given number q of items, the bounds become more narrow as the sample size n becomes larger.

In the case of an individual i for an increasing number q of items, the normal approximation for $\hat{\theta}_i$ is relatively straightforward. Here

$$\hat{\theta}_i = \hat{\theta}_{kM} = g_k(\hat{\boldsymbol{\beta}})$$

if $Y_{i+} = k$. Let

$$\sigma_{ij}^2 = p_{ij}(1 - p_{ij})$$

be the variance of Y_{ij} given θ_i, let

$$\sigma_{i+}^2 = \sum_{j=1}^{q} \sigma_{ij}^2$$

be the variance of Y_{i+} given θ_i, and let

$$\sigma(\hat{\theta}_i) = 1/\sigma_{i+}.$$

Then it is a fairly straightforward matter to verify that $\hat{\theta}_i - \theta_i - (Y_{i+} - p_{i+})/\sigma_{i+})$ is of order q^{-1}. It follows that $(\hat{\theta}_i - \theta_i)/\sigma(\hat{\theta}_i)$ converges in distribution to a standard normal random variable. In addition, for r a finite integer, the r-dimensional vector with coordinates $(\hat{\theta}_i - \theta_i)/\sigma(\hat{\theta}_i)$ for $1 \leq i \leq r$ converges in distribution to a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{I}$.

Given the assumptions on the empirical distribution of the β_j, the asymptotic standard deviation $\sigma(\hat{\theta}_i)$ is readily shown to satisfy the condition that $q^{1/2}\sigma(\hat{\theta}_i)$ converges to $[1/E(w(\theta_i - \beta^*))]^{1/2}$ for each i.

In this case, the normal approximation for $\hat{\theta}_i$ does hold, so that approximate confidence intervals are available. The probability that

$$\hat{\theta}_i - z\hat{\sigma}(\hat{\theta}_i) < \theta_i < \hat{\theta}_i + z\hat{\sigma}(\hat{\theta}_i)$$

approaches $1 - \alpha$ if

$$\hat{\sigma}(\theta_i) = 1/\hat{\sigma}_{i+},$$

$$\hat{\sigma}^2_{i+} = \sum_{j=1}^{q} \hat{\sigma}^2_{ij},$$

and

$$\hat{\sigma}^2_{ij} = \hat{p}_{ij}(1 - \hat{p}_{ij}).$$

In practice, the normal approximation suggests limits on accuracy of estimation. Note that $\sigma^2_{ij} \leq \frac{1}{4}$, so that $\sigma(\hat{\theta}_i) \geq 2/q^{1/2}$. For a test with $q = 60$ items, $\sigma(\hat{\theta}_i) \geq 0.258$, while for a test with $q = 78$ items, $\sigma(\hat{\theta}_i) \geq 0.226$. For the SAT I Math, the observed estimates $\hat{\sigma}(\hat{\theta}_i)$ are all at least 0.309 and are greater than 1 in some instances in which $\hat{\theta}_i$ is finite. In the case of the Verbal exam, the minimum estimated standard deviation is 0.259 and the maximum for $\hat{\theta}_i$ finite is greater than 1. These large estimates also suggest limitations in the quality of the normal approximation.

Some obvious limitations are evident for the normal approximation for the $\hat{\theta}_i$. Observe that, for $1 \leq i < k \leq q$, the probability that $\hat{\theta}_i$ and $\hat{\theta}_k$ are unequal is the probability

that $Y_{i+} \neq Y_{k+}$. Thus the probability that, for randomly chosen i and k, $\hat{\theta}_i \neq \hat{\theta}_k$ is the probability that $Y_{1+} \neq Y_{2+}$. This probability is the Gini concentration

$$1 - \sum_{k=0}^{q} P_k^2$$

of Y_{1+} (Gini, 1912). An unbiased estimate of this measure is

$$\frac{n}{n-1}\left[1 - \sum_{k=0}^{q}(n_k/n)^2\right].$$

In the SAT I Math, this estimate is 0.976, while for the Verbal examination, the estimate is 0.980. In contrast, if each $\hat{\theta}_i$ had an independent normal distribution, then the probability that $\hat{\theta}_i \neq \hat{\theta}_k$ would be 1 for $i \neq k$.

Normal approximations for $\hat{H}_J$ and $\hat{H}_M$ are somewhat unsatisfactory in practice due to the relative large estimation biases involved.

1.6 Generalized Residuals

Generalized residuals based on JMLE can be considered for the Rasch model as a tool for detection of model deficiencies (Haberman, 1978, chap. 5). Several possible approaches exist, but none is very satisfactory. For a simple possibility, consider fixed constants d_{ij} for $1 \leq i \leq n$ and $1 \leq j \leq q$. Assume that d_{ij} is not additive in i and j, so that no a_i, $1 \leq i \leq n$, and b_j, $1 \leq j \leq q$, exist such that $d_{ij} = a_i + b_j$ for $1 \leq i \leq n$ and $1 \leq j \leq q$. The raw generalized residual corresponding to the d_{ij} is

$$e = \sum_i \sum_j d_{ij}(Y_{ij} - \hat{p}_{ij}) = O - \hat{E},$$

where

$$O = \sum_i \sum_j d_{ij} Y_{ij}$$

and

$$\hat{E} = \sum_i \sum_j d_{ij} \hat{p}_{ij}$$

is the estimated expected value of O. The adjusted generalized residual is

$$z = e/\hat{\sigma}(e),$$

where $\hat{\sigma}^2(e)$ is the minimum of

$$\sum_i \sum_j \hat{\sigma}_{ij}^2 (d_{ij} - a_i - b_j)^2$$

for real a_i and b_j such that $b_1 = 0$.

For a very simple example, consider a check of whether a set A of N_A examinees has unusual behavior on a set B of N_B items. A possible choice is $d_{ij} = 1$ for i in A and j in B and $d_{ij} = 0$ otherwise. If N_A/n is very small and $\hat{\sigma}_{iA}^2$ is $\sum_{j \in B} \hat{\sigma}_{ij}^2$, then $\hat{\sigma}^2(e)$ is well approximated by

$$\sum_{i \in A} \frac{\hat{\sigma}_{iA}^2 (\hat{\sigma}_{i+}^2 - \hat{\sigma}_{iA}^2)}{\hat{\sigma}_{i+}^2}.$$

In customary asymptotic theory, z has an approximate standard normal distribution if the model holds and if

$$[\sigma(e)]^{-1} \max_i \max_j |d_{ij}|$$

approaches 0, where $\sigma^2(e)$ is the minimum of

$$\sum_i \sum_j \sigma_{ij}^2 (d_{ij} - a_i - b_j)^2$$

for real a_i and b_j such that $b_1 = 0$ (Haberman, 1978). In the context of joint estimation, one must consider the case in which the sample size n and the number of items q both approach ∞ and the d_{ij} depend on n and q. It is clearly necessary that the number of nonzero d_{ij} approach ∞.

Unfortunately, the requirement on the number of nonzero d_{ij} creates a problem with the linear approximations to the joint maximum-likelihood estimates on which the generalized residuals are based. Because $p_{kjM} - p_{kjC}$ is of order q^{-1}, the difference $\hat{p}_{ij} - p_{ij}$ cannot be expected to be of order less than q^{-1} even if the sample size is quite large. With some calculation that exploits the fact that $\hat{p}_{ij} = w(\hat{\mu}_{ij})$ for the bounded and continuous function w defined in Section 1.4, it is possible to show that z converges in distribution to a standard normal random variable if $q^{-1}u/\sigma(e)$ approaches 0, where u is the minimum of $\sum_{i=1}^n \sum_{j=1}^q |d_{ij} - a_i - b_j|$ for a_i and b_j real. This condition is quite restrictive in practice.

For example, recall the case with the set A of examinees and the set B of items. Then the condition on $q^{-1}u/v$ only holds if $N_A N_B/q^2$ approaches 0. Nonetheless, one rather

simple application is of some interest. For the SAT I data under study, consider a Rasch model in which the Verbal and Math tests are combined together into a test with 138 items. Let the last 60 items come from the Math test. For individual h, consider the number of correct responses on the Math test. In this fashion, one might consider $d_{ij} = 1$ for $i = h$ and $j \geq 79$ and $d_{ij} = 0$ otherwise. Thus

$$O = \sum_{j=79}^{138} Y_{ij}$$

and

$$E = \sum_{j=79}^{138} \hat{p}_{ij}.$$

Because the number n of items is very large, $\sigma(e)$ is nearly equal to $\hat{\sigma}^2_{iM}\hat{\sigma}^2_{iV}/\hat{\sigma}^2_{i+}$ if

$$\hat{\sigma}^2_{iV} = \sum_{j=1}^{78} \hat{\sigma}_{ij}$$

and

$$\hat{\sigma}^2_{iM} = \sum_{j=79}^{138} \hat{\sigma}_{ij}.$$

Although concerns about the accuracy of the normal approximation are clearly in order if $\sigma(e)$ is relatively small, it is noteworthy that this form of residual analysis is quite adequate to indicate severe problems with the combined Rasch model. Some 144,082 examinees exist with generalized residuals that exceed 2 in magnitude out of 446,603 examinees with a positive value of v. If the Rasch model really held, then the normal approximation implies that only about 23,000 examinees would be expected to have generalized residuals that exceed 2 in absolute value. Of course, no reasonable person would expect a combined Rasch model to apply to tests as different as the Verbal and Math tests. Nonetheless, it is useful to note that analysis of generalized residuals for individual examinees can detect this problem in a substantial fraction of all examinees.

1.7 Model Error

Some further complications arise when the model is not simply assumed to be true. There does not appear to be any treatment of this case in the literature, but existing

arguments can be modified easily in this case (Haberman, 1977b; Haberman, 1989; Gilula & Haberman, 1994; Gilula & Haberman, 1995).

