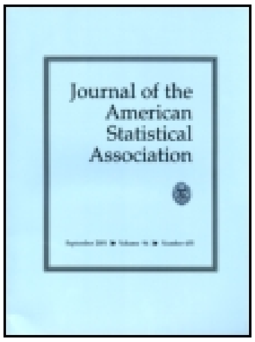

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Generalized Residuals for General Models for Contingency Tables With Application to Item Response Theory

Shelby J. Haberman & Sandip Sinharay

To cite this article: Shelby J. Haberman & Sandip Sinharay (2013) Generalized Residuals for General Models for Contingency Tables With Application to Item Response Theory, Journal of the American Statistical Association, 108:504, 1435-1444, DOI: 10.1080/01621459.2013.835660

To link to this article: https://doi.org/10.1080/01621459.2013.835660

Accepted author version posted online: 25 Aug 2013.
Published online: 19 Dec 2013.

Generalized Residuals for General Models for Contingency Tables With Application to Item Response Theory

Shelby J. HABERMAN and Sandip SINHARAY

Generalized residuals are a tool employed in the analysis of contingency tables to examine possible sources of model error. They have typically been applied to log-linear models and to latent-class models. A general approach to generalized residuals is developed for a very general class of models for contingency tables. To illustrate their use, generalized residuals are applied to models based on item response theory (IRT) models. Such models are commonly applied to analysis of standardized achievement or aptitude tests. To obtain a realistic perspective on application of generalized residuals, actual testing data are employed.

KEY WORDS: Marginal distributions; Marginal estimation; Maximum likelihood; Normal approximations; Rasch model; Three-parameter logistic model; Two-parameter logistic model.

1. INTRODUCTION

Generalized residuals are a tool for examination of sources of error in models for contingency tables. They have been applied to log-linear (Haberman 1976, 1978, 1979) and latent-class models (Haberman 1979, chap. 10). Special cases of generalized residuals have been used for more than 50 years for analysis of two-way and three-way contingency tables (Yates 1948; Cochran 1954, 1955; Armitage 1955; Mantel and Haenszel 1959). Because the Rasch (1960) model commonly used to analyze examinee responses in educational tests corresponds to a log-linear model (Tjur 1982), generalized residuals have been applied to this model in a number of cases (Glas and Verhelst 1989, 1995; Haberman 2004). Nonetheless, generalized residuals have not been applied to more general models for contingency tables, and their application to sparse contingency tables has received relatively limited study. In this article, generalized residuals are developed for very general models for contingency tables. To illustrate their use in a case not covered by the theory previously developed for log-linear models and latent-class models, they are applied to latent-structure models employed in IRT (Birnbaum 1968; Lord 1980; Holland and Rosenbaum 1986; Hambleton, Swaminathan, and Rogers 1991; Junker 1993) in which the underlying latent variable or vector has a normal distribution.

In Section 2, generalized residuals are defined for general models for contingency tables, and their properties are examined. Models considered are comparable in generality to models used in contingency table analysis for general discussions of maximum likelihood estimation (Rao 1973, pp. 359–360), and the class of generalized residuals studied is also more general than has previously been discussed. In addition, an alternative form of generalized residuals is proposed that is readily applied to such sparse tables with relatively easy computations.

To develop the proposed application, a basic discussion of models for item response theory (IRT) is provided in Section 3. In this discussion, some plausible generalized residuals are developed for models for IRT. In Section 4, generalized residuals are applied to data from educational tests. Conclusions and recommendations are provided in Section 5.

2. GENERALIZED RESIDUALS

Generalized residuals are applied to contingency tables (Haberman 1976, 1978, 1979). These residuals use a sample of size $N \geq 1$ to compare a linear combination O_N of observed relative frequencies from a contingency table to an estimated expectation $\hat{E}_N$ of O_N based on a model for the observed contingency table. To obtain an approximate standard normal deviate, one divides the difference $O_N - \hat{E}_N$ by an estimate s_N of its asymptotic standard deviation to obtain the generalized residual

$$z_N = \begin{cases} \dfrac{O_N - \hat{E}_N}{s_N}, & s_N > 0, \\ 0, & s_N = 0, \end{cases} \tag{1}$$

which has an approximate standard normal distribution if the sample size is large and if the model tested is valid. Thus, a large value of $|z_N|$ indicates a failure of the model under study.

