PSYCHOMETRIKA—VOL. 61, NO. 3, 509–528
SEPTEMBER 1996

ANALYSIS OF RESIDUALS FOR THE MULTINOMIAL ITEM RESPONSE MODEL

MARK REISER

ARIZONA STATE UNIVERSITY

Using the item response model as developed on the multinomial distribution, asymptotic variances are obtained for residuals associated with response patterns and first-, and second-order marginal frequencies of manifest variables. When the model does not fit well, an examination of these residuals may reveal the source of the poor fit. Finally, a limited-information test of fit for the model is developed by using residuals defined for the first-, and second-order marginals. Model evaluation based on residuals for these marginals is particularly useful when the response pattern frequencies are sparse.

Key words: asymptotic standard error, latent trait, limited information, factor analysis.

1. Introduction

In applications of the item response model, goodness of fit is typically assessed by using the the Pearson and likelihood ratio statistics, which have approximate chi-square distributions in large samples. This paper presents associated methods for the analysis of residuals in the two parameter item response model. A diagnostic developed here is an adjusted residual for a cell in the cross-classification of the responses to the manifest variables. For k manifest variables, the cross-classification has 2^k cells, and if the model fits poorly, the set of adjusted residuals can be used to attribute lack of fit to one or more of the associated response patterns.

When the number of manifest variables is large, say greater than seven or eight, and the sample size is small to moderate, other approaches are needed, because the table of frequencies for the cross-classification of item responses becomes so sparse that the chi-square distribution is not a valid asymptotic approximation for traditional goodness-of-fit statistics. For this situation, the examination of residuals associated with the first-, and second-order marginal frequencies is proposed for assessing how well the model fits each variable or pair of variables. The first-, and second-order marginals are also used to develop a limited-information test of fit for the model, and this test is compared to a similar one found in Christoffersson (1975) and Muthén (1978).

Throughout this paper, the focus is on residual analysis as it may be carried out in association with traditional goodness-of-fit tests. A general review of assessing the fit of item-response models is given in Hambleton, Swaminathan, and Rogers (1991).

Item Response Model

The two-parameter item response model is presented in this section. According to this model, the probability of the response to a manifest variable, sometimes also referred to as an item, can be given by a logistic item response function:

The author would like to thank Yasuo Amemiya and Joseph Lucke for helpful suggestions. This research was supported by a Research Incentive Grant from Arizona State University.

Requests for reprints should be sent to Mark Reiser, Department of Economics, Arizona State University, Tempe, Arizona 85287-3806.

0033-3123/96/0900-93103$00.75/0

$$\pi(Y_i = 1|\boldsymbol{\beta}_i', X = x) = (1 + \exp(-\beta_{i0} - \beta_{i1}x))^{-1}, \tag{1}$$

where Y_i represents the response to item i,

$\beta_{\mathrm{i}0}$ = intercept parameter for item i,
β_{i1} = slope parameter for item i,
$\boldsymbol{\beta}_i' = (\beta_{0\mathrm{i}}, \beta_{1\mathrm{i}})$, and
x = value taken on by latent random variable X.

Since

$$\pi(Y_i = 0|\boldsymbol{\beta}_i', X = x) = 1.0 - \pi(Y_i = 1|\boldsymbol{\beta}_i', X = x),$$

it follows that

$$\pi(Y_i = y_i|\boldsymbol{\beta}_i', x) = \pi(Y_i = 1|\boldsymbol{\beta}_i', x)^{y_i}[1.0 - \pi(Y_i = 1|\boldsymbol{\beta}_i', x)]^{1-y_i}.$$

It is assumed that, *conditional* upon the latent variable, responses to the manifest variables are independent. Let **Y** represent a random vector of responses to the items, with element Y_i, and let **y** represent a realized value of **Y**. Then

$$\pi(\mathbf{Y} = \mathbf{y}|\boldsymbol{\beta}, x) = \prod_{i=1}^{k} \pi(Y_i = 1|\boldsymbol{\beta}, x)^{y_i}[1 - \pi(Y_i = 1|\boldsymbol{\beta}, x)]^{1-y_i}, \tag{2}$$

$$\text{where } \boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \beta_{i1} \\ \beta_{02} & \beta_{12} \\ \beta_{03} & \beta_{13} \\ \vdots & \vdots \\ \beta_{0k} & \beta_{1k} \end{bmatrix}.$$

Finally, the probability of response pattern s, say, is obtained by taking the expected value of the conditional probability over the distribution of X in the population, and is sometimes called the marginal probability:

$$\pi_s(\boldsymbol{\beta}) = \pi(\mathbf{Y} = \mathbf{y}_s|\boldsymbol{\beta}) = \int_{-\infty}^{\infty} \pi(\mathbf{Y} = \mathbf{y}_s|\boldsymbol{\beta}, x)f(x)\,dx, \tag{3}$$

where $f(x)$ is the density function of X in the population of respondents. The notation $\pi_s(\boldsymbol{\beta})$ is introduced here for use later where a compact notation is required.

If **V** represents a T-dimensional multinomial random vector of frequencies associated with the response patterns, the distribution of **V** is given by

$$\pi(\mathbf{V} = \mathbf{n}) = n! \prod_{s=1}^{T} \frac{[\pi_s(\boldsymbol{\beta})]^{n_s}}{n_s!}, \tag{4}$$

where **n** = vector of observed frequencies,

n_s = element s of **n**, and
n = total sample size = $\Sigma_{s=1}^{T} n_s$.

Probit model is a term sometimes associated with the normal distribution. Given the normal distribution for $f(x)$, and the logit function in Expression (1), Bartholomew (1987) refers to Expression (3) as a logit/probit model. A necessary condition for the identification of the model is $k \geq 3$. The maximum likelihood solution for parameter estimates based on (4), but with a probit function in place of (1), was first given by Bock and Lieberman (1970). Bock and Aitken (1981) refer to estimation based on the likelihood function that follows from (4) as the marginal maximum likelihood technique. Holland (1990) and Cressie and Holland (1983) also use the marginal maximum likelihood method in item response models. The model essentially gives a full-information factor analysis of dichotomous variables; the presentation here, however, is confined to a single factor. See Takane and de Leeuw (1987) for a thorough treatment of the relationship between the item response model and dichotomous variable factor analysis.

2. Residuals

The adequacy of the model may be assessed with goodness-of-fit tests for the multinomial distribution that are based on traditional likelihood ratio and Pearson statistics. Lack of fit may be due to a misspecification in any aspect of the model given above. Two key features of the model are the item response curve given in (1) and conditional independence given in (2). The remainder of the paper is focused primarily on analysis of residuals for detecting lack of fit due to failure of conditional independence. While the logistic response curve has been used with enormous success in models for categorical variables, lack of fit may be due to an error in the specification of this curve. For inquiries where the choice of item response curve is an issue, the methodology given by Stout (1987) is recommended as a complement to the methodology developed here.

To form residuals for the cells of the multinomial vector of response patterns, there are several possible approaches, including examining $\hat{p}_s - \pi_s(\hat{\boldsymbol{\beta}})$ directly, where $\hat{p}_s = n_s/n$ is element s of $\hat{\mathbf{p}}$, the vector of sample proportions, $\hat{\boldsymbol{\beta}}$ is an estimator for the parameter matrix, and $\pi_s(\hat{\boldsymbol{\beta}})$ is the estimated expected proportion for cell s. For the multinomial model, it has been traditional to examine standardized residuals (Cochran, 1954). Let

$$r_s = \frac{\hat{p}_s - \pi_s(\hat{\boldsymbol{\beta}})}{(\pi_s(\hat{\boldsymbol{\beta}}))^{1/2}} \tag{5}$$

then $n^{1/2} r_s$ is the standardized residual. $n\ \Sigma_s\ r_s^2$ is equal to the Pearson chi-square goodness-of-fit statistic. Under some circumstances, the set of these residuals may be useful for finding cells that are not well fit by the model. However, since the distribution of $n^{1/2}r_s$ is not necessarily $N(0, 1)$, it is sometimes difficult to assess the significance of the magnitude of the standardized residual. Therefore, it is useful to divide the statistic by its standard error:

$$\frac{n^{1/2}r_s}{\hat{\sigma}_s},$$

yielding the adjusted residual (Haberman, 1973), which has an approximate $N(0, 1)$ distribution in large samples.