Consider the case in which the Y_{ij} need not satisfy the Rasch model. Assume that the $\mathbf{Y}_i$ are independent and identically distributed and the probability $p_J(\mathbf{c})$ that $\mathbf{Y}_i = \mathbf{c}$ is positive for each q-dimensional vector $\mathbf{c}$ with each coordinate 0 or 1. Let P_k still be the probability that $Y_{1+} = k$, and let m_{kjC} be the conditional probability that $Y_{1j} = 1$ given that $Y_{1+} = k$ for $1 \leq j \leq q$ and $0 \leq k \leq q$. Note that for $k = 0$, $m_{kjC} = 0$, while for $k = q$, $m_{kjC} = 1$. Let p_j^Y be the probability that $Y_{1j} = 1$. In this case, the same arguments used to verify existence of joint maximum-likelihood estimates also imply the existence of θ_{kM} and β_{jM} such that $\beta_{1M} = 0$,

$$p_{kjM} = \text{lgt}(\theta_{kM} - \beta_{jM}),$$
$$\sum_{j=1}^{q} p_{kjM} = \sum_{j=1}^{q} m_{kjC} = k,$$

and

$$\sum_{k=1}^{q-1} P_k p_{kjM} = \sum_{k=1}^{q-1} P_k m_{kjC} = p_j^Y - P_q.$$

One may define H_J and H_M as in the case in which the Rasch model is valid.

For q fixed, it remains true that $\hat{\beta}_j$ converges almost surely to β_{jM} and $\hat{\theta}_{kM}$ converges almost surely to θ_{kM}. For asymptotic normality, define $\mathbf{V}^+$ and $\mathbf{W}$ as in the case of the Rasch model correct. Then it is a relatively straightforward matter to demonstrate that $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_M)$ converges in distribution to a multivariate normal random vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{K}\mathbf{W}^-\mathbf{V}^+\mathbf{W}^-\mathbf{K}^T$. It remains true that $\hat{H}_J$ converges almost surely to H_J and $\hat{H}_M$ converges almost surely to H_M.

The results for q fixed are readily generalized to the case of q increasing by use of arguments similar to those required for the case in which the model is correct. Wording of results must be modified to some degree because β_{jM} depends on q. It is simplest to assume that the β_{jM} are uniformly bounded as q increases, as is the case if the Rasch model is valid. To avoid somewhat pathological cases, it is also helpful to assume, as is the case for the Rasch model, that positive constants α_1 and α_2 exist such that, for any $q \geq 1$,

$$\alpha_1 \mathbf{x}^T\mathbf{x} \leq \mathbf{x}^T\mathbf{V}^+\mathbf{x} \leq \alpha_2 \mathbf{x}^T\mathbf{x}$$

whenever $\sum_{j=1}^{q} x_j = 0$. It remains true that $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_M|$ converges in probability to 0. Asymptotic normality results are also available, although some simplifications do not apply. Let $\sigma_e^2(\hat{\beta}_j)$ be row j and column j of of the matrix $\mathbf{K}\mathbf{W}^{-}\mathbf{V}\mathbf{W}^{-}\mathbf{K}^T$. Then the standardized value $n^{1/2}(\hat{\beta}_j - \beta_{jC})/\sigma_e(\hat{\beta}_j)$ converges in distribution to a standard normal deviate.

On the whole, use of JMLE appears to have an intermediate status. Parameter estimation is feasible to the extent that the $\hat{\beta}_j$ and $\hat{\theta}_i$ do approximate the quantities they estimate under realistic testing conditions if the Rasch model holds. On the other hand, severe limitations exist on basic tools for statistical inference such as approximate confidence intervals and generalized residuals. It appears difficult to advocate use of JMLE except for preliminary estimation of parameters.

2 Conditional Maximum Likelihood

Conditional maximum-likelihood estimation is applicable to the Rasch model (Andersen, 1973a). In this case, inference is conditional on the observed examinee sums Y_{i+}. For $\mathbf{c}$ in $\Gamma(k)$ and for $0 \leq k \leq q$, the conditional probability $p_{JC}(\mathbf{c})$ that $\mathbf{Y}_i = \mathbf{c}$ given that $Y_{i+} = k$ satisfies

$$p_{JC}(\mathbf{c}) = p_J(\mathbf{c})/P_k.$$

Under the Rasch model,

$$P_k = s_k(\boldsymbol{\beta}) \int_{-\infty}^{\infty} e^{k\theta} \Psi(\boldsymbol{\beta}, \theta) dD(\theta),$$

so that

$$p_{JC}(\mathbf{c}) = \exp(-\boldsymbol{\beta}^T \mathbf{c})/s_k(\boldsymbol{\beta}) \tag{17}$$

does not depend on the distribution function D of the ability θ_1. The conditional log likelihood function is then

$$\ell_C(\mathbf{p}_{JC}) = \sum_{i=1}^{n} \log p_{JC}(\mathbf{Y}_i)$$

for the array $\mathbf{p}_{JC}$ of $p_{JC}(\mathbf{c})$ for $\mathbf{c}$ in Γ. Thus

$$\ell_C(\boldsymbol{\beta}) = -\boldsymbol{\beta}^T \mathbf{Y}_+ - \sum_{k=0}^{q} n_k \log s_k(\boldsymbol{\beta}).$$

Because

$$\mathbf{Y}_{+j} = \sum_{k=0}^{q} f_{kj},$$

$\ell_C(\boldsymbol{\beta})$ is determined by the f_{kj}, and inferences again may be based on the collapsed table. The relationship of conditional and marginal maximum likelihood is relatively simple. Let $\mathbf{P}$ be the vector with coordinates P_k for $0 \leq k \leq q$, and let

$$\ell_S(\mathbf{P}) = \sum_{k=0}^{q} n_k \log P_k$$

be the marginal log likelihood for the examinee totals Y_{i+}, $1 \leq i \leq n$, under the unrestricted model that $Y_{i+} = k$ with probability P_k for some nonnegative P_k such that $\sum_{k=0}^{q} P_k = 1$. Then

$$\ell(\mathbf{p}_J) = \ell_C(\mathbf{p}_{JC}) + \ell_S(\mathbf{P}).$$

Let ℓ_{CM} be the maximum of $\ell_C(\mathbf{p}_{JC})$ under the constraint that (17) holds for some $\boldsymbol{\beta}$ with $\beta_1 = 0$. Let ℓ_{SM} be the maximum

$$\sum_{k=0}^{q} n_k \log(n_k/n)$$

of $\ell_S(\mathbf{P})$. Then

$$\ell_M \leq \ell_{CM} + \ell_{SM}.$$

As discussed in Section 3, it is commonly true that ℓ_M is the sum of $\ell_{CM} + \ell_{SM}$.

The conditional maximum-likelihood estimate $\hat{\boldsymbol{\beta}}_C$, if it exists, satisfies $\hat{\beta}_{1C} = 0$,

$$\hat{p}_{JC}(\mathbf{c}) = \exp(-\hat{\boldsymbol{\beta}}_C^T \mathbf{c})/s_k(\hat{\boldsymbol{\beta}}_C)$$

for $\mathbf{c}$ in $\Gamma(k)$ and $0 \leq k \leq q$, and

$$\ell_C(\hat{\mathbf{p}}_{JC}) = \ell_{CM}.$$

The notable feature here is that the conditional log-likelihood function does not involve the common distribution function D of the examinee abilities θ_i. If $\hat{\boldsymbol{\beta}}_C$ exists, then it satisfies the conditional maximum-likelihood equations

$$\hat{m}_{kjC} = m_{kj}(\hat{\boldsymbol{\beta}}_C)$$

and

$$\sum_{k=0}^{q} n_k \hat{m}_{kjC} = Y_{+j}$$

for $1 \leq j \leq q$. Conversely, if $\boldsymbol{\beta}'_C$ satisfies the conditions that $\beta'_{1C} = 0$ and

$$\sum_k n_k m_{kj}(\boldsymbol{\beta}'_C) = Y_{+j},$$

then $\boldsymbol{\beta}'_C$ is a conditional maximum-likelihood estimate of $\boldsymbol{\beta}$. Provided that $n_0 + n_q < n$, no more than one conditional maximum-likelihood estimate $\hat{\boldsymbol{\beta}}_C$ exists. If $n_0 + n_q$ is n, then $\ell_C(\boldsymbol{\beta})$ is constant, so that any q-dimensional vector with first coordinate 0 is a conditional maximum-likelihood estimate. In this instance, the arbitrary choice of $\hat{\beta}_{jC} = 0$ for all j may be made.

Existence of conditional maximum-likelihood estimates is an issue, although normally a much less important one than in the case of joint estimation. Let $N(\mathbf{c})$, $\mathbf{c}$ in $\Gamma(k)$, $0 \leq k \leq q$, be the number of examinees i with $Y_{ij} = c_j$ for each item j. To find existence conditions for $\hat{\boldsymbol{\beta}}_C$ for the case of $n_0 + n_q < n$, let δ_{ab} be the Kronecker δ function with $\delta_{ab} = 1$ for $a = b$ and $\delta_{ab} = 0$ for $a \neq b$. The following theorem is then available.

Theorem 9 *The conditional maximum likelihood estimate $\hat{\boldsymbol{\beta}}_C$ fails to exist if, and only if, α_j, $1 \leq j \leq q$, and γ_k, $0 < k < q$, can be found such that the following conditions hold:*

1. *If k is in K, $0 < k < q$, and $\mathbf{c}$ is in $\Gamma(k)$, then*

$$\tau(\mathbf{c}) = \sum_j \delta_{c(j)1}\alpha_j + \gamma_k \leq 0.$$

2. *For some integer k in K such that $0 < k < q$ and some $\mathbf{c}$ in $\Gamma(k)$, $\tau(\mathbf{c}) < 0$.*

3. *The product $N(\mathbf{c})\tau(\mathbf{c}) = 0$ for all $\mathbf{c}$ in $\Gamma(k)$ and k in K such that $0 < k < q$*

(Haberman, 1974, chap. 2). Obviously, $\hat{\boldsymbol{\beta}}_C$ must exist if $N(\mathbf{c}) > 0$ for each $\mathbf{c}$ in $\Gamma(k)$ and each k in K such that $0 < k < q$. An equivalent result is available with a somewhat different appearance.