Theorem 1, a new theorem proven in the Appendix, provides the basis for application of generalized residuals to general models for contingency tables. It provides both computational formulas for the estimated asymptotic standard deviation s_N and general conditions under which generalized residuals have approximate standard normal distributions if the model under study holds and the sample size is large. In addition,

Shelby J. Haberman, Educational Testing Service, Rosedale Road, Princeton, NJ 08541 (E-mail: *shaberman@ets.org*). Sandip Sinharay, Educational Testing Service, Rosedale Road, Princeton, NJ 08541. We are grateful to Neil Dorans, Hongwen Guo, Andreas Oranje, Frederic Robin, Matthias von Davier, the editor, the associate editor, and the two anonymous reviewers for their helpful advice; to Jill Carey, Behroz Maneckshana, Anthony Giunta, Rui Gao, Kevin Larkin, Aleta Sclan, and Yi-Hsuan Lee for their help with the data; and to Ruth Greenwood for her help with copyediting. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D120006 to Educational Testing Service as part of the Statistical and Research Methodology in Education Initiative. The opinions expressed in this article are those of the authors and do not represent views of the Institute, the U.S. Department of Education, or Educational Testing Service. Note: Sandip Sinharay was at ETS when he performed most of the work for this article. He is currently at CTB/McGraw-Hill (E-mail: *sandip_sinharay@ctb.com*).

Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Theory and Methods
DOI: 10.1080/01621459.2013.835660

an alternative estimated asymptotic standard deviation s_{LN} is provided that can be used in place of s_N in circumstances in which calculation of s_{LN} is more straightforward than is calculation of s_N. In the statement of the theorem, the common convention is followed that primes denote transposes and vectors such as gradients are treated as matrices with one column. An index N for sample size is used to permit more precise discussion of limits. The conditions for the theorem are standard in analysis of contingency tables (Rao 1973, pp. 359–360). The theorem is related to result 2 by Reiser (1996); however, the theorem presented in this article permits weights of relative frequencies to depend on parameters and presents an alternative asymptotic standard deviation s_{LN}. The emphasis on root mean squared errors from weighted least squares is notable in terms of computation of generalized residuals.

Theorem 1. Let q be a positive integer, let $\mathcal{X}$ in R^q be a finite set with at least two elements, and let $\mathbf{X}_i$, $i \geq 1$, be independent and identically distributed random vectors with values in $\mathcal{X}$. Let $p(\mathbf{x})$, $\mathbf{x}$ in $\mathcal{X}$, be the probability that $\mathbf{X}_i = \mathbf{x}$.

Let k be a positive integer, and let the parameter space Ω be a nonempty open set of k-dimensional vectors. Let the probability mappings $\pi(\mathbf{x}; \cdot)$, $\mathbf{x}$ in $\mathcal{X}$, be continuously differentiable functions from the parameter space Ω to the real interval (0, 1) such that

$$\sum_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x}; \boldsymbol{\omega}) = 1, \ \boldsymbol{\omega} \in \Omega. \tag{2}$$

Assume that for some $\boldsymbol{\omega}_0$ in Ω,

$$p(\mathbf{x}) = \pi(\mathbf{x}; \boldsymbol{\omega}), \ \mathbf{x} \in \mathcal{X}, \tag{3}$$

holds for $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. In addition, assume that $\boldsymbol{\omega}_t$, $t \geq 1$, converges to $\boldsymbol{\omega}_0$ whenever $\boldsymbol{\omega}_t$, $t \geq 1$, is a sequence in Ω such that $\pi(\mathbf{x}; \boldsymbol{\omega}_t)$, $t \geq 1$, converges to $\pi(\mathbf{x}; \boldsymbol{\omega}_0)$ for all $\mathbf{x}$ in $\mathcal{X}$.