Consider the vector-valued function of $\mathbf{p}$ and $\boldsymbol{\beta}$:

$$\mathbf{h}(\mathbf{p}, \boldsymbol{\beta}) = \mathbf{D}(\boldsymbol{\pi}(\boldsymbol{\beta}))^{-1/2}(\mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\beta})), \tag{6}$$

where $\boldsymbol{\pi}(\boldsymbol{\beta})$ = vector of multinomial probabilities as a function of $\boldsymbol{\beta}$,
and $\mathbf{D}(\boldsymbol{\pi}(\boldsymbol{\beta}))$ = diagonal matrix with element (s, s) equal to $\pi_s(\boldsymbol{\beta})$.

The T dimensional vector of residuals, $\mathbf{r}$, is obtained from the function $\mathbf{h}(\mathbf{p}, \boldsymbol{\beta})$ when $\mathbf{p}$ takes on the value $\hat{\mathbf{p}}$ and $\boldsymbol{\beta}$ takes on the value $\hat{\boldsymbol{\beta}}$. For the general multinomial model, the asymptotic covariance matrix of $\mathbf{r}$ has been given by Haberman (1973) and Rao (1973).

Cochran (1955) also considered asymptotic variances for residuals. The following result applies this line of work on the multinomial model to the item response model.

Result 1. Assuming the regularity conditions given by Birch (1964a),

$$n^{1/2}\mathbf{r} \xrightarrow{L} N(\mathbf{0}, \boldsymbol{\Omega}_{\mathbf{r}}), \tag{7}$$

where $\boldsymbol{\Omega}_{\mathbf{r}} = \mathbf{I} - \boldsymbol{\pi}^{1/2}(\boldsymbol{\pi}')^{1/2} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$, $\xrightarrow{L}$ indicates convergence in Law

$\mathbf{A} = \mathbf{D}(\boldsymbol{\pi})^{-1/2}\, \partial \boldsymbol{\pi}(\boldsymbol{\beta})/\partial \text{ vec } \boldsymbol{\beta}$, evaluated at the true parameter values,

$\text{vec } \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$ with intercepts stacked on top of slopes, and

$\boldsymbol{\pi}^{1/2}$ = vector with elements given by square root of true proportions.

Result 1 follows directly from Haberman (1973) and Rao (1973).

Expression (7) applies to the item response model presented above, with $\boldsymbol{\pi}(\boldsymbol{\beta})$ given by expression (4) and an element of $\partial\boldsymbol{\pi}(\boldsymbol{\beta})/\partial\text{vec}\boldsymbol{\beta}$ given by

$$\frac{\partial \pi_s(\boldsymbol{\beta})}{\partial \beta_{hi}} = \int [y_{is} - (1 + \exp(-\beta_{0i} - \beta_{1i}x))^{-1}]\left(\prod_{j=1}^{k} \pi(Y_j = y_{js}|\boldsymbol{\beta}, x)\right) \frac{\partial(\beta_{0i} + \beta_{1i}x)}{\partial \beta_{hi}}\, dF(x),$$

where y_{is} is element i of $\mathbf{y}_s$. This integral must be evaluated by numerical quadrature.

Define $\boldsymbol{\Sigma}_{\mathbf{r}}$ to be the covariance matrix of $\mathbf{r}$, with estimator $\hat{\boldsymbol{\Sigma}}_{\mathbf{r}}$ defined by

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{r}} = n^{-1}\boldsymbol{\Omega}_{\mathbf{r}}\big|_{\boldsymbol{\pi} = \boldsymbol{\pi}(\hat{\boldsymbol{\beta}}), \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}}.$$

The square root of the diagonal elements from $\hat{\boldsymbol{\Sigma}}_{\mathbf{r}}$ can be used as estimated standard errors for calculating the adjusted residuals. $n^{1/2}r/\hat{\sigma}_r$ is the adjusted residual evaluated at the maximum likelihood estimates, $\boldsymbol{\pi}(\hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{\beta}}$. Anderson (1990) uses adjusted residuals for the one-parameter or Rasch item response model. The results here for the two-parameter model are more general.

Example 1

For an example of the use of adjusted residuals in assessing the adequacy of a model, data on sex role expectations is taken from a previous analysis by Duncan (1979). In the 1953 Detroit Area Study, a sample of 257 mothers were asked the following question regarding sex role expectations: "Here are some things that might be done by a boy or a girl. Suppose the person were 13 years old. As I read each of these to you, I would like you to tell me if it should be done as a regular task by a boy, by a girl, or by both.

(1) Shoveling walks
(2) Washing the car
(3) Dusting furniture
(4) Making beds" (Duncan, 1979, p. 252).

Responses of "boy" to Items 1 and 2 and "girl" to Items 3 and 4 were coded as "0" and are referred to as the traditional answers. Responses of "both", which Duncan refers to as the egalitarian answer, were coded as "1" for each item. Duncan deleted the "very few respondents" giving item cross-gender responses, for example, answers of "boy" to Items 3 or 4, or "girl" to Items 1 or 2.

For this data, the logit/probit model of Expressions (3) and (4) does not show a very good fit. The likelihood ratio statistic G^2 has a value of 25.7 on seven degrees of freedom, $p < 0.01$. Although the fit is poor, an examination of the standardized residuals shown in

Table 1

Standardized and Adjusted Residuals for Sex Role Data

Response Pattern	n_s	$n\hat{\pi}_s$	Standardized Residual	Adjusted Residual
0000	86	91.18	-0.54	-2.80
1000	20	18.88	0.26	0.58
0100	12	11.25	0.23	0.44
1100	8	4.19	1.86	2.09
0010	7	5.53	0.63	1.45
1010	2	4.21	-1.08	-1.25
0110	4	3.80	0.10	0.12
1110	2	3.43	-0.77	-1.02
0001	24	23.09	0.19	0.46
1001	12	10.55	0.45	0.55
0101	8	8.21	-0.07	-0.09
1101	1	5.17	-1.83	-2.46
0011	21	13.24	2.13	3.00
1011	7	13.69	-1.81	-3.11
0111	8	14.92	-1.79	-3.19
1111	35	15.69	1.84	3.79

Table 1 reveals only one large value, 2.13, which is associated with response pattern (0011). In his original ânalysis of this data, Duncan was looking for evidence to support a theory of ideological responding—namely, that some respondents would always give traditional answers, regardless of the situation, while others would always give egalitarian answers, regardless of the situation. Those who answered from a nonideological basis were called situational responders.

The assumption of conditional independence in (2) implies that conditional on the latent variable X, the manifest variables are independent. Conditional independence subsumes an assumption of local independence, the principle that responding to any item should not influence responses to the other items. Sometimes specific features of items, such as similar wording, may induce violations of local independence. The essence of the response consistency model, as it has been called, is that there is a complete failure of local independence for two domains of the population. Within these domains, which are identified with the response patterns 0000 and 1111, respondents give answers to some questions by invoking an ideology, and once invoked, that ideology determines that any other questions on the same topic will be answered in exactly the same way, without regard to considerations about specific situations presented in the questions. If the answers to items

i and j are represented as $Y_i = y_i$ and $Y_j = y_j$, then for an ideological responder, $P(Y_i = Y_j) = 1.0$ for $i \neq j$. This probability is not consistent with the model given in (3) and (4).