Theorem 10 *The conditional maximum likelihood estimate $\hat{\boldsymbol{\beta}}_C$ exists if, and only if, there is no real a and b such that the following conditions hold:*

1. *a and b are not integers,*

2. $1 < a < q - 1$ *and* $n_0 < b < n - n_0$,

3. $f_{kj} = 0$ *for* $k < a$ *and* $Y_{+j} < b$,

4. $f_{kj} = n_k$ *for* $k > a$ *and* $Y_{+j} > b$,

5. k *in* K *and* j *exist such that* $n_k > 0$ *and either* $0 < k < a$ *and* $Y_{+j} < b$ *or* $a < k < q$ *and* $Y_{+j} > b$.

Proof. A simple modification of existing proofs for JMLE is required (Haberman, 1977b, pp. 821–822).

Existence results presented here appear to be consistent with but simpler than those previously available (Fischer, 1981).

The fundamental change relative to JMLE is that n_0 and n_q need not be 0. Instead of the requirement that Y_{+j} not be 0 in order for joint maximum-likelihood estimates to exist, it is now necessary that Y_{+j} exceeds n_0 but is less than $n - n_0$. It is important to observe that conditional maximum-likelihood estimates exist if unconditional maximum-likelihood estimates exist.

Extended conditional maximum-likelihood estimates may be considered if $\hat{\boldsymbol{\beta}}_C$ does not exist. An array $\hat{\mathbf{p}}_{JC}$ of extended conditional maximum-likelihood estimates $\hat{p}_{JC}(\mathbf{c})$ of $p_J(\mathbf{c})$, $\mathbf{c}$ in Γ, exists such that, whenever $\ell_C(\mathbf{p}_{JC})$ approaches ℓ_{CM} for $\mathbf{p}_{JC}$ such that (17) holds and $\beta_1 = 0$, $p_{JC}(\mathbf{c})$ approaches $\hat{p}_{JC}(\mathbf{c})$ for $\mathbf{c}$ in $\Gamma(k)$ and either k in K or k equal 0 or q. The convention is adopted that $\hat{p}_{JC}(\mathbf{c}) = k!(q-k)!/q!$ if $n_0 + n_q = n$. If $\hat{\boldsymbol{\beta}}_C$ exists, then the new definition reduces to the conventional definition of $\hat{p}_{JC}(\mathbf{c})$. If $\hat{m}_{kjC}$ is the sum of $\hat{p}_{JC}(\mathbf{c})$ for $\mathbf{c}$ in $\Gamma(k)$ with $c_j = 1$, then

$$\sum_{k=0}^{q} n_k \hat{m}_{kjC} = Y_{+j}$$

and

$$\sum_{j=1}^{q} \hat{m}_{kjC} = 1$$

for k in K. If the conditional maximum-likelihood estimate $\hat{\boldsymbol{\beta}}_C$ exists, then $\hat{m}_{kjC} = m_{kj}(\hat{\boldsymbol{\beta}}_C)$ for k in K. Various conventions can be considered to define $\hat{\boldsymbol{\beta}}_C$ in the case in which no conditional maximum-likelihood estimate exists for $\boldsymbol{\beta}$.

Given the estimates $\hat{\beta}_{jC}$, it is possible to estimate the examinee abilities θ_i. For each i, the log likelihood for θ_i given the $\hat{\beta}_{jC}$ is

$$\sum_{j=1}^{q}\{Y_{ij}\log \mathrm{lgt}(\theta_i - \hat{\beta}_{jC}) + (1 - Y_{ij})\log[1 - \mathrm{lgt}(\theta_i - \hat{\beta}_{jC})]\}.$$

The maximum is achieved by solving the equation

$$\hat{p}_{i+C} = Y_{i+}$$

for

$$\hat{p}_{ijC} = \mathrm{lgt}(\hat{\theta}_{iC} - \hat{\beta}_{jC})$$

and

$$\hat{p}_{i+C} = \sum_{j=1}^{q}\hat{p}_{ijC}.$$

If $Y_{i+} = q$, $\hat{\theta}_{iC} = \infty$, while if $Y_{i+} = 0$, $\hat{\theta}_{iC} = -\infty$.

2.1 Large-sample Properties

For q fixed, if the Rasch model is valid and n becomes large, then $\hat{\boldsymbol{\beta}}$ is a consistent and asymptotically normal estimate for $\boldsymbol{\beta}$ (Andersen, 1973a; Haberman, 1977a). The argument in the second citation permits generalization to the case of q increasing. Let

$$V_{jj'C} = V_{jj'}(\boldsymbol{\beta}) = \sum_{k=1}^{q-1} n_k v_{kjj'}(\boldsymbol{\beta})$$

be the conditional covariance of Y_{+j} and $Y_{+j'}$ given n_k for $0 \le k \le q$, so that $V_{jj'C} < 0$ if $j \ne j'$ and if $n_0 + n_q < n$ and V_{jjC} is positive for $n_0 + n_q < n$. Let $\mathbf{V}_C = \mathbf{V}(\boldsymbol{\beta})$ be the q by q matrix with row j and column j' equal to $V_{jj'}(\boldsymbol{\beta})$ for $1 \le j \le q$ and $1 \le j' \le q$. This matrix is of rank $q - 1$ if $n_0 + n_q < n$. Let $\mathbf{V}^*$ be the expected value of $n^{-1}\mathbf{V}_C$, so that $\mathbf{V}^*$ is obtained from $\mathbf{V}_C$ by substitution of P_k for n_k. Note that if Y_{ij}^* is the random variable equal to $Y_{ij} - m_{kjC}$ for $Y_{i+} = k$ and if $\mathbf{Y}_i^*$ is the vector with coordinates Y_{ij}^* for $1 \le j \le q$, then $\mathbf{V}^*$ is the covariance matrix of $\mathbf{Y}_i^*$ for each observation i. Arguments rather similar to those applied in the case of joint maximum-likelihood estimation may also be applied to conditional maximum-likelihood estimation. If the number q of items is fixed, then $\hat{\boldsymbol{\beta}}_C$

converges almost surely to $\boldsymbol{\beta}$ and $n^{1/2}(\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta})$ converges in distribution to a multivariate normal random variable with mean $\mathbf{0}$ and covariance matrix $\mathbf{K}\mathbf{V}^{*-}\mathbf{K}^T$. If q approaches ∞, then $\max_{1\leq j\leq q}|\hat{\beta}_{jC} - \beta_j|$ converges in probability to 0. For an integer $r \geq 1$, let $\hat{\boldsymbol{\beta}}_{rC}$ be the r-dimensional vector with coordinates $\hat{\beta}_j$ for $1 \leq j \leq r$, and let $\boldsymbol{\beta}_r$ be the r-dimensional vector with coordinates β_j for $1 \leq j \leq r$. Then $n^{1/2}(\hat{\boldsymbol{\beta}}_{rC} - \boldsymbol{\beta}_r)$ converges in distribution to a multivariate normal random vector with mean $\mathbf{0}$ and the covariance matrix $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_r)$ encountered in the discussion of the normal approximation for $n^{1/2}(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\beta}_{rM})$. The gain for conditional estimation is quite major, for the asymptotic approximations involve the actual parameters of interest, namely the β_j, rather than the β_{jM} parameters. As in JMLE, it should be noted that $\boldsymbol{\Sigma}(\hat{\boldsymbol{\beta}}_r)$ is the limit of the matrix formed from the first r rows and columns of $\mathbf{K}\mathbf{V}^{*-}\mathbf{K}^T$. Let $\hat{\mathbf{V}}$ be $\mathbf{V}(\hat{\boldsymbol{\beta}}_C)$. Then both for q fixed and q increasing, asymptotic confidence intervals for parameters such as β_j are easily constructed by estimation of the asymptotic standard deviation $s(\hat{\beta}_{jC})$ of $\hat{\beta}_{jC}$ by the square root of the jth row and jth column of $\mathbf{K}\hat{\mathbf{V}}^{-}\mathbf{K}^T$.

If q approaches ∞, then the asymptotic properties of $\hat{\theta}_{iC}$ are essentially the same as those for $\hat{\theta}_i$ as far as consistency, asymptotic normality, and approximate confidence intervals are concerned. Estimation of the distribution of θ_1 can be implemented in essentially the same fashion as in JMLE by substitution of $\hat{\theta}_{iC}$ for $\hat{\theta}_i$.

Estimation of the entropy measures H_C and H_U involves relatively little difficulty, for H_C may be estimated by

$$\hat{H}_{CR} = -\frac{1}{nq}\ell_{CM},$$

H_+ may be estimated by

$$\hat{H}_+ = -\frac{1}{nq}\ell_{SM},$$

and H_U may be estimated by

$$\hat{H}_{UR} = \hat{H}_{CR} + \hat{H}_+.$$

For q constant, $\hat{H}_{CR}$ converges almost surely to H_C, $\hat{H}_+$ converges almost surely to H_+, and $\hat{H}_{UR}$ converges almost surely to H_U. For q increasing, $\hat{H}_{CR} - H_C$, $\hat{H}_+ - H_+$, and $\hat{H}_{UR} - H_U$ all converge in probability to 0. Normal approximations are readily available, at least if q/n approaches 0. Let $\sigma(\hat{H}_U)$ be the standard deviation of $q^{-1}\log p_J(\mathbf{Y}_1)$, and let $\sigma(\hat{H}_C)$ be

the standard deviation of $q^{-1} \log p_{JC}(\mathbf{Y}_1)$. For a fixed number q of items, $n^{1/2}(\hat{H}_{UR} - H_U)$ converges in distribution to a normal random variable with mean 0 and variance $\sigma^2(\hat{H}_{UR})$, and $n^{1/2}(\hat{H}_{CR} - H_C)$ converges in distribution to a normal random variable with mean 0 and variance $\sigma^2(\hat{H}_{CR})$. For q increasing, one may exploit the conditional independence of the Y_{ij} given θ_i to examine the distribution of the average