For each element $\mathbf{x}$ of $\mathcal{X}$, let $\nabla\pi(\mathbf{x}; \cdot)$ be the gradient function of $\pi(\mathbf{x}; \cdot)$, so that for $\boldsymbol{\omega}$ in Ω, $\nabla\pi(\mathbf{x}; \boldsymbol{\omega})$ is the gradient at $\boldsymbol{\omega}$ of $\pi(\mathbf{x}; \cdot)$. For $\boldsymbol{\omega}$ in Ω, let

$$\mathbf{J}(\boldsymbol{\omega}) = \sum_{\mathbf{x}\in\mathbf{X}} [\pi(\mathbf{x}; \boldsymbol{\omega})]^{-1} \nabla\pi(\mathbf{x}; \boldsymbol{\omega})[\nabla\pi(\mathbf{x}; \boldsymbol{\omega})]' \tag{4}$$

denote the Fisher information per observation at $\boldsymbol{\omega}$. Assume that $\mathbf{J}(\boldsymbol{\omega}_0)$ is positive definite.

The log-likelihood per observation ℓ_N is

$$\ell_N(\boldsymbol{\omega}) = N^{-1} \sum_{i=1}^{N} \log \pi(\mathbf{X}_i; \boldsymbol{\omega}), \ \boldsymbol{\omega} \in \Omega. \tag{5}$$

The maximum log-likelihood per observation $\hat{\ell}_N$ is the supremum of ℓ_N.

Let the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\omega}}_N$ of $\boldsymbol{\omega}$ be any function of the $\mathbf{X}_i$, $1 \leq i \leq N$, with range in Ω such that $\ell_N(\hat{\boldsymbol{\omega}}_N) = \hat{\ell}_N$ whenever ℓ_N achieves its supremum on Ω. Let $\hat{\pi}_N(\mathbf{x}) = \pi(\mathbf{x}; \hat{\boldsymbol{\omega}}_N)$ for $\mathbf{x}$ in $\mathcal{X}$.

For each $\mathbf{x}$ in $\mathcal{X}$, let $d(\mathbf{x}; \cdot)$, $\mathbf{x}$ in $\mathcal{X}$, be a real continuous function on Ω, and let $\hat{d}_N(\mathbf{x})$ be its MLE $d(\mathbf{x}; \hat{\boldsymbol{\omega}}_N)$. Let

$$E(\boldsymbol{\omega}) = \sum_{\mathbf{x}\in\mathbf{X}} d(\mathbf{x}; \boldsymbol{\omega})\pi(\mathbf{x}; \boldsymbol{\omega}), \ \boldsymbol{\omega} \in \Omega, \tag{6}$$

be the expectation of $d(\mathbf{X}_i; \boldsymbol{\omega})$, $i \geq 1$, if (3) holds for $\boldsymbol{\omega}$. Let $\hat{E}_N = E(\hat{\boldsymbol{\omega}}_N)$ be the MLE of the expectation $E(\boldsymbol{\omega})$. Let

$$O_N = N^{-1} \sum_{i=1}^{N} \hat{d}_N(\mathbf{X}_i) \tag{7}$$

be the sample mean of the $\hat{d}_N(\mathbf{X}_i)$.

For $\boldsymbol{\omega}$ in Ω, let $\sigma^2(\boldsymbol{\omega})$ be the minimum over $\mathbf{y}$ in R^k of the weighted sum of squares

$$\begin{aligned} S(\mathbf{y}; \boldsymbol{\omega}) = \sum_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x}; \boldsymbol{\omega})\{d(\mathbf{x}, \boldsymbol{\omega}) - E(\boldsymbol{\omega}) \\ - \mathbf{y}'[\pi(\mathbf{x}; \boldsymbol{\omega})]^{-1} \nabla\pi(\mathbf{x}; \boldsymbol{\omega})\}^2. \end{aligned} \tag{8}$$

Let $\hat{\sigma}^2_{LN}$ be the minimum over $\mathbf{y}$ in R^k of the sum of squares

$$S_{LN}(\mathbf{y}) = N^{-1} \sum_{i=1}^{N} \{\hat{d}_N(\mathbf{x}) - O_N - \mathbf{y}'[\hat{\pi}_N(\mathbf{x})]^{-1} \nabla\pi(\mathbf{x}; \hat{\boldsymbol{\omega}}_N)\}^2. \tag{9}$$

Let $s_N = \sigma(\hat{\boldsymbol{\omega}}_N)/N^{1/2}$, and let $s_{LN} = \hat{\sigma}_{LN}/N^{1/2}$. Define the generalized residual z_N as in (1). Let the Louis generalized residual (Louis 1982) be

$$z_{LN} = \begin{cases} (O_N - \hat{E}_N)/s_{LN}, & s_{LN} > 0, \\ 0, & s_{LN} = 0. \end{cases} \tag{10}$$

If $\sigma^2(\boldsymbol{\omega}_0) > 0$, then z_N, $N \geq 1$, and z_{LN}, $N \geq 1$, both converge in distribution to a standard normal random variable.