Although not every respondent answering with a completely consistent response pattern does so on the basis of ideology, one piece of evidence in support of ideological responding would be a poor fit for the entirely consistent response patterns (0000 and 1111) under some baseline model. Judging on the basis of the *standardized residuals*, there would seem to be little support for the contention that the poor fit of the logit/probit model could be attributed to the consistent response patterns. However, by examining the *adjusted residuals*, it becomes apparent that the consistent response patterns are indeed poorly fit by the model. Pattern (1111), in fact, has the largest residual. It appears that using the logit/probit model of a single continuous latent variable as a baseline, egalitarian responders are under represented by the model, and consistent traditional responders are over represented. Although adjusted residuals were not available for Duncan to examine, he used a model with latent classes for consistent responders to obtain a good fit for the data. See also Tuch (1981).

This example has shown that the traditional standardized residual may be completely inadequate for identifying response patterns that are poorly fit by the IRT model. Moreover, the large adjusted residuals were associated with response patterns in a manner suggested by a substantive theory that the responses of some individuals to attitudinal questions cannot be described by the model of Expressions (3) and (4) due to the influences of ideology. It is noteworthy that several large residuals are present for patterns besides those for consistent responding. These other large residuals suggest that response consistency may not be the only model that is appropriate for the data. See Reiser and Schuessler (1990) for further discussion. Configural Frequency Analysis (von Eye, 1990) is essentially equivalent to the use of the residuals in this example.

First- and Second-Order Marginal Frequencies

When the number of manifest variables is large (greater than seven or eight), and the sample size is moderate, there are often so many cells in the vector **V** that have low expected values (even with a sample as large as, say, 3000), that the chi-square approximation for the distribution of the likelihood ratio and Pearson goodness-of-fit statistics is not valid (Koehler, 1986; Reiser & VandenBerg, 1994; Tate & Hyer, 1973). In such a situation, it may be more reasonable to assess the fit of the model on the first- and second-order marginal frequencies (i.e., the univariate and bivariate distributions of the manifest variables), rather than the T cells of the entire vector **V**. The first- and second-order marginal frequencies of the manifest variables are almost always substantially larger than zero, even in small data sets; and when sparseness is present, tests that are a function of the marginals usually perform better than tests based on the full cross-classification (Agresti, Lipsitz & Lang, 1992; Agresti & Yang, 1987). In this discussion, the reference to marginal distributions indicates the distribution of manifest variable(s) with respect to the other *manifest variables*. The use of the term marginal in this section should not be confused with the usage in section 1, where it referred to the distribution of the manifest variables relative to the *latent* variable.

In the remainder of this section, adjusted residuals are developed for the first- and second-order marginal frequencies. Christoffersson (1975) and Muthén (1978, 1988) give a method for *estimation* of parameters and a large sample test of fit using only the first- and second-order marginals. Results from the method given below will be compared to results obtained by their methods.

Since **Y** is a vector of discrete variables, the first, and second-order marginal distributions for elements of **Y** can be obtained by ordinary summation. Under the model the first-order marginal proportion for variable i can be defined by

$$P_i(\boldsymbol{\beta}) = P(Y_i = 1|\beta) = \sum_s y_{is}\pi_s(\boldsymbol{\beta}),$$

where y_{is} is element i of response pattern s, and takes on the value 0 or 1; the true first-order marginal proportion is given by

$$P_i = P(Y_i = 1) = \sum_s y_{is}\pi_s,$$

where π_s is the true proportion for cell s.

Under the model, the second-order marginal proportion for variables Y_i and Y_j can be defined as

$$P_{ij}(1, 1|\boldsymbol{\beta}) = P(Y_i = 1, Y_j = 1|\boldsymbol{\beta}) = \sum_s y_{is}y_{js}\pi_s(\boldsymbol{\beta}),$$

and the true second-order marginal proportion is given by

$$P_{ij}(1, 1) = P(Y_i = 1, Y_j = 1) = \sum_s y_{is}\underline{y_{js}\pi_s}.$$

The summation across the response patterns to realize the marginal proportions represents a transformation of the joint proportions in the vector $\boldsymbol{\pi}$. The summation can be implemented via multiplication by matrix **M**, which has elements consisting of 0's and 1's. The number of rows in **M** is equal to $\frac{1}{2}k(k + 1)$, which is the number of first- and second-order marginals, and the number of columns is 2^k, which is the number of cells in $\boldsymbol{\pi}$. Elements of **M** are determined from the response patterns associated with the cells of $\boldsymbol{\pi}$. In order to obtain the first-order marginals, elements of **M** are assigned values so that the proportions are summed for all cells where y_i has the value "1" in the associated response pattern. Similarly, to obtain the second-order marginals, elements of **M** are assigned values so that the proportions are summed for all cells where both y_i and y_j have the value "1". Specifically, for $j = 1, 2, \ldots k$; $i = j, j + 1 \ldots k$; $s = 1, 2, \ldots T$; and $\ell = (j - 1)k + 0.5(j)(j - 1) + i$, element ℓs of **M** is given by

$$m_{\ell s} = \begin{cases} 1 & \text{if } y_{is} = y_{js} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Define $\boldsymbol{\Psi}$ as the $k \cdot k$ matrix of first- and second-order marginal proportions, with element $P_{ij}(1, 1)$ equal to a first-order marginal if $i = j$ and a second-order marginal if $i \neq j$. Let vech($\boldsymbol{\Psi}$) represent the diagonal and lower off diagonal elements of $\boldsymbol{\Psi}$ stacked into a vector, then

$$\text{vech}(\boldsymbol{\Psi}) = \mathbf{M}\boldsymbol{\pi}.$$

Since the matrix **M** consists of fixed 0's and 1's, the first- and second-order marginals in $\boldsymbol{\Psi}$ represent linear combinations of the multinomial proportions in $\boldsymbol{\pi}$.

There are a number of ways to define a residual for the first- or second-order marginal. Some possibilities are

$$e = \hat{P}_{ij}(1, 1) - P_{ij}(1, 1|\hat{\boldsymbol{\beta}}),\ e = \frac{\hat{P}_{ij} - P_{ij}(1, 1|\hat{\boldsymbol{\beta}})}{P_{ij}(1, 1|\hat{\boldsymbol{\beta}})^{1/2}} \text{ and } e = \frac{\hat{P}_{ij} - P_{ij}(1, 1|\hat{\boldsymbol{\beta}})}{P_{ij}(1, 1|\hat{\boldsymbol{\beta}})},$$

$$\text{where } P_{ij}(1, 1|\hat{\boldsymbol{\beta}}) = P(Y_i = 1, Y_j = 1|\hat{\boldsymbol{\beta}}) = \mathbf{m}'_\ell\boldsymbol{\pi}(\hat{\boldsymbol{\beta}}),$$

$$\hat{P}_{ij}(1, 1) = \hat{P}(Y_i = 1, Y_j = 1) = \mathbf{m}'_\ell\hat{\mathbf{p}},$$

and $\mathbf{m}'_\ell$ is row ℓ of matrix $\mathbf{M}$ defined earlier. Although the definition given above in the middle has a similar form to the traditional standardized residual, the unstandardized residual, $e = \hat{P}_{ij}(1, 1) - P_{ij}(1, 1|\hat{\boldsymbol{\beta}})$, will be used because the presentation of results in subsequent sections will be simplified considerably by that choice. Although these results will be based on the unstandardized residual, the results are valid for the standardized residual as well.