$$q^{-1} \log p_J(\mathbf{Y}_i) = q^{-1} \sum_{j=1}^{q} H(Y_{ij}, p_{ij})$$

given θ_i and the distribution of Y_{i+} given θ_i. One may demonstrate that both $\sigma(\hat{H}_{CR})$ and $\sigma(\hat{H}_{UR})$ converge to the standard deviation τ of $\kappa(\theta_1)$, where

$$\kappa(t) = E(H(\mathrm{lgt}(t - \beta^*), \mathrm{lgt}(t - \beta^*))).$$

In this case, both $n^{1/2}(\hat{H}_{UR} - H_U)$ and $n^{1/2}(\hat{H}_{CR} - H_C)$ converge in distribution to a normal random variable with mean 0 and variance τ^2. These results are readily applied to construction of approximate confidence intervals for H_C and H_U (Gilula & Haberman, 1995). It is sometimes worth noting that

$$\sigma^2(\hat{H}_U) = q^{-2} \boldsymbol{\beta}^T \mathbf{V}^* \boldsymbol{\beta} + \sigma^2(z)$$

for a random variable z with value

$$q^{-1} \left\{ \sum_{j=1}^{q} \beta_j m_{kjC} + \log[s_k(\boldsymbol{\beta})/P_k)] \right\}$$

for $Y_{1+} = k$. Estimation of $\sigma^2(\hat{H}_U)$ and $\sigma^2(\hat{H}_C)$ is straightforward. Let

$$\hat{p}_J(\mathbf{c}) = \hat{p}_{JC}(\mathbf{c}) n_k / n$$

for $\mathbf{c}$ in $\Gamma(k)$, and let

$$\hat{p}_{JCi} = \hat{p}_{JC}(\mathbf{Y}_i)$$

and

$$\hat{p}_{Ji} = \hat{p}_J(\mathbf{Y}_i).$$

Then $\sigma^2(\hat{H}_U)$ may be estimated by

$$\hat{\sigma}^2(\hat{H}_{UR}) = n^{-1} \sum_{i=1}^{n} [-q^{-1} \log \hat{p}_{Ji} - \hat{H}_{UR}]^2,$$

and $\sigma^2(\hat{H}_C)$ may be estimated by

$$\hat{\sigma}^2(\hat{H}_{CR}) = n^{-1}\sum_{i=1}^{n}[-q^{-1}\log\hat{p}_{JCi} - \hat{H}_{CR}]^2.$$

For comparisons, it is often useful to consider entropy measures under the assumptions that $\beta_j = 0$, so that item difficulties are all the same and under the assumptions that θ_i is constant, so that examinees all have the same ability. Under the model of constant item difficulties, the conditional probability $p_{JC}(\mathbf{c})$ must be $k!(q-k)!/q!$ for $\mathbf{c}$ in $\Gamma(k)$, so that the expected log penalty per item

$$H_{CA} = q^{-1}\sum_{k=0}^{q} P_k \log\left[\frac{q!}{k!(q-k)!}\right]$$

for $\mathbf{Y}_1$ given Y_{1+} and the expected logarithmic penalty per item for Y_{1+} is

$$H_{UA} = H_{CA} + H_{+}.$$

The obvious estimates are

$$\hat{H}_{CA} = (nq)^{-1}\sum_{k=0}^{q} n_k \log\left[\frac{q!}{k!(q-k)!}\right]$$

and

$$\hat{H}_{UA} = \hat{H}_{CA} + \hat{H}_{+},$$

respectively. Normal approximations and approximate confidence intervals are easily obtained (Gilula & Haberman, 1994; Gilula & Haberman, 1995). For the case of no ability effects, the Rasch model then is equivalent to the model so that the Y_{ij} are independently distributed. In this case, the expected logarithmic penalty per item for $\mathbf{Y}_{1+}$ is

$$H_{UI} = q^{-1}\sum_{j=1}^{q} H(p_j^Y, p_j^Y).$$

The obvious estimate is

$$\hat{H}_{UI} = q^{-1}\sum_{j=1}^{q} H(f_{+j}/n, f_{+j}/n).$$

2.2 The Newton-Raphson Algorithm

The Newton-Raphson algorithm is readily applied to computation of $\hat{\boldsymbol{\beta}}_C$ when the ordinary conditional maximum-likelihood estimate exists (Andersen, 1972). One begins with a preliminary approximation $\boldsymbol{\beta}_0$ to $\hat{\boldsymbol{\beta}}_C$, typically $\hat{\boldsymbol{\beta}}$. One then uses the iterations

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t - \mathbf{K}[\mathbf{V}(\boldsymbol{\beta}_t)]^-[\mathbf{f}_+ - \mathbf{m}_+(\boldsymbol{\beta}_t)],$$

where $\mathbf{f}_+$ is the q-dimensional vector with elements

$$f_{+j} = \sum_{k=1}^{q-1} f_{kj} = Y_{+j} - n_q$$

and $\mathbf{m}_+(\boldsymbol{\beta}_t)$ is the q-dimensional vector with elements

$$m_{+j}(\boldsymbol{\beta}_t) = \sum_{k=1}^{q-1} n_k m_{kjC}(\boldsymbol{\beta}_t).$$

In typical cases, $\boldsymbol{\beta}_t$ converges quite rapidly to $\hat{\boldsymbol{\beta}}_C$.

This algorithm has traditionally been difficult to apply for large values of q; however, considerable simplification in computations may be achieved by exploitation of the random variables U_{kj} previously used to study $m_{kjC} = m_{kj}(\boldsymbol{\beta})$ and $v_{kjj'C} = v_{kjj'}(\boldsymbol{\beta})$. Recall that, for $1 \leq k \leq q-1$,

$$m_{kjC} = \frac{p_{kjC} P(U_{k+} - U_{kj} = k-1)}{P(U_{k+} = k)}.$$

Note also that

$$v_{kjjC} = m_{kjC} - m_{kjC}^2$$

and

$$v_{kjj'C} = \frac{p_{kjC} p_{kj'C} P(U_{k+} - U_{kj} - U_{kj'} = k-2)}{P(U_{k+} = k)} - m_{kjC} m_{kj'C}$$

for $j \neq j'$. Given $\boldsymbol{\beta}$, computation of θ_{kC} in the definition of p_{kjC} may be accomplished by use of the Newton-Raphson algorithm customarily used with maximum-likelihood estimation in log-linear models (Haberman, 1978). For an initial approximation θ_{k0} of θ_{kC}, one uses the iteration

$$\theta_{k(t+1)} = \theta_{kt} + \frac{k - \sum_{j=1}^q \mathrm{lgt}(\theta_{kt} - \beta_j)}{\sum_{j=1}^q w(\theta_{kt} - \beta_j)}.$$

The θ_{kt} normally converge rapidly to θ_{kC}. There is no need for a precise approximation to θ_{kC}. Given θ_{kC}, p_{kjC} is easily computed. At this point, computations reduce to the problem of finding the probability that the sum of independent Bernoulli variables has a specified value. A simple recursion formula is quite adequate for this purpose. It suffices to consider $P(U_{k+} = k)$. Arguments for $U_{k+} - U_{kj}$, the sum of $U_{kj'}$ for $j' \neq j$ are essentially the same. Similar remarks apply to $U_{k+} - U_{kj} - U_{kj'}$ for $j \neq j'$. Let a_{khi} be the probability that $\sum_{j=1}^{i} U_{kj} = h$ for $0 \leq h \leq i$ and $1 \leq i \leq q$. Then $a_{k01} = 1 - p_{k1C}$ and $a_{k11} = p_{k1C}$. For $0 \leq i < q$ and $h = 0$,

$$a_{k0(i+1)} = a_{k0i}(1 - p_{k(i+1)C}).$$

For $h = i + 1$,

$$a_{k(i+1)(i+1)} = a_{kii}p_{k(i+1)C}.$$

For $1 \leq h \leq i$,

$$a_{kh(i+1)} = a_{k(h-1)i}p_{k(i+1)C} + a_{khi}(1 - p_{khi}).$$

Thus $P(U_{k+} = k) = a_{kkq}$. In the course of calculations, it is helpful to note that there is never a need for a_{khi} for $h > k$ or for $h < k - q + i$. Minor changes in the algorithm are appropriate for $k > q/2$. In this case, p_{kiC} is replaced by $1 - p_{kiC}$ in the algorithm, and $P(U_{k+} = k)$ is then $a_{k(q-k)q}$. Alternative approaches to computations can also be considered that are comparable in terms of computational labor, although these methods do not involve the scaling procedures used here to prevent sums from becoming excessively small (Gustafsson, 1980; Liou, 1994).

Given that the table of f_{kj} has already been prepared and given that starting values based on joint estimation are employed, computation time for the Verbal test with $q = 78$ was 10 seconds and for the Math test with $q = 60$ was 3 seconds on an IBM NetVista M41 computer with 512 megabytes of RAM running Windows NT. The Intel Pentium 4 processor ran at 1.8 gigahertz. The stopping rule was that $|\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_t|$ was no greater than 0.00001. The starting values based on JMLE were reasonably successful. The maximum difference between joint and conditional maximum-likelihood estimates of a parameter β_j was 0.097 for the Verbal test and 0.126 for the Math test. These differences are sufficiently small to justify use of joint estimation for starting values for conditional estimation. At the

same time, the differences are rather large relative to the estimated asymptotic standard deviations of the parameters. For instance, $\hat{\beta}_{jC}$ has an estimated asymptotic standard deviation no greater than 0.008 in the Verbal and Math examples.

It should be emphasized that the Newton-Raphson algorithm is relatively stable, so that a choice of crude starting values will not normally prevent convergence, although it may lead to somewhat slower computations. Because the computational cost of JMLE is very low, this approach to starting values appears appropriate.