In many applications, $d(\mathbf{x}, \cdot)$ is a constant $d_0(\mathbf{x})$ for each $\mathbf{x}$ in $\mathcal{X}$, so that O_N is the observed average $N^{-1}\sum_{i=1}^{N} d_0(\mathbf{X}_i)$ for any value of the MLE $\hat{\boldsymbol{\omega}}_N$. For this case for the generalized residual z_N, this result has been obtained with slightly less rigorous conditions (Reiser 1996).

It is always possible to obtain an alternative parameterization of the model that leads to simplified formulas. Consider the following theorem proven in the Appendix.

Theorem 2. In Theorem 1, for each $\mathbf{x}$ in $\mathcal{X}$ let $\lambda(\mathbf{x}; \cdot)$ on Ω be a continuously differentiable function with gradient function $\nabla\lambda(\mathbf{x}; \cdot)$. For $\boldsymbol{\omega}$ in Ω and $\mathbf{x}$ in $\mathcal{X}$, let

$$\pi(\mathbf{x}; \boldsymbol{\omega}) = [\gamma(\boldsymbol{\omega})]^{-1} \exp[\lambda(\mathbf{x}; \boldsymbol{\omega})], \tag{11}$$

where

$$\gamma(\boldsymbol{\omega}) = \sum_{\mathbf{x}\in\mathcal{X}} \exp[\lambda(\mathbf{x}; \boldsymbol{\omega})]. \tag{12}$$

Let

$$\mathbf{m}(\boldsymbol{\omega}) = \sum_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x}; \boldsymbol{\omega})\nabla\lambda(\mathbf{x}; \boldsymbol{\omega}), \tag{13}$$

so that $\mathbf{m}(\boldsymbol{\omega})$ is the expectation of $\nabla\lambda(\mathbf{X}_i; \boldsymbol{\omega})$, $i \geq 1$, if (3) holds. Then for $\boldsymbol{\omega}$ in Ω, the Fisher information is

$$\mathbf{J}(\boldsymbol{\omega}) = \sum_{\mathbf{x}\in\mathbf{X}} \pi(\mathbf{x}; \boldsymbol{\omega})[\nabla\lambda(\mathbf{x}; \boldsymbol{\omega}) - \mathbf{m}(\boldsymbol{\omega})][\nabla\lambda(\mathbf{x}; \boldsymbol{\omega}) - \mathbf{m}(\boldsymbol{\omega})]', \tag{14}$$

so that $\mathbf{J}(\boldsymbol{\omega}_0)$ is the covariance matrix of $\nabla\lambda(\mathbf{X}_i; \boldsymbol{\omega}_0)$ if (3) holds for $\boldsymbol{\omega}$ in Ω. Let $\mathbf{0}_k$ be the k-dimensional vector with all elements 0. Then the Fisher information at $\boldsymbol{\omega}$ in Ω is positive definite if no k-dimensional vector $\mathbf{u}$ other than $\mathbf{0}_k$ satisfies the constraint that $\mathbf{u}'\nabla\lambda(\mathbf{x}, \boldsymbol{\omega})$ is constant for all $\mathbf{x}$ in $\mathcal{X}$.