The vector of unstandardized residuals for the first- and second-order marginals is defined by the following transformation:

$$\mathbf{e} = \mathbf{M}(\hat{\mathbf{p}} - \boldsymbol{\pi}(\hat{\boldsymbol{\beta}})). \tag{8}$$

Result 2. Assuming the regularity conditions given by Birch (1964a),

$$n^{1/2}\mathbf{e} \overset{L}{\to} N(\mathbf{0}, \boldsymbol{\Omega}_\mathrm{e}),$$

$$\text{where } \boldsymbol{\Omega}_\mathrm{e} = \mathbf{M}\boldsymbol{\Omega}_\mathrm{u}\mathbf{M}',$$

$$\boldsymbol{\Omega}_\mathrm{u} = \mathbf{D}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}' - \mathbf{G}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{G}',$$

$$\text{and } \mathbf{G} = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\beta})}{\partial \mathrm{vec}\boldsymbol{\beta}}.$$

The validity of Result 2 can be shown by an application of the multivariate delta method (the method of statistical differentials). It can be seen from expression (8) that the elements of $\mathbf{e}$ are linear combinations of the unstandardized residuals, $\mathbf{u} = \hat{\mathbf{p}} - \boldsymbol{\pi}(\hat{\boldsymbol{\beta}})$, associated with the multinomial proportions. The covariance matrix of $\mathbf{e}$ can be found by starting with the covariance matrix for $\mathbf{u}$, which closely resembles (7).

Define the following vector valued function of $\mathbf{p}$ and $\boldsymbol{\pi}(\boldsymbol{\beta})$:

$$\mathbf{h}(\mathbf{p}, \boldsymbol{\pi}(\boldsymbol{\beta})) = \mathbf{p} - \boldsymbol{\pi}(\boldsymbol{\beta}).$$

Then $\mathbf{u} = \mathbf{h}(\hat{\mathbf{p}}, \boldsymbol{\pi}(\hat{\boldsymbol{\beta}}))$, and by Theorem 14.6-2 of Bishop, Fienberg and Holland (1975),

$$\mathbf{u} \overset{L}{\to} N\left(\mathbf{0}, \left(\frac{\partial \mathbf{h}}{\partial \mathbf{p}}\right)\Sigma_{\hat{p}}\left(\frac{\partial \mathbf{h}}{\partial \mathbf{p}'}\right)\right). \tag{9}$$

The use of expression (9) requires the partial derivative of $\mathbf{h}(\mathbf{p}, \boldsymbol{\pi}(\boldsymbol{\beta}))$ with respect to $\mathbf{p}$ and an expression for $\boldsymbol{\Sigma}_{\hat{\mathbf{p}}}$. Proceeding to obtain the necessary expressions,

$$n^{1/2}(\hat{\mathbf{p}} - \boldsymbol{\pi}) \overset{L}{\to} N(\mathbf{0}, \mathbf{D}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'), \tag{10}$$

by Theorem 14.3.4 in Bishop, Fienberg and Holland, which gives $\boldsymbol{\Sigma}_{\hat{\mathbf{p}}}$.

The partial derivative of $\mathbf{h}(\mathbf{p}, \boldsymbol{\pi}(\boldsymbol{\beta}))$ with respect to $\mathbf{p}$ follows from the chain rule:

$$\frac{\partial \mathbf{h}}{\partial \mathbf{p}} = \mathbf{I} - \frac{\partial \boldsymbol{\pi}}{\partial\, \mathrm{vec}\, \boldsymbol{\beta}} \frac{\partial\, \mathrm{vec}\, \boldsymbol{\beta}(\mathbf{p})}{\partial \mathbf{p}}.$$

$\boldsymbol{\beta}$ as a function of $\mathbf{p}$ is not known explicitly, but the existence of that function can be established by the Implicit Function Theorem. Using this approach, Birch (1964a) obtains

$$\frac{\partial\, \mathrm{vec}\, \boldsymbol{\beta}(\mathbf{p})}{\partial \mathbf{p}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{D}(\boldsymbol{\pi})^{-1/2}.$$

Then, with $\mathbf{G} = \partial\boldsymbol{\pi}/\partial\, \mathrm{vec}\, \boldsymbol{\beta}$,

$$\frac{\partial \mathbf{h}}{\partial \mathbf{p}} = \mathbf{I} - \mathbf{G}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{D}(\boldsymbol{\pi})^{-1/2},$$

when evaluated at $\mathbf{p} = \boldsymbol{\pi}$, the true value.

Finally, applying these results to (9),

$$n^{1/2}\mathbf{u} \xrightarrow{L} N(\mathbf{0}, \boldsymbol{\Omega}_{\mathbf{u}}),$$

where

$$\boldsymbol{\Omega}_{\mathbf{u}} = (\mathbf{I} - \mathbf{G}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{D}(\boldsymbol{\pi})^{-1/2})(\mathbf{D}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'))(\mathbf{D}(\boldsymbol{\pi})^{-1/2}\mathbf{A}\mathbf{G}'(\mathbf{A}'\mathbf{A})^{-1} - \mathbf{I}).$$

After multiplying, and using $\mathbf{A}'\boldsymbol{\pi}^{1/2} \underline{=} \Sigma\ \partial\boldsymbol{\pi}(\boldsymbol{\beta})/\partial \text{ vec } \boldsymbol{\beta} = 0$, the expression simplifies as follows:

$$\boldsymbol{\Omega}_{\mathbf{u}} = \mathbf{D}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}' - \mathbf{G}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{G}'. \tag{11}$$

Now returning to the residuals on the first- and second-order marginals, which are linear combinations of the elements in **u**, Result 2 follows from (11) and Result 6a.1(ii) of Rao (1973, p. 383).

Define $\boldsymbol{\Sigma}_{\mathbf{e}}$ to be the asymptotic covariance matrix of the first- and second-order residuals, with estimator $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$ defined by

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{e}} = n^{-1}\mathbf{M}(\mathbf{D}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}' - \mathbf{G}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{G}')\mathbf{M}'\big|_{\boldsymbol{\pi} = \boldsymbol{\pi}(\hat{\boldsymbol{\beta}}), \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} \tag{12}$$

$n\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$ is consistent for $\boldsymbol{\Omega}_{\mathbf{e}}$.

The rank of $\boldsymbol{\Sigma}_{\mathbf{e}}$ will be needed in the next section in order to determine degrees of freedom for the distribution of a quadratic form, so the rank is discussed now. $\boldsymbol{\Sigma}_{\mathbf{e}}$ can be factored as follows:

$$\mathbf{M}\mathbf{D}(\boldsymbol{\pi})^{-1/2}(\mathbf{I} - \boldsymbol{\pi}^{1/2}(\boldsymbol{\pi}')^{1/2} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')\mathbf{D}(\boldsymbol{\pi})^{-1/2}\mathbf{M}'.$$

Under the condition that all $\pi_s > 0$, $\mathbf{D}(\boldsymbol{\pi})^{1/2}$ is a square, full-rank diagonal matrix, so the rank of $\boldsymbol{\Sigma}_{\mathbf{e}}$ will be determined by the rank of $(\mathbf{I} - \boldsymbol{\pi}^{1/2}(\boldsymbol{\pi}')^{1/2} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')$ and the rank of **M**. $(\mathbf{I} - \boldsymbol{\pi}^{1/2}(\boldsymbol{\pi}')^{1/2} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')$ is idempotent with rank equal to trace which is equal to $2^k - 2k - 1$. **M** generally has full row rank equal to $\frac{1}{2}k \cdot (k + 1)$. Therefore the rank of $\boldsymbol{\Sigma}_{\mathbf{e}}$ will in general be equal to min $[2^k - 2k - 1, \frac{1}{2}k \cdot (k + 1)]$. Certain configurations of true proportions can reduce the rank. For example, for any i and j where $\beta_{0i} = \beta_{0j}$ and $\beta_{1i} = \beta_{1j}$, the rank of $\boldsymbol{\Sigma}_{\mathbf{e}}$ will be reduced by 2.0.