Estimates of H_C and H_U are readily obtained, together with estimated asymptotic standard deviations. For the Math test, $\hat{H}_{CR} = 0.41667$, and $\hat{H}_{UR} = 0.48040$. The estimated asymptotic standard deviations of these statistics are quite small, about 0.00063. For the Verbal test, $\hat{H}_{CR} = 0.47497$, and $\hat{H}_{UR} = 0.52630$. The estimated asymptotic standard deviations are about 0.00071. There is a substantial difference between the estimated expected log penalties for the Rasch model and for models in which the Rasch model holds and either ability parameters are constant or item difficulties are constant. For the Math test, $\hat{H}_{UI} = 0.55715$ and $\hat{H}_{UA} = 0.55492$. For the Verbal test, $\hat{H}_{UI} = 0.58440$ and $\hat{H}_{UA} = 0.58355$.

2.3 Generalized Residuals

In contrast to JMLE, generalized residuals are readily available for CMLE. In a very general formulation, for each examinee i, real weights $d_i(\mathbf{c})$ are assigned for q-dimensional vectors $\mathbf{c}$ with coordinates 0 or 1. L For $\mathbf{Y}_i$ in $\Gamma(k)$, the estimated conditional expected value of $d_i(\mathbf{Y}_i)$ given $Y_{i+} = k$ is

$$\hat{D}_i = \sum_{\mathbf{c}\in\Gamma(k)} d_i(\mathbf{c})\hat{p}_{JC}(\mathbf{c}).$$

The observed sum

$$O = \sum_{i=1}^{n} d_i(\mathbf{Y}_i)$$

is then compared to the estimated conditional expected value

$$\hat{E} = \sum_{i=1}^{n} \hat{D}_i.$$

Verification of regularity conditions is not trivial, but it is somewhat easier in the case in which $q/n^{1/2}$ approaches 0, a condition that appears appropriate for the Math and Verbal tests.

Application of adjusted generalized residuals is relatively straightforward, although some computation is required to obtain the estimated asymptotic variance $\hat{\sigma}(e)$ of $e = O - \hat{E}$. If $Y_{i+} = k$, then let

$$\hat{d}_i = \sum_{\mathbf{c} \in \Gamma(k)} d_i(\mathbf{c}) \hat{p}_{JC}(\mathbf{c}),$$

let

$$\hat{u}_i = \sum_{\mathbf{c} \in \Gamma(k)} [d_i(\mathbf{c}) - \hat{d}_i]^2 \hat{p}_{JC}(\mathbf{c}),$$

let $\hat{\mathbf{m}}_{Ci}$ be the q-dimensional vector with coordinates $m_{kj}(\hat{\boldsymbol{\beta}}_C)$ for j from 1 to q, and let

$$\hat{\mathbf{g}}_i = \sum_{\mathbf{c} \in \Gamma(k)} (\mathbf{c} - \hat{m}_{Ci}) d_i(c) \hat{p}_{JC}(\mathbf{c}).$$

Let

$$\hat{\mathbf{h}} = \sum_{i=1}^{n} \hat{\mathbf{g}}_i.$$

Then

$$\hat{\sigma}^2(e) = \sum_{i=1}^{n} \hat{u}_i - \hat{\mathbf{h}}^T \hat{\mathbf{V}}^- \hat{\mathbf{h}}.$$

The adjusted residual

$$z = e/\hat{\sigma}(e)$$

converges in distribution to a standard normal random variable if $\mathbf{h}^T(\mathbf{V}^*)^-\mathbf{h}/\sigma(e)$ is bounded above and if

$$[\sigma(e)]^{-1} \max_{1 \le i \le n} \max_{0 \le k \le q} \max_{\mathbf{c} \in \Gamma(k)} |d_i(c)|$$

approaches 0, where $\sigma^2(e)$ is defined by the equations

$$\bar{d}_{ik} = \sum_{\mathbf{c} \in \Gamma(k)} d_i(\mathbf{c}) p_{JC}(\mathbf{c}),$$

$$u_i = \sum_{k=0}^{q} \sum_{\mathbf{c} \in \Gamma(k)} [d_i(\mathbf{c}) - \bar{d}_{ik}]^2 p_J(\mathbf{c}),$$

$\mathbf{m}_{kC}$ is the q-dimensional vector with coordinates $m_{kjC} = m_{kj}(\boldsymbol{\beta})$ for j from 1 to q,

$$\mathbf{g}_i = \sum_{k=0}^{q} \sum_{\mathbf{c}\in\Gamma(k)} (\mathbf{c} - \mathbf{m}_{kC}) d_i(c) p_{JC}(\mathbf{c}),$$

$$\mathbf{h} = \sum_{i=1}^{n} \mathbf{g}_i,$$

and

$$\sigma^2(e) = \sum_{i=1}^{n} u_i - \mathbf{h}^T (\mathbf{V}^*)^- \mathbf{h}.$$

A good example of an adjusted residual to consider is the sum

$$O_j = \sum_{i=1}^{n} Y_{ij} Y_{i+} = \sum_{k=0}^{q} k f_{kj}.$$

The corresponding estimated expected value is

$$\hat{E}_j = \sum_{k=0}^{q} k n_k \hat{m}_{kjC},$$

and the unadjusted residual is

$$e_j = O_j - \hat{E}_j.$$

In effect, this sum leads to examination of the difference between the estimates of the point-biserial correlation of Y_{1j} and Y_{1+} derived with and without the Rasch model. Let

$$Y_{\cdot+} = n^{-1} \sum_{i=1}^{n} Y_{i+} = n^{-1} \sum_{k=0}^{q} k n_k,$$

and let

$$U = \sum_{i=1}^{n} (Y_{i+} - Y_{\cdot+})^2 = \sum_{k=0}^{q} n_k (k - Y_{\cdot+})^2.$$

The standard estimate of the point-biserial correlation is

$$\frac{O_j - Y_{+j} Y_{\cdot+}}{[Y_{+j}(1 - Y_{+j}/n)U]^{1/2}}.$$

Under the model, the estimated point-biserial correlation is

$$\frac{\hat{E}_j - Y_{+j} Y_{\cdot+}}{[Y_{+j}(1 - Y_{+j}/n)U]^{1/2}}.$$

For any constant c,

$$O_j - \hat{E}_j = \sum_{k=0}^{q}(k-c)f_{kj} - \sum_{k=0}^{q}(k-c)n_k\hat{m}_{kjC},$$

Let $\hat{v}_{kjj'}$ be $v_{kjj'}(\bar{\boldsymbol{\beta}}_C)$. It is a straightforward matter to verify that

$$\hat{\sigma}^2(e_j) = \sum_{k=0}^{q} n_k(k - Y_{\cdot+})^2\hat{v}_{kjj} - \hat{\mathbf{h}}_j^T\hat{\mathbf{V}}^-\hat{\mathbf{h}}_j,$$

where coordinate j' of $\hat{\mathbf{h}}_j$ is

$$\sum_{k=0}^{q} n_k(k - Y_{\cdot+})v_{kjj'}.$$

The adjusted residual for item j is then

$$z_j = e_j/\hat{\sigma}(e_j).$$

Both application to the Math test and application to the Verbal test provide overwhelming evidence that the Rasch model cannot hold. In the Math test, the largest adjusted residual in magnitude is $z_{57} = -158.08$, and only 8 items are associated with an adjusted residual less than 10 in absolute value. In the Verbal test, z_{57} is -144.54, and only 10 adjusted residuals z_j have magnitude less than 10.

The very large adjusted residuals are associated with substantial differences between observed and fitted point-biserial correlations. In the Math test, the observed point-biserial correlation for item 57 is 0.319. The fitted value is 0.482. In the Verbal test, the observed point-biserial correlation for item 57 is 0.234, while the fitted value is 0.401.

2.4 Model Error and Log Penalties

As with use of JMLE, modifications in results for CMLE must be made if the Rasch model does not hold. Arguments are quite similar in nature to those for JMLE, so details are omitted. As in Section 1.7, let the $\mathbf{Y}_i$ be independent and identically distributed, and let the joint probability $p_J(\mathbf{c})$ that $\mathbf{Y}_i = \mathbf{c}$ be positive for each q-dimensional vector $\mathbf{c}$ with coordinates c_j equal to 0 or 1. The conditional probability that $\mathbf{Y}_1 = \mathbf{c}$ in $\Gamma(k)$ given that $\mathbf{Y}_1 = k$ is then $p_{JC}(\mathbf{c}) = p_J(\mathbf{c})/P_k$. A unique q-dimensional vector $\boldsymbol{\beta}$ with coordinates β_j

exists such that $\beta_1 = 0$ and

$$\sum_{k=0}^{q} P_k m_{kj}(\hat{\boldsymbol{\beta}}_C) = p_j^Y .$$

This definition of $\boldsymbol{\beta}$ is consistent with the previous definition of $\boldsymbol{\beta}$ if the Rasch model holds. Let

$$p_{JCR}(\mathbf{c}) = \exp(-\boldsymbol{\beta}^T \mathbf{c})/s_k(\mathbf{c}),\ \mathbf{c} \in \Gamma(k).$$

Then the expected log penalty per response

$$H_{CR} = -q^{-1} \sum_{k=0}^{q} P_k \sum_{\mathbf{c}\in\Gamma(k)} p_{JC}(\mathbf{c}) \log p_{JCR}(\mathbf{c})$$

does not exceed the conditional expected log penalty per response

$$-q^{-1} \sum_{k=0}^{q} P_k \sum_{\mathbf{c}\in\Gamma(k)} p_{JC}(\mathbf{c}) \log p'_{JC}(\mathbf{c})$$

if

$$p'_{JC}(\mathbf{c}) = \exp(-\boldsymbol{\gamma}^T \mathbf{c})/s_k(\boldsymbol{\gamma}),\ \mathbf{c} \in \Gamma(k),$$

for a q-dimensional vector $\boldsymbol{\gamma}$ with $\gamma_1 = 0$. The minimum expected log penalty per response H_{CR} is at least as great as the conditional entropy H_C of $\mathbf{Y}_1$ given Y_{1+}, with equality if, and only if, $p_{JC}(\mathbf{c}) = p_{JCR}(\mathbf{c})$. This equality surely holds if the Rasch model holds. Customary arguments for log-linear models show that

$$p_{JR}(\mathbf{c}) = P_k p_{JCR}(\mathbf{c})$$

satisfies the condition that the expected log penalty per response

$$H_{UR} = H_{CR} + H_+ = -q^{-1} \sum_{k=0}^{q} \sum_{\mathbf{c}\in\Gamma(k)} p_J(\mathbf{c}) \log p_{JR}(\mathbf{c})$$

does not exceed

$$-q^{-1} \sum_{k=0}^{q} \sum_{\mathbf{c}\in\Gamma(k)} p_J(\mathbf{c}) \log p'_J(\mathbf{c})$$

for

$$p'_J(\mathbf{c}) = \exp(\alpha_k - \boldsymbol{\gamma}^T \mathbf{c}),\ \mathbf{c} \in \Gamma(k),$$

for a q-dimensional vector $\boldsymbol{\gamma}$ with $\gamma_1 = 0$ and some real α_k (Haberman, 1974, chap. 2) for which

$$\sum_{k=0}^{q} \sum_{\mathbf{c}\in\Gamma(k)} p'_J(\mathbf{c}) = 1.$$

Clearly $H_{UR} \geq H_U$. The condition $H_{UR} = H_U$ holds if the Rasch model holds.