As in the previous section, estimated standard errors for the residuals can be obtained by taking square roots of the diagonal elements of $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$.

Example 2

In this example, the responses of 917 German workers to eight questions regarding job satisfaction are analyzed. The items are from Krebs and Schuessler (1987), and have also been studied by Reiser and Schuessler (1990). The eight items and their means (proportion keyed responses) are as shown in Table 2.

Fitting the logit/probit model described earlier gives a likelihood ratio statistic of $G^2 = 150$, on, ostensibly, 239 degrees of freedom. However, a large number of response patterns are unrepresented in the data set. Of the $2^8 = 256$ response patterns, 150 showed no cases and another 50 showed only one. Under these circumstances, many expected cell frequencies will be very close to zero, and the tabulated significance level of G^2, as taken from a chi-square distribution, will not be valid. Simulation results from Reiser and VandenBerg (1994) show that for eight variables and a sample size of around 1000, the G^2

Table 2

Job Satisfaction Items

No.	Direction	Item	Mean	KEY
1	P	Satisfied	.88	A
2	N	Like more freedom	.62	D
3	P	Workers feel they belong	.83	A
4	N	Too little variety	.82	D
5	P	Job gives me a chance	.69	A
6	N	Little share in deciding	.69	D
7	P	Job means more than money	.75	A
8	N	Must be better places to work	.25	D

statistics has low power due to sparseness. One possible correction for this problem would be to combine small frequencies, which for this data yields 39 degrees of freedom. The adjustment of combining cells works well if there is only a small number of cells with low expected values. When there is a large number of small frequencies, as here, combining them is not satisfactory, because the data used for the test of fit does not even resemble the data used for estimation of the model unless there is a natural way to carry out the combinations (Agresti, 1990, p. 247). In this situation, both tests, on the original 256 response patterns and on the combined frequencies, have very little validity for assessing the fit of the model.

Because the frequencies for the response patterns are so sparse, it may be more useful to examine how well the model reproduces the first- and second-order marginal frequencies, which are almost always substantially larger than zero. These frequencies, as well as discrepancies between observed and expected values under the model are shown in Table 3, parts (a) and (b). Most of the differences between observed and expected values are small, but some differences, with values around 10 and 13, appear to be large. By looking only at the frequencies, it is difficult to know how large differences should be in order to be considered too large. In order to assess the magnitudes of the differences, standardized residuals and adjusted residuals are shown in Table 3, parts (c) and (d). None of the standardized residuals shown in Table 3(c) are large; but when judged relative to their standard errors, so that each has an approximate N(0, 1) distribution as in Table 3(d), some of them are quite large. Among the several large adjusted residuals, -2.817, which corresponds to the second-order marginal frequency between variables two and seven, has the largest magnitude. The probability that a standard normal deviate exceeds 2.817 in absolute value is only 0.0048. Given that this residual was selected from the total of 36 residuals in the table, one cannot conclusively state that the model of a single latent variable does not fit the data, because with simultaneous inference on several residuals, the large magnitudes may be due just to chance. However, it seems unlikely that the model adequately accounts for the association between several pairs of variables, and in the next section some methods are considered to assess the magnitude of the residuals jointly. It is notable that the adjusted residuals for the first-order marginals are quite small. Any large residuals associated with the first-order marginals would have been an indication that the item response curve (1) may have been misspecified. Since the first-order marginals are fit so well, there is no specific evidence among the residuals that lack of fit could be due to use of the two-parameter logistic item response curve. As mentioned previously, other methods may be more appropriate for inquiries as to the choice of item response curve.

Table 3

Observed Frequencies and Residuals

(a)	Matrix of Observed First-, and Second-Order Frequencies							
	1	2	3	4	5	6	7	8
1	107							
2	76	352						
3	39	90	155					
4	58	114	61	169				
5	71	163	79	109	283			
6	66	199	89	127	169	344		
7	58	111	71	86	106	128	229	
8	100	305	131	155	249	291	200	689
(b)	Matrix of Residuals for First-, and Second-Order Frequencies							
1	1.55							
2	4.53	1.40						
3	0.90	1.96	1.12					
4	0.121	-1.47	0.34	2.67				
5	4.37	1.41	0.56	0.88	1.77			
6	8.03	10.31	-0.88	6.87	3.78	1.59		
7	7.17	-13.61	10.62	4.49	-3.53	1.29	1.26	
8	3.81	6.87	-2.29	1.73	3.62	-4.21	5.20	0.30

A multivariate procedure for assessing the magnitude of the residuals as a set is available in a QPLOT. Figure 1 shows such a plot, where each normalized (adjusted) residual is plotted against its normalized quantile. The large negative residual, −2.817, is not included in the figure, resulting in a display where a line drawn through the residuals would with slope roughly equal to 1.0. When the residuals fall along this 45° line, the pattern is usually interpreted as an indication of a moderate fit. However, if the large negative value were included, a line through the residuals would have a slope substantially greater than 1.0, so the QPLOT indicates a moderate to poor fit for the model. Strictly speaking, the QPLOT method assumes independently distributed residuals, an assumption which does not hold for the data in Figure 1. Although the plot seems useful here, a more rigorous method would be desirable. Such a method is given in the next section. Ludlow

Table 3 (cont.)

Observed Frequencies and Residuals

(c)	Standardized residual: $n^{1/2}\hat{E}^{-1/2}(O - \hat{E})$							
1	0.146							
2	0.536	0.075						
3	0.145	0.209	0.087					
4	0.158	-0.140	0.044	0.209				
5	0.535	0.110	0.063	0.087	0.107			
6	0.933	0.750	0.093	0.630	0.296	0.087		
7	1.006	-1.218	1.367	0.497	0.335	0.116	0.086	
8	0.389	0.400	-0.199	0.137	0.230	-0.245	0.373	0.014
(d)	Adjusted Residuals: $\hat{\sigma}^{-1}n^{\frac{1}{2}}$							
1	1.724							
2	1.307	1.420						
3	0.262	0.457	1.684					
4	0.536	-0.414	0.111	1.501				
5	1.330	0.306	1.386	0.270	0.513			
6	**2.466**	**2.217**	0.218	**1.980**	0.887	1.399		
7	**1.972**	**-2.817**	**2.446**	1.292	0.779	0.287	1.687	
8	1.469	1.579	-0.662	0.547	0.934	-1.036	1.279	0.812

(1986) gives a much more extensive treatment on graphical analysis of residuals in item response models.

3. A Limited-Information Test of Fit

We may define differences associated with the first-, and second-order marginals such that

$$P_i = P_i(\boldsymbol{\beta}) + \delta_i, \text{ and}$$

$$P_{ij}(1, 1) = P_{ij}(1, 1|\boldsymbol{\beta}) + \delta_{ij}.$$

The error terms defined above may be placed into the vector $\boldsymbol{\delta}$, where

$$\boldsymbol{\delta} = \mathbf{M}(\boldsymbol{\pi} - \boldsymbol{\pi}(\boldsymbol{\beta})),$$

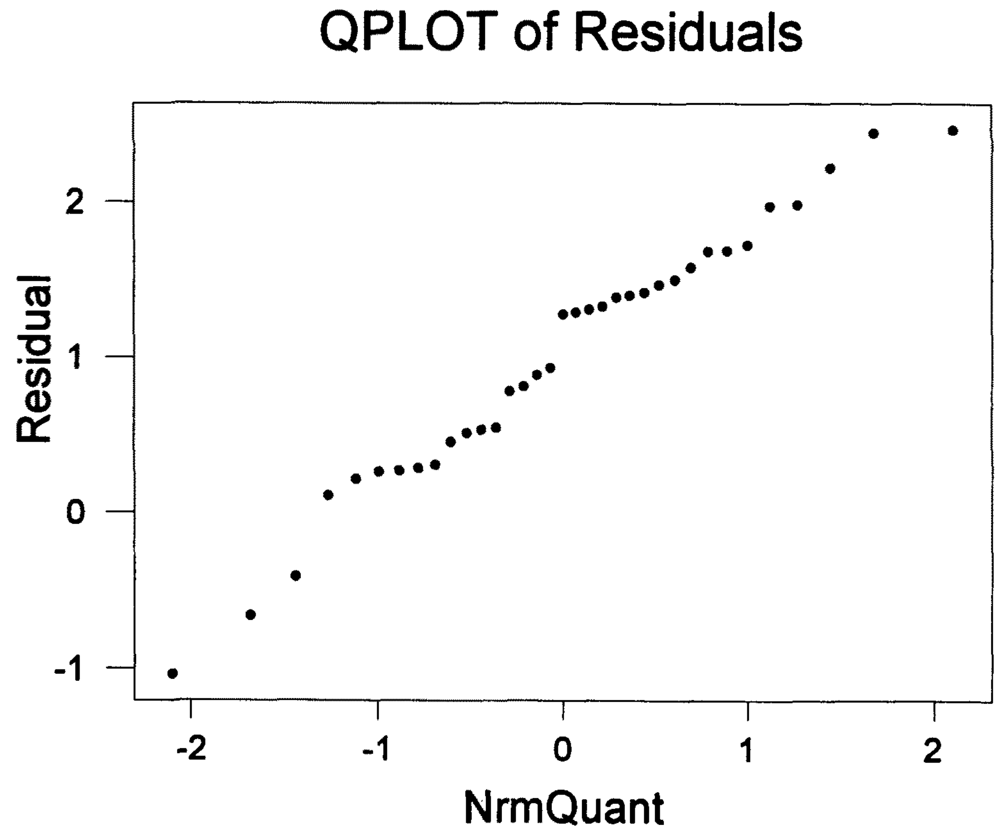

FIGURE 1.
Plot of normalized residuals versus normalized quantiles.

A more precise method to assess the residuals as a set would be to test the null hypothesis that $\boldsymbol{\delta}$, the vector of difference, is equal to the zero. The third main result of this paper gives a statistic that may be used for this test.

Result 3. Under the null hypothesis H_0: $\boldsymbol{\delta} = 0$, and assuming the regularity conditions of Birch (1964a), the statistic

$$X_w^2 = \mathbf{e}'\hat{\Sigma}_e^{-1}\mathbf{e} \tag{13}$$

has a limiting chi-square distribution.

Since it has already been established that the limiting distribution of $\mathbf{e}$ is multivariate normal, and since $\hat{\Sigma}_e$ is converging stochastically to Σ_e, the limiting distribution of X_w^2 as $n \rightarrow \infty$ is the χ^2-distribution, by Theorem 2.4.5 of Anderson (1984). The degrees of freedom will be determined by the rank of Σ_e, which was given in the previous section as generally equal to min $[\frac{1}{2}k \cdot (k+1), 2^k - 2k - 1]$. In some cases, as discussed above, the degrees of freedom will be reduced. For $k \leq 4$, Σ_e will be singular. The calculation of X_w^2 requires an estimator for Σ_e, which may be based on (12) as given in the previous section.

In practice, there may be high multicollinearity between the linear combinations that produce the second-order marginals when some items have nearly identical parameters. In this case, it is advantageous to reformulate the model with equality restrictions on the appropriate parameters, and to reduce the dimension of Σ_e. Multicollinearity may also be present among the linear combinations that produce the first-order marginals and the basis vectors that represent the intercept parameters. In such a case, it is more practical to calculate X^2_w based on residuals for only the second-order marginals. Also, if only second-order marginals are used when $k = 4$, then $\mathbf{\Sigma}_e$ will be full rank.

Example 2. Continued

In this application, X^2_w, which represents the statistic calculated on the first- and second-order marginals, is equal to 61.64 on 36 degrees of freedom (p < 0.005), indicating a poor fit and a sharp discrepancy from the likelihood ratio test considered earlier. Given the poor fit and the residuals shown in Figure 1, it seems that the single factor model provides a poor representation of the relationship between several item pairs, particularly Item 2 and Item 7. The aspects of work satisfaction tapped by these item pairs require at least a second latent variable which, as suggested by item content, may be associated with self-fulfillment.

The procedure described above represents full-information estimation of parameters with a limited-information test of fit. Christoffersson (1975) and Muthén (1978) give methods where both estimation and the test of fit are based on limited information. Using the either the FADIV (Andersson, Christoffersson, Muthén, 1974) or LISCOMP (Muthén, 1988) computer programs for limited-information estimation/limited-information test of fit, a chi-square value of 36.85 on 20 degrees of freedom is obtained for a one-factor model ($0.01 < p < 0.05$). Although the test in the preceding paragraph has a much lower p-value, the decision regarding the model would be the same for both tests at the $\alpha = 0.05$ level. Even though these two limited-information tests are both calculated on residuals associated with the first- and second-order marginals, they are not equivalent and do not have the same degrees of freedom. It should be noted that while adding parameters to the model may change the value taken on by the statistic given in (13), doing so will not necessarily change the degrees of freedom for the distribution of the statistic under H_0.

In most applications, these two tests can be expected to yield the same conclusion. However, in smaller samples with a large number of variables, the tests may often give different results. Simulations by Reiser and VandenBerg (1994) using limited-information estimation/limited-information test of fit show that as the model becomes larger, larger sample sizes are needed for the asymptotic chi-square approximation to be valid. Similar results are found in Muthén and Kaplan (1992). Simulations now in progress using the method developed above—full information estimation/limited-information test of fit—show that the asymptotic chi-square approximation for the test statistic performs well even with a large number of variables in smaller samples.

Any significant differences in performance between the test developed above and the test from FADIV or LISCOMP could be due to the use of a different estimator for the covariance matrix of the residuals, and/or to the use of different information when calculating parameter estimates. Since the estimator for the covariance matrix given here is based on the maximum likelihood estimator of $\boldsymbol{\pi}$, it may be more efficient in moderate size samples than the estimator given by Christoffersson (1975), which is based on the sample proportions. See Agresti (1990, p. 182.) This estimated covariance matrix for the residuals appears directly in the expression for the test statistic calculated on the marginals, so more efficiency in estimating the covariance matrix could bring about superior performance for the test statistic given in (13). Parameter estimates also influence the magnitude of the test statistic, primarily through the calculation of residuals. Since the full-information parameter estimator presented here may have a smaller mean square error than the limited-information parameter estimator, the residual calculated from it would, in turn, have smaller mean square error. Less error in the estimation of the residual could also contribute to superior performance for the test given in (13). More research is needed on the comparison of these two limited-information tests. Another estimator that could be considered for the covariance matrix of the residuals has been given by Küsters (1990).