If the number q of items is fixed, then it is a straightforward matter to verify that $\hat{\boldsymbol{\beta}}_C$ converges almost surely to $\boldsymbol{\beta}$ and that $n^{1/2}(\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta})$ converges in distribution to a multivariate normal random variable with zero mean and with covariance matrix

$$\mathbf{K}\mathbf{V}^{*-}\mathbf{V}'\mathbf{V}^{*-}\mathbf{K}^T.$$

The matrix $\mathbf{V}'$ is the expected conditional covariance matrix of $\mathbf{Y}_1$ given Y_{1+}. This formula is consistent with the formula obtained if the Rasch model holds, for $\mathbf{V}'$ is then $\mathbf{V}^*$.

If the number q of items increases, then, as in the case of JMLE, wording of results must be modified due to dependence of β_j on q. It is simplest to assume that the β_j are uniformly bounded as q increases, as is the case if the Rasch model is valid. Assume that positive constants α_1 and α_2 exist such that

$$\alpha_1\mathbf{x}^T\mathbf{x} \leq \mathbf{x}^T\mathbf{V}'\mathbf{x} \leq \alpha_2\mathbf{x}^T\mathbf{x}$$

for any $q \geq 1$ and any q-dimensional vector $\mathbf{x}$ such that $\sum_{j=1}^{q} x_j = 0$. Then $|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|$ converges in probability to 0. If $\sigma_e^2(\hat{\beta}_{jC})$ is row j and column j of of the matrix $\mathbf{K}\mathbf{V}^{*-}\mathbf{V}'\mathbf{V}^{*-}\mathbf{K}^T$, then the standardized value $n^{1/2}(\hat{\beta}_j - \beta'_{jC})/\sigma_e(\hat{\beta}_{jC})$ converges in distribution to a standard normal deviate. It should be noted that results for an invalid model do approach those for a valid model as the m_{kjC} approach $m_{kj}(\boldsymbol{\beta})$ for some $\boldsymbol{\beta}$ such that $\beta_1 = 0$.

The minimum conditional expected penalty H_{CR} may be estimated by $\hat{H}_{CR}$. For q fixed, $\hat{H}_{CR}$ converges to H_{CR} with probability 1, and $\hat{H}_{UR}$ converges to H_{UR} with probability 1. For q increasing, $\hat{H}_{CR} - H_{CR}$ and $\hat{H}_{UR} - H_{UR}$ converge in probability to 0. Normal approximations are readily available, at least if q/n approaches 0. Let $\sigma(\hat{H}_{UR})$ be the standard deviation of $q^{-1}\log p_{JR}(\mathbf{Y}_1)$, and let $\sigma(\hat{H}_{CR})$ be the standard deviation of $q^{-1}\log p_{JCR}(\mathbf{Y}_1)$. For a fixed number q of items, $n^{1/2}(\hat{H}_{UR} - H_U)$ converges in distribution to a normal random variable with mean 0 and variance $\sigma^2(\hat{H}_{UR})$, and $n^{1/2}(\hat{H}_{CR} - H_C)$ converges in distribution to a normal random variable with mean 0 and variance $\sigma^2(\hat{H}_{CR})$.

For q increasing, results are not quite as convenient as in the case in which the Rasch model holds. Provided that $\sigma(\hat{H}_{CR})$ and $\sigma(\hat{H}_{UR})$ are bounded below by a positive constant, $n^{1/2}(\hat{H}_{UR} - H_U)/\sigma(\hat{H}_{UR})$ and $n^{1/2}(\hat{H}_{CR} - H_C)/\sigma(\hat{H}_{CR})$ both converge in distribution to a standard normal random variable. These results are readily applied to construction of approximate confidence intervals for H_C and H_U (Gilula & Haberman, 1995). Indeed, the same approach to approximate confidence intervals that applies if the Rasch model holds continues to apply even if the model is not valid.

For a rather small number q of items, estimation of the entropies H_C and H_U may be accomplished without any assumptions concerning validity of the Rasch model. One may estimate H_C by

$$\hat{H}_C = -(nq)^{-1} \sum_{k=1}^{q-1} \sum_{\mathbf{c} \in \Gamma(k)} N(\mathbf{c}) \log[N(\mathbf{c})/n_k]$$

and H_U by

$$\hat{H}_U = -(nq)^{-1} \sum_{k=1}^{q-1} \sum_{\mathbf{c} \in \Gamma(k)} N(\mathbf{c}) \log[N(\mathbf{c})/n].$$

For fixed q, $\hat{H}_C$ converges almost surely to H_C, and $\hat{H}_U$ converges almost surely to H_U. Thus a comparison with the estimates $\hat{H}_{CR}$ and $\hat{H}_{UR}$ indicates the loss of predictive power due to use of the Rasch model. Approximate confidence intervals for the difference $H_{CR} - H_C = H_{UR} - H_U$ are readily available (Gilula & Haberman, 1994; Gilula & Haberman, 1995). In addition, formal chi-square tests for validity of the Rasch model are available in this case. The standard likelihood-ratio chi-square statistic

$$L^2 = 2nq(\hat{H}_U - \hat{H}_{UR}) = 2nq(\hat{H}_{CR} - \hat{H}_C)$$

converges in distribution to a chi-square random variable with $2^q - 2q$ degrees of freedom. If $H_U > H_{UR}$, so that the Rasch model fails, then L^2/n converges almost surely to $2(H_U - H_{UR}) > 0$.

Application of the chi-square test is limited to relatively small subtests of the Math and Verbal tests. Presumably $np_J(\mathbf{c})$ should be at least 1 for $\mathbf{c}$ in $\Gamma(k)$ for $1 \leq k \leq q-1$ and 80 percent of such $np_J(\mathbf{c})$ should be at least 5 (Cochran, 1954). In practice, the sample size n should be relatively large compared to 2^q. Nonetheless, it should be noted that the Rasch

model for a complete Math or Verbal test can only hold if it holds for a subtest consisting of a selected group of items. For the first five questions in the Verbal test, $\hat{H}_{UR}$ is 0.411 and $\hat{H}_U$ is 0.410, so that the actual information loss appears small. Nonetheless, overwhelming evidence exists that the Rasch model cannot hold, for $L^2 = 4,175.5$, and there are only $2^5 - 2(5) = 22$ degrees of freedom. This pattern is repeated for other selections of items. For the first five Math items, $\hat{H}_{UR} = 0.344$, $\hat{H}_U = 0.343$, and $L^2 = 4,580.5$. For the last five Verbal items, $\hat{H}_{UR} = 0.597$, $\hat{H}_U = 0.593$, and $L^2 = 12,257.7$. For the last five Math items, $\hat{H}_{UR} = 0.604$, $\hat{H}_U = 0.602$, and $L^2 = 7,239.4$.

An alternative approach to studying lack of fit can be based on a simple variant of the Rasch model in which it is assumed that, for some q-dimensional vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ with initial coordinates 0,

$$p_{JC}(\mathbf{c}) = \frac{\exp\{-[\boldsymbol{\beta} + (k - q/2)\boldsymbol{\gamma}]^T \mathbf{c}\}}{s_k(\boldsymbol{\beta} + q^{-1}(k - q/2)\boldsymbol{\gamma})}$$

for $\mathbf{c}$ in $\Gamma(k)$. This variant includes the Rasch model as a special case, as is evident by consideration of $\boldsymbol{\gamma}$ equal to $\mathbf{0}$. This model may be analyzed through the same conditional arguments used in the Rasch model. The resulting estimates $\hat{\boldsymbol{\beta}}_{CL}$, $\hat{\boldsymbol{\gamma}}_{CL}$, and

$$\hat{m}_{kjCL} = m_{kj}(\hat{\boldsymbol{\beta}} + q^{-1}(k - q/2)\hat{\boldsymbol{\gamma}})$$

satisfy the constraints

$$\sum_{k=0}^{q} n_k \hat{m}_{kjCL} = Y_{+j},$$

$$\sum_{k=0}^{q} (k - q/2) n_k \hat{m}_{kjCL} = \sum_{k=0}^{q} (k - q/2) f_{kj},$$

and

$$\sum_{j=1}^{q} \hat{m}_{kjCL} = k.$$

A notable feature for the new model is that fitted and observed point-biserial correlations of Y_{ij} and Y_{i+} are the same. Given that the ability variable θ_1 has positive variance, consistency and asymptotic normality results are readily established both for the case in which the model holds and for the case in which the model fails, and the Newton-Raphson algorithm remains applicable. Under the new model, the minimum expected logarithmic

penalty per observation for prediction of $\mathbf{Y}_1$ by Y_{1+} is

$$H_{CL} = -q^{-1} \sum_{k=0}^{q} P_k \sum_{\mathbf{c} \in \Gamma(k)} p_{JC}(\mathbf{c}) \log p_{JCL}(\mathbf{c}),$$