A comparison of results from the limited-information test to results on a full-information procedure, such as the likelihood ratio test, should also be considered. In many instances, the limited-information test developed in the previous section may produce the same conclusion as a full-information test. On the other hand, it may produce a different result, and if it does so, the difference could be attributed to one or more of the following reasons: sparseness in the joint frequencies (as seen in the preceding example), third- and higher-order effects that are not reflected in the first- and second-order marginals, or differential power of the tests. Sparseness in the joint frequencies tends to invalidate the chi-square approximation for the distribution of the full-information statistics. When the number of variables is small relative to the sample size, sparseness is not an issue in the discrepancies between results from the full-, and limited-information tests, so then only higher-order effects and differential power are relevant. Finally, if neither sparseness nor higher-order effects are present, the limited- and full-information tests may still not have the same power against a false null hypothesis.

The performance of the limited-information test relative to the full-information test of the model will depend to a large degree on the *collapsibility* of the contingency table that corresponds to the joint distribution of the manifest variables. Loosely formulated, collapsibility invokes the notion that inferences regarding certain parameters may be performed without loss of information in smaller marginal tables. Although the phrase *without loss of information* seems to imply without loss of efficiency, existing definitions of collapsibility (Bishop, Feinberg, & Holland, 1975; Kreiner, 1987) focus on parameter values rather than efficiency of estimators. Thus, it seems that in this context, the phrase *without loss of information* should be interpreted to mean without change of information.

Collapsibility can be defined in terms of odds-ratios. The 2^k table of manifest variables can be considered as a set of $2 \times 2 \times Q$ tables, where $Q = 2^{k-2}$. There would be $\frac{1}{2}k \cdot (k - 1)$ such tables, one for each pair of manifest variables. Let the variables in this three dimensional table be labeled as Y_i, Y_j, and C_t. C_t represents a compound variable obtained by a cross-classification of the manifest variables excluding Y_i and Y_j.

Define the odds ratio conditional on C_t as

$$OR_{ijt} = \frac{\{P(Y_{j2}|Y_{i1}, C_t)/P(Y_{j1}|Y_{i1}, C_t)\}}{\{P(Y_{j2}|Y_{j2}, C_t)/P(Y_{j1}|Y_{i2}, C_t)\}},$$

and define the marginal odds ratio as

$$OR_{ij.} = \frac{\{P(Y_{j2}|Y_{i1})/P(Y_{j1}|Y_{i1})\}}{\{P(Y_{j2}|Y_{i2})/P(Y_{j1}|Y_{i2})\}}.$$

The $2 \times 2 \times Q$ table for variables Y_i and Y_j is collapsible if

$$OR_{ijt} = \alpha_{ij} = OR_{ij.},$$

where α_{ij} indicates a constant pertaining to the table for variables Y_i, Y_j and C_t. The 2^k table of manifest variables is collapsible to the set of second-order marginals if each $2 \times 2 \times Q$ table is collapsible in the manner just defined. Note that this definition of collapsibility is not equivalent to the one given by Bishop, Feinberg, and Holland (1975).

Collapsibility to second-order marginals should be a sufficient condition for the test of H_0: $\boldsymbol{\delta} = 0$ on the marginals to be obtained from the same information as the full-information test. It would not be a necessary condition for both tests to give the same result, however, because if the model of (3) and (4) does not fit the joint frequencies and if collapsibility does not hold, the model could still be rejected by the test on the marginal frequencies. Whether or not H_0: $\boldsymbol{\delta} = 0$ is rejected under this condition is an empirical question that will depend in each case on the extent to which the presence of both higher-order effects in the joint frequencies and confounding result in marginal frequencies that do not satisfy the conditional independence stated in (2). To the extent that any higher-order association effects that do not correspond to the model are not well represented in the marginals, there may be an increased Type II error rate for the test on the marginals as interpreted as a test of fit for the model. In this sense, the test on the second-order marginals is weaker than the full-information test.

In a case where the model of (3) and (4) is an incorrect representation of the relationships among variables, and the inadequacy of the model is detectable only in the high-order interactions among the Y variables, there should be a large discrepancy between the results (i.e., the p-values) from the full-information and limited-information tests. The model would normally be rejected by the full-information test, but if the number of variables is large, the full-information test will suffer from the effects of sparseness, and the result may have no validity. In this latter situation, a comparison of the full- and limited-information test results would not be useful; but a comparison of limited-information to full-information parameter estimates may be informative for the following reason: If the model is not correct, and the table of manifest variables is not collapsible to the second-order marginals, the relationships in the joint frequencies may be distorted in the table of marginal frequencies. In this circumstance, parameter estimates based on the collapsed table will not necessarily be equivalent to full-information estimates based on the joint frequencies. Therefore, large discrepancies between full- and limited-information parameter estimates should be viewed as a warning that third- or higher-order effects may be present and that the power of the limited-information test may not be adequate.

When neither sparseness nor higher-order effects are present, there is still a question of the power of the statistical test on the marginals versus the power of the test on the joint proportions. Results from Birch (1964b) based on Theorem 3 of Lehman (1959, p. 136) establish that under the condition of no third- or higher-order interactions, some tests based on the second-order marginal are uniformly most powerful. Although the test developed in section 3 uses first- as well as second-order marginals, simulations now in progress support the applicability of Birch's result; that is, the limited-information test appears to be more powerful than the full-information test if interactions higher than the second-order are not present. Thus, even under the condition of collapsibility for the table of manifest variables as discussed above, there would not be an equivalence between the full- and limited-information tests.

4. Discussion and Conclusions

For assessing the adequacy of a latent variable model, an examination of the adjusted residuals associated with the response patterns may be very useful if the number of manifest variables is small. These residuals can be used in conjunction with the chi-square goodness-of-fit tests.

Because of the sample sizes typically encountered in empirical studies, the adequacy of the model is difficult to assess with traditional fit statistics when the number of manifest variables is larger than, say, seven or eight. Because the frequencies associated with the response patterns become so low, the chi-square approximation for traditional goodness-of-fit tests is not valid, and the residuals associated with the response patterns are also less useful. Because of the low frequencies, the variances of the standardized residuals are all near 1.0, and the adjusted residuals are nearly identical to the standardized residuals. Under this condition, large values for the residuals are often found for cells with an expected value near zero and an observed value of 1.0. The main problem is the low frequencies, but even if the expected value of the count is large, it can be difficult to interpret the meaning of a poor fit to a complicated response pattern such as (11001001).

When the number of variables is large relative to the sample size, it is therefore more useful to examine residuals calculated for the second-order marginal frequencies (or proportions), as demonstrated in example two. These residuals can be combined into a single statistic for testing goodness of fit, and they may be examined individually or in subsets for the purposes of interpreting departures from the model and developing modifications to the model, such as including another latent variable. Applying the statistic given in (13) to subsets of first- and second-order marginals would be useful for testing the model to specific manifest variables. For example, the model could be tested for a single manifest variable by selecting all second-order marginals that involve that item. If there were eight manifest variables, this test for one variable would involve seven second-order marginals, and if the first-order marginal was included, the test would have eight degrees of freedom. The results of such a test could be used to determine if a particular variable should be dropped in order to obtain a set of items for which the model would provide a good description of the inter-item associations.

Based on the discussion in the previous section, it is appropriate to rely on the first-, and second-order marginals for assessing the model if there is no additional information contained in the higher-order marginals or cross-classifications. Since variables entered into an IRT analysis are chosen on the expectation that all items are indicators of the same latent variable, it would be unusual to find interactions above the second-order among such pre-selected variables. Results from an empirical study by Salomaa (1990) support this view. Salomaa's results are from simulation studies carried out with data sets obtained by generating responses for categorical manifest variables from the model of one or more continuous latent variable. The models and parameter values used in generating the data were taken from studies that applied latent variable models to real data. Then using the generated data, a hierarchical logit model with only first-, and second-order effects was fit to each data set, and in virtually all cases, the logit model provided a very good fit. Since the first-, and second-order marginal frequencies are sufficient statistics for the model of first- and second-order effects, Salomaa's results indicate that the first-, and second-order marginals contain essentially all of the information regarding parameter estimates and goodness of fit in data that corresponds to the model of a continuous latent variable.

It should be noted that Salomaa's results do not imply that the first- and second-order marginals are sufficient statistics (in the mathematical sense of Fisher's definition) for the model of a continuous latent variable, nor do the results imply that if some other model is correct, instead of the continuous latent variable model, that the first- and second-order marginals still contain essentially all relevant information. The significance of the results for applications of the item response model is that in commonly encountered data, the first- and second-order marginals represent virtually the same information as the joint frequencies. Therefore, methods which use the limited information in the first- and second-order marginals will usually give the same results as methods which use the full information in the joint frequencies. Some exceptions to this rule have been discussed in this paper. The limited-information test may actually be more powerful, if no higher-order interactions are present, and it is certainly more useful under conditions of sparseness in the joint frequencies.

Summary of Recommended Methodology

The methodology for analyzing residuals in applications of the item response model is summarized in the following recommendations, which are based on results from this paper and from Reiser and VandenBerg (1994).

Sparseness Not Present

When sparseness is not present, the test of fit for the item response model can be carried out by using both the full-information and limited-information methods. If the sample size is small relative to the number of variables, the limited-information test given in this paper performs more closely to the stated Type I error rate than other alternatives. Possible results are as follows:

1. The full-information and limited-information tests both indicate that H_0 cannot be rejected. In this case, the redundancy in the conclusions is reassuring, and the usual precautions regarding the possibility of a Type II error still apply.
2. Both the full-information and limited-information tests indicate that H_0 should be rejected. Again, the redundancy is reassuring, and the usual precautions about the possibility of a Type I error still apply. Adjusted residuals calculated on the response patterns can be used as a guide for detecting where the model fits poorly. Also, the statistic given in section 3 may be used to test a subset of items by selecting second-order marginals associated with specific items. If there is doubt about the choice of the function used to represent the item response curve, researchers may want to proceed with methodology from Stout (1987).
3. The limited-information test indicates the opposite conclusion from the full-information test. The researcher must choose which test he or she feels is more appropriate for the application under consideration. In order to understand the discrepancy between the tests, adjusted residuals could be calculated for both the response pattern frequencies and the second-order marginal frequencies. It is more likely that the limited-information test would be the one to indicate that H_0 should be rejected, since it is more powerful under the condition of no third-, or higher-order interactions. If the full-information test indicates reject H_0, but the limited-information test shows the opposite, a large discrepancy in the p-values may be a reflection of the presence of third-, or higher-order interactions in the joint frequencies of the manifest variables.

Sparseness Present

If sparseness is present in the data, and if the chi-square distribution is used to obtain p-values for the traditional goodness-of-fit tests based on the likelihood ratio and Pearson statistics, then the p-values may not be valid. If sparseness is severe, the chi-square distribution should not be used to obtain p-values for the traditional statistics. A normal approximation may be appropriate under conditions of sparseness, but this avenue has not been investigated for item response models.

Because of the problems mentioned above, the presence of sparseness necessitates placing more reliance on the limited-information test. Sparse cell frequencies imply that the sample size is small relative to the number of manifest variables, and the limited-information test given in this paper will perform more closely to the stated Type I error rate than other available alternatives. The test has two possible outcomes:

1. The limited-information test indicates reject H_0. The usual level of caution about the probability of a Type I error should be kept in mind. Adjusted residuals calculated on the second-order marginals will be useful as a guide for detecting where the model fits poorly. The fit for individual items may be assessed by using the statistic given in section 3 with a subset of second-order marginals.
2. The limited-information test indicates do not reject H_0. In the majority of cases, it can be assumed that the model gives an adequate representation of both the joint and marginal frequencies. However, in addition to the usual precautions regarding

Type II errors, the possibility that the second-order marginal frequencies do not represent the same information as the joint frequencies should also be taken into account. To check on this possibility, it may be useful to compare full-information parameter estimates to the limited-information parameter estimates. Large discrepancies in the comparison would indicate the presence of either confounding or higher-order interactions in the joint frequencies; if such discrepancies are found, the result of the limited-information test should be interpreted with caution, because the test would have low power for detecting departures from the model that are apparent among higher-order interactions but which are not well represented in the second-order marginals.

References

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.

Agresti, A., Lipsitz, S., & Lang, J. B. (1992). Comparing marginal distributions of large sparse contingency tables. *Computational Statistics & Data Analysis, 14,* 55–73.

Agresti, A., & Yang, M. C. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics & Data Analysis, 5,* 9–21.

Anderson, T. W. (1984). *An introduction to multivatiate statistical analysis*. (2nd ed.). New York: Wiley.

Andersen, E. B. (1990). *The statistical analysis of categorical data*. New York: Springer-Verlag.

Andersson, C. G., Christoffersson, A., & Muthén , B. (1974). *FADIV: A computer program for factor analysis of dichotomized variables*. Sweden: University of Uppsala, Department of Statistics.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. New York: Oxford University Press.

Birch, M. W. (1964a). The detection of partial association, I: the 2 × 2 case. *Journal of the Royal Statistical Society, Series B, 26,* 313–324.

Birch, M. W. (1964b). A new proof of the Pearson-Fisher Theorem. *Annals of Mathematical Statistics, 35,* 818–824.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivatiate analysis: Theory and practice*. Cambridge, MA.: MIT Press.

Bock, R. D., & Aitken, M. (1981). Marginal Maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika, 35,* 179–197.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40,* 5–32.

Cochran, W. G. (1954). Some methods for strengthening the common chi-square tests. *Biometircal Journal, 10,* 417–451.

Cochran, W. G. (1955). A test of a linear function of the deviations between observed and expected numbers. *Journal of the American Statistical Association, 50,* 377–397.

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

Duncan, O. D. (1979). Indicators of sex typing: Traditional and egalitarian, situational and ideological responses. *American Journal of Sociology, 85,* 251–260.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics, 29,* 205–220.

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

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–601.

Koehler, K. J. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *Journal of the American Statistical Association, 81,* 336–344.

Krebs, D., & Schuessler, K. (1987). *Soziole Empfindungen* [Social life feelings]. New York: Campus Verlag.

Kreiner, S. (1987). *Collapsibility of multidimensional contingency tables: Theorems, algorithms and programs*. Copenhagen, Denmark: The Danish Institute for Educational Research.

Küsters, U. (1990). A note on sequential ML estimates and their asymptotic covariances. *Statistical Papers, 31,* 131–145.

Lehman, E. L. (1959). *Testing statistical hypotheses*. New York: Wiley.

Ludlow, L. H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement, 10,* 217–229.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43,* 551–560.

Muthén, B. (1988). *LISCOMP*. Mooresville, IN.: Scientific Software.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: a note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19–30.

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

Reiser, M., & Schuessler, K. (1990). A hierarchy for some latent structure models. *Sociological Methods & Research, 19*, 419–465.

Reiser, M., & VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology, 47*, 85–107.

Salomaa, H. (1990). Factor analysis of dichotomous data. Helsinki, Finland: Statistical Society.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589–617.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408.

Tate, M. W., & Hyer, L. A. (1973). Inaccuracy of the chi-squared test of goodness of fit when expected frequencies are small. *Journal of the American Statistical Association, 68*, 836–841.

Tuch, S. A. (1981). Analyzing recent trends in prejudice toward blacks: insights from latent class models. *American Journal of Sociology, 87*, 130–141.

von Eye, A. (1990). *Configural frequency analysis*. London: Cambridge University Press.

Manuscript received 3/4/93
Final version received 3/27/95