where $p_{JCL}(\mathbf{c})$ is defined so that

$$p_{JCL} = \frac{\exp\{-[\boldsymbol{\beta}_L + q^{-1}(k - q/2)\boldsymbol{\gamma}_L]^T \mathbf{c}\}}{s_k(\boldsymbol{\beta}_L + q^{-1}(k - q/2)\boldsymbol{\gamma}_L)},$$

$$m_{kjCL} = m_{kj}(\boldsymbol{\beta}_L + (k - q/2)\boldsymbol{\gamma}_L),$$

$$\sum_{k=0}^{q} P_k m_{kjCL} = p_j^Y,$$

$$\sum_{k=0}^{q} (k - q/2) P_k m_{kjCL} = \sum_{k=0}^{q} (k - q/2) P_k m_{kjC},$$

and

$$\sum_{j=1}^{q} m_{kjCL} = k.$$

The corresponding minimum expected log penalty for prediction of $\mathbf{Y}_1$ is

$$H_{UL} = H_{CL} + H_+.$$

If

$$\hat{p}_{JCL}(\mathbf{c}) = \frac{\exp\{-[\hat{\boldsymbol{\beta}}_{CL} + q^{-1}(k - q/2)\hat{\boldsymbol{\gamma}}_{CL}]^T \mathbf{c}\}}{s_k(\hat{\boldsymbol{\beta}} + q^{-1}(k - q/2)\hat{\boldsymbol{\gamma}})}$$

for $\mathbf{c}$ in $\Gamma(k)$, then one obtains the estimated expected log penalties

$$\hat{H}_{CL} = -(nq)^{-1} \sum_{i=1}^{n} \log \hat{p}_{JCL}(\mathbf{Y}_i)$$

and

$$\hat{H}_{UL} = \hat{H}_{CL} + \hat{H}_+.$$

For the Verbal test, $\hat{H}_{UL} = 0.52098$ is rather close to $\hat{H}_{UR} = 0.52630$. For the Math test, $\hat{H}_{UL} = 0.47500$ is rather close to $\hat{H}_{UR} = 0.48040$. Despite the closeness, there is overwhelming evidence based on these comparisons that the Rasch model cannot hold.

To verify this claim, consider the likelihood-ratio chi-square statistic

$$L_1^2 = 2nq(\hat{H}_{UR} - \hat{H}_{UL}) = 2nq(\hat{H}_{CR} - \hat{H}_{CL}).$$

If the Rasch model holds and q is fixed, then a straightforward application of general results for log-linear models permits a demonstration that L_1^2 converges in distribution to a chi-square random variable with $q-1$ degrees of freedom (Haberman, 1974, chap. 4). A more complicated case has q increasing but q^2/n approaching 0. In this case, one may show that $(L_1^2 - q + 1)/[2(q-1)]^{1/2}$ converges in distribution to a standard normal random variable (Haberman, 1977b; Haberman, 1977a; Portney, 1988). The normal approximation supports use of the chi-square approximation, for a chi-square random variable χ_ν^2 with ν degrees of freedom satisfies the condition that $(\chi_\nu^2 - \nu)/(2\nu)^{1/2}$ converges in distribution to a standard normal random variable as ν approaches ∞.

For the case under study, the Math test yields $L_1^2 = 289,401$ on 59 degrees of freedom, and the Verbal test yields L_1^2 of 370,648 on 77 degrees of freedom, so that the test statistics provide overwhelming evidence that the Rasch model does not hold. Note that it has been shown that the Rasch model is not valid for either test; however, no demonstration has been made that the model error is very large in terms of prediction of the response vector $\mathbf{Y}_1$.

Other comparable likelihood-ratio chi-square tests can and have been constructed (Andersen, 1973b). The particular test chosen has been emphasized because the number of degrees of freedom is relatively small and regularity conditions appear reasonable for the sample sizes and numbers of items under consideration.

3 Latent Structures and Log-linear Models

There is a subtle difficulty encountered in Section 2 in distinguishing between a Rasch model and a log-linear model. For a Rasch model,

$$p_J(\mathbf{c}) = P_k \exp(-\boldsymbol{\beta}'\mathbf{c})/s_k(\boldsymbol{\beta}) \tag{18}$$

for some q-dimensional vector $\boldsymbol{\beta}$ with $\beta_1 = 0$ and some distribution function D for which

$$P_k = s_k(\boldsymbol{\beta}) \int_{-\infty}^{\infty} e^{k\theta} \Psi(\boldsymbol{\beta}, \theta) dD(\theta). \tag{19}$$

In the corresponding log-linear model,

$$p_J(\mathbf{c}) = \exp(\alpha_k - \boldsymbol{\beta}^T \mathbf{c}),\ \mathbf{c} \in \Gamma(k),$$

for a q-dimensional vector $\boldsymbol{\beta}$ with $\beta_1 = 0$ and some real α_k for which

$$\sum_{k=0}^{q} \exp(\alpha_k) s_k(\boldsymbol{\beta}) = \sum_{k=0}^{q} \sum_{\mathbf{c} \in \Gamma(k)} p_J(\mathbf{c}) = 1 \tag{20}$$

(Tjur, 1982). If (18) holds, then the log-linear model holds with

$$\alpha_k = \log[P_k / s_k(\boldsymbol{\beta})] = \log \int_{-\infty}^{\infty} e^{k\theta} \Psi(\boldsymbol{\beta}, \theta) dD(\theta). \tag{21}$$

Thus the Rasch model implies the log-linear model. If Z is a positive random variable such that the distribution function of $\log Z$ has value

$$\frac{\int_{-\infty}^{x} \Psi(\boldsymbol{\beta}, \theta) dD(\theta)}{\int_{-\infty}^{\infty} \Psi(\boldsymbol{\theta}, \theta) dD(\theta)}$$

for x in R, then $\exp(\alpha_k - \alpha_0) = E(Z^k)$.

If the log-linear model holds and a positive random variable Z exists such that $E(Z^k) = \exp(\alpha_k - \alpha_0)$ for $0 \leq k \leq q$, then the Rasch model holds (Cressie & Holland, 1983), for one may let Δ be the distribution function of $\log Z$ and let D be the distribution function such that

$$D(x) = \frac{\int_{-\infty}^{x} [\Psi(\boldsymbol{\beta}, \theta)]^{-1} d\Delta(\theta)}{\int_{-\infty}^{\infty} [\Psi(\boldsymbol{\beta}, \theta)]^{-1} d\Delta(\theta)}$$

for real x. In this case, the equation

$$[\Psi(\boldsymbol{\beta}, \theta)]^{-1} = \sum_{k=0}^{q} e^{k\theta} s_k(\boldsymbol{\beta})$$

and (20) imply that

$$\exp(-\alpha_0) = \int_{-\infty}^{\infty} [\Psi(\boldsymbol{\beta}, \theta)]^{-1} d\Delta(\theta).$$

It then follows that (21) holds, so that (19) holds.

Classical results concerning existence of moments of positive random variables may be used to indicate whether a particular choice of α_k, $0 \leq k \leq q$, corresponds to a suitable positive random variable Z such that $\exp(\alpha_k - \alpha_0) = E(Z^k)$ for $1 \leq k \leq q$. It essentially suffices to consider whether two matrices are positive definite or nonnegative definite. If q is even, let $s = q/2$ and $r = s + 1$. If q is odd, let $r = s = (q + 1)/2$. Let the log-linear model hold. Let $\mathbf{A}$ be the r by r matrix with row i and column j equal to $\exp(\alpha_{i+j-2} - \alpha_0)$,

and let $\mathbf{B}$ be the s by s matrix with row i and column j equal to $\exp(\alpha_{i+j-1} - \alpha_0)$. The Rasch model can only hold if $\mathbf{A}$ and $\mathbf{B}$ are nonnegative definite, and the Rasch model can only hold for a continuous distribution function D of θ_1 if $\mathbf{A}$ and $\mathbf{B}$ are positive definite. On the other hand, if $\mathbf{A}$ and $\mathbf{B}$ are positive definite, then the Rasch model holds for some distribution function D (Cressie & Holland, 1983). In addition, the observed P_k and $\boldsymbol{\beta}$ are consistent with a distribution function D corresponding to a random variable with mass confined to no more than s points (Karlin & Studden, 1966, pp. 44, 173). The observed P_k are also consistent with other distribution functions D (Lindsay et al., 1991).

These results lead to a relatively simple relationship between unconstrained maximum-likelihood estimates and conditional maximum-likelihood estimates. If the conditional maximum-likelihood estimate $\hat{\boldsymbol{\beta}}_C$ exists, if each n_k is positive, if $\hat{P}_k = n_k/n$, if

$$\hat{\alpha}_k = \log[\hat{P}_k / s_k(\hat{\beta}_C)],$$

and if

$$\exp(\hat{\alpha}_k - \hat{\alpha}_0) = E(Z_k), 0 \leq k \leq q, \tag{22}$$

for some positive random variable Z, then $\ell_M = \ell_{CM} + \ell_{SM}$, and $\hat{\boldsymbol{\beta}}_C$ is the unique marginal maximum-likelihood estimate $\hat{\boldsymbol{\beta}}_J$ of $\boldsymbol{\beta}$. If $\hat{D}_J$ is a distribution function such that

$$\hat{P}_k = s_k(\hat{\boldsymbol{\beta}}_C) \int_{-\infty}^{\infty} e^{k\theta} \Psi(\hat{\boldsymbol{\beta}}_C, \theta) d\hat{D}(\theta),$$

then $\hat{D}_J$ is a marginal likelihood estimate of D. The estimate $\hat{D}_J$ is not uniquely defined. If no positive random variable Z exists such that (22) holds, then $\hat{\beta}_C$ need not be an unconstrained marginal maximum-likelihood estimate of $\boldsymbol{\beta}$.

As a practical matter, the difference between a conditional maximum-likelihood estimate of $\boldsymbol{\beta}$ and an unconstrained marginal maximum-likelihood estimate of $\boldsymbol{\beta}$ appears to be small. If the Rasch model holds, the number q of items is fixed, and $\mathbf{A}$ and $\mathbf{B}$ are positive definite, then standard continuity properties of eigenvalues imply that the probability approaches 1 that $\hat{\boldsymbol{\beta}}_C$ is the unique unconstrained marginal maximum-likelihood estimate of $\boldsymbol{\beta}$ (Wilkinson, 1965, chap. 2). This situation can also apply if $n/2^q$ becomes large; however, this condition is clearly not relevant for the SAT I data under study. More typically, if q is large, then some of the products nP_k are quite small, even if n is very

large. Thus it is possible that the probability does not approach 1 that $\hat{\boldsymbol{\beta}}_C$ is the unique unconstrained marginal maximum-likelihood estimate of $\boldsymbol{\beta}$. This result does not prevent effective estimation of $\boldsymbol{\beta}$ or D. The large-sample properties of the $\hat{\boldsymbol{\beta}}_C$ are basically the same properties obtained if maximum-likelihood is applied in the case in which the θ_i are known. It remains the case that the empirical distribution function $\hat{D}$ does approximate the distribution function D.

There does exist a possibility that the log-linear model holds but the Rasch model does not, and tests of goodness of fit, which are really based on the log-linear model, will not detect this situation.

4 Conclusions

The results derived in the preceding sections suggest that CMLE provides an effective approach for analysis of the Rasch model for dichotomous items even in cases in which both the sample size and the number of items are large. An efficient approach for computation of conditional maximum-likelihood estimates has been derived that is considerably faster than previous algorithms. Standard large-sample approximations for the distributions of conditional maximum-likelihood estimates have been shown to apply, so that asymptotic confidence intervals are available. In addition, methods for residual analysis have been developed based on CMLE, and formal tests of goodness of fit have been produced that have effectively demonstrated lack of fit for the SAT I data under study.

Efforts have also been made to apply JMLE under realistic conditions. Results have been somewhat less satisfactory. Significant problems of asymptotic bias of joint maximum-likelihood estimates were detected for the SAT I data, generalized residual analysis was much more limited, goodness of fit was much less readily tested, and the analyses based on expected logarithmic penalty were more affected by bias problems. Nonetheless, JMLE was found to provide an effective approach for calculation of starting values for CMLE, and it was adequate for study of individual deviations from the Rasch models.

The analysis has also considered evaluation of the value of the Rasch model in terms of effective prediction of the pattern of item responses. This analysis has been based on

the criterion of expected logarithmic penalty. It has suggested but not conclusively proven that, even though the Rasch model is not valid for the SAT I data under study, the error of the Rasch model is relatively small. Further work on alternative models is needed to consider this issue more thoroughly; however, the basic issue is that small errors in models are readily detected if sample sizes are sufficiently large.

The methods of analysis developed in this report can be applied more generally. The most immediate generalizations involve the Rasch model for polytomous responses. The JMLE approach has application to 2PL and 3PL models; although presumably the limitations observed for the Rasch model will be even more severe in these cases.

The alternative to the Rasch model used in testing goodness of fit is not a latent-structure model; however, it is easily implemented and has potential use by itself. Of particular interest is the question of how this model compares to common psychometric models in terms of predictive power. Such comparisons will require analysis of marginal models for the case of large sample sizes and large numbers of items. Obviously much work remains.

References

Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B*, *34*, 42–54.

Andersen, E. B. (1973a). *Conditional inference and models for measuring*. Copenhagen: Mentalhygiejnisk Forskningsinstitut.

Andersen, E. B. (1973b). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, *10*, 417–451.

Cramér, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

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

Esseen, C.-G. (1945). Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law. *Acta Mathematica*, *77*, 1–125.

Fischer, G. H. (1981). On the existence and uniqueness of maximum likelihood estiamtes in the Rasch model. *Psychometrika*, *46*, 59–77.

Gilula, Z., & Haberman, S. J. (1994). Models for analyzing categorical panel data. *Journal of the American Statistical Association*, *89*, 645–656.

Gilula, Z., & Haberman, S. J. (1995). Prediction functions for categorical panel data. *The Annals of Statistics*, *23*, 1130–1142.

Gini, C. (1912). *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statische*. Bologna, Italy: Cuppini.

Gustafsson, J. E. (1980). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement*, *40*, 377–385.

Haberman, S. J. (1974). *The analysis of frequency data*. Chicago: University of Chicago Press.

Haberman, S. J. (1977a). Log-linear models and frequency tables with small expected cell counts. *The Annals of Statistics*, *5*, 1148–1169.

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

Haberman, S. J. (1978). *Analysis of qualitative data* (Vol. 1). New York: Academic Press.

Haberman, S. J. (1989). Concavity and estimation. *The Annals of Statistics*, *17*, 1631–1661.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Karlin, S., & Studden, W. J. (1966). *Tchebycheff systems: With applications in analysis and statistics*. New York: Interscience.

Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, *27*, 887–906.

Lindsay, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation for the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, *86*, 96–107.

Liou, M. (1994). More on the computation of Higher-order derivatives of the elementary symmetric functions in the Rasch model. *Applied Psychological Measurement*, *18*, 53–62.

Loomis, L. A., & Sternberg, S. (1968). *Advanced calculus*. Reading, MA: Addison-Wesley.

Portney, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, *16*, 356–366.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York: John Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Stewart, G. W. (1973). *Introduction to matrix computations.* New York: Academic Press.

Tjur, T. (1982). A connection between Rasch's item analysis model and a multiplicative Poisson model. *Scandinavian Journal of Statistics*, *9*, 23–30.

Wilkinson, J. H. (1965). *The algebraic eigenvalue problem.* Oxford, UK: Clarendon Press.

Appendix
Proofs of Results

Proof that $v_{jj'k}(\boldsymbol{\beta}) < 0$ ***if*** $j \neq j'$ ***and*** $0 < k < q$**.**

For any random variables A and B with values 0 and 1, if $P(A = a, B = b) = \tau_{ab}$, then the covariance of A and B is

$$\begin{aligned} & \tau_{11}(\tau_{11} + \tau_{10} + \tau_{01} + \tau_{00}) - (\tau_{11} + \tau_{10})(\tau_{11} + \tau_{01}) \\ = \;& \tau_{11}\tau_{00} - \tau_{01}\tau_{10}. \end{aligned}$$

For $k = 1$, it is trivially true that $v_{jj'k}(\boldsymbol{\beta}) < 0$ for $j \neq j'$, for $Y_{ij'} = Y_{ij} = 1$ cannot hold if $Y_{i+} = 1$. For $k > 1$, consider j and j' such that $1 \leq j < j' \leq q$. Then the probability that $Y_{ij} = a$ and $Y_{ij'} = b$ given that $Y_{i+} = k$ is

$$h_{abjj'k}(\boldsymbol{\beta}) = \frac{\sum_{\mathbf{c} \in \Gamma'(a,b,j,j',k)} \exp(-\boldsymbol{\beta}'\mathbf{c})}{\sum_{\mathbf{c} \in \Gamma(k)} \exp(-\boldsymbol{\beta}'\mathbf{c})},$$

where $\Gamma'(a, b, j, j', k)$ consists of $\mathbf{c}$ in $\Gamma(k)$ such that $c_j = a$ and $c_{j'} = b$. Thus the conditional covariance of Y_{ij} and $Y_{ij'}$ given $Y_{i+} = k$, $2 \leq k < q$ is $h_{11jj'k}h_{00jj'k} - h_{01jj'k}h_{10jj'k}$. Let $G(j, j', k)$ be the population of q-dimensional vectors $\mathbf{d}$ with nonnegative integer coordinates $d_h \leq 2$ such that $\sum_{h=1}^{q} d_h = 2k$ and $d_j = d_{j'} = 1$. If $\mathbf{c}$ is in $\Gamma'(1, 1, j, j', k)$ and $\mathbf{e}$ is in $\Gamma'(0, 0, j, j', k)$, then $\mathbf{c} + \mathbf{e}$ is in $G(j, j', k)$. Similarly, if $\mathbf{c}$ is in $\Gamma'(1, 0, j, j', k)$ and $\mathbf{e}$ is in $\Gamma'(0, 1, j, j', k)$, then $\mathbf{c} + \mathbf{e}$ is in $G(j, j', k)$. For each $\mathbf{d}$ in $G(j, j', k)$, let $\upsilon(\mathbf{d})$ be half the number of coordinates $d_h = 1$ for $1 \leq h \leq q$, $h \neq j$, and $h \neq j'$. Note that $\upsilon(\mathbf{d})$ must be an integer. To any $\mathbf{d}$ in $G(j, j', k)$ correspond

$$\xi_1(\mathbf{d}) = \frac{[2\upsilon(\mathbf{d})]!}{\{[\upsilon(\mathbf{d})]!\}^2}$$

pairs of $\mathbf{c}$ in $\Gamma'(1, 0, j, j', k)$ and $\mathbf{e}$ in $\Gamma'(0, 1, j, j', k)$ such that $\mathbf{c} + \mathbf{e}$ is in $G(j, j', k)$ and

$$\xi_2(\mathbf{d}) = \frac{[2\upsilon(\mathbf{d})]!}{[\upsilon(\mathbf{d}) + 1]![\upsilon(\mathbf{d}) - 1]!}$$

pairs of $\mathbf{c}$ in $\Gamma'(1, 1, j, j', k)$ and $\mathbf{e}$ in $\Gamma'(0, 0, j, j', k)$ such that $\mathbf{c} + \mathbf{e}$ is in $G(j, j', k)$. Thus

$$v_{jj'k}(\boldsymbol{\beta}) = \frac{\sum_{\mathbf{d} \in G(j,j',k)} [\xi_2(\mathbf{d}) - \xi_1(\mathbf{d})] \exp(\boldsymbol{\beta}'\mathbf{d})}{\left[\sum_{\mathbf{c} \in \Gamma(k)} \exp(\boldsymbol{\beta}'\mathbf{c})\right]^2}.$$

The ratio

$$\frac{\xi_2(\mathbf{d})}{\xi_1(\mathbf{d})} = \frac{\upsilon(\mathbf{c})}{\upsilon(\mathbf{c})+1} < 1,$$

so that $v_{jj'k}(\boldsymbol{\beta}) < 0$.