The log-likelihood per observation ℓ_N satisfies

$$\ell_N(\boldsymbol{\omega}) = -\log\gamma(\boldsymbol{\omega}) + N^{-1}\sum_{i=1}^{N}\lambda(\mathbf{X}_i;\boldsymbol{\omega}),\ \ \boldsymbol{\omega}\in\Omega. \tag{15}$$

For $\boldsymbol{\omega}$ in Ω, the weighted sum of squares

$$S(a,\mathbf{y};\boldsymbol{\omega}) = \sum_{\mathbf{x}\in\mathcal{X}}\pi(\boldsymbol{\omega})[d(\mathbf{x};\boldsymbol{\omega}) - a - \mathbf{y}'\nabla\lambda(\mathbf{x};\boldsymbol{\omega})]^2, \tag{16}$$

a real, $\mathbf{y}$ in R^k, has minimum $\sigma^2(\boldsymbol{\omega})$, and $\sigma^2(\boldsymbol{\omega}) > 0$ if, and only if, no $\mathbf{u}$ in R^k exists such that $d(\mathbf{x};\boldsymbol{\omega}) - \mathbf{u}'\nabla\lambda(\mathbf{x};\boldsymbol{\omega})$ is constant for all $\mathbf{x}$ in $\mathcal{X}$.

If $\ell_N(\hat{\boldsymbol{\omega}}_N)) = \hat{\ell}_N$, then $\hat{\sigma}^2_{LN}$ is the minimum of

$$S_{LN}(a,\mathbf{y}) = N^{-1}\sum_{i=1}^{N}[\hat{d}_N(\mathbf{X}_i) - a - \mathbf{y}'\nabla\lambda(\mathbf{X}_i;\hat{\boldsymbol{\omega}}_N)]^2, \tag{17}$$

a real, $\mathbf{y}$ in R^k, and $\hat{\sigma}^2_{LN} > 0$ if, and only if, no $\mathbf{u}$ in R^k exists such that $\hat{d}_N(\mathbf{X}_i) - \mathbf{u}'\nabla\lambda(\mathbf{X}_i;\hat{\boldsymbol{\omega}}_N)$ is constant for $1 \le i \le N$.

In Theorem 2, the formulas for $\sigma^2(\boldsymbol{\omega})$ and $\hat{\sigma}^2_{LN}$ have the advantage that they are routine formulas from regression analysis that do not require computation of $\mathbf{m}(\boldsymbol{\omega})$ or $\mathbf{m}(\hat{\boldsymbol{\omega}}_N)$.

It is always possible to let $\lambda(\mathbf{x}, \cdot)$ be $\log\pi(\mathbf{x};\cdot)$ for all $\mathbf{x}$ in $\mathcal{X}$; however, alternative definitions of the functions $\lambda(\mathbf{x};\cdot)$ are often more convenient. A simple application of (11) arises in log-linear models. For $\mathbf{x}$ in $\mathcal{X}$, let $\mathbf{T}(\mathbf{x})$ be a k-dimensional vector, let Ω be R^k, and let

$$\lambda(\mathbf{x};\boldsymbol{\omega}) = \boldsymbol{\omega}'\mathbf{T}(\mathbf{x}),\ \ \mathbf{x}\in\mathcal{X}, \boldsymbol{\omega}\in R^k. \tag{18}$$

Then the resulting model is a log-linear model, and $\nabla\lambda(\mathbf{x};\boldsymbol{\omega})$ is $\mathbf{T}(\mathbf{x})$ for $\mathbf{x}$ in $\mathcal{X}$. In this case, if no k-dimensional vector $\mathbf{u} \ne \mathbf{0}_k$ exists such that $\mathbf{u}'\mathbf{T}(\mathbf{x})$ is constant over $\mathbf{x}$ in $\mathcal{X}$, then a unique $\boldsymbol{\omega}_0$ in Ω exists such that (3) holds for $\boldsymbol{\omega}$ equal to $\boldsymbol{\omega}_0$. In addition, $\boldsymbol{\omega}_t$, $t \ge 1$, converges to $\boldsymbol{\omega}_0$ whenever $\boldsymbol{\omega}_t$, $t \ge 1$, is a sequence in Ω such that $\pi(\mathbf{x};\boldsymbol{\omega}_t)$, $t \ge 1$, converges to $\pi(\mathbf{x};\boldsymbol{\omega}_0)$ for all $\mathbf{x}$ in $\mathcal{X}$. The Fisher information matrix $\mathbf{J}_n(\boldsymbol{\omega})$ is positive definite for all $\boldsymbol{\omega}$ in Ω, and $\sigma^2(\boldsymbol{\omega})$ is positive for all $\boldsymbol{\omega}$ in Ω if no $\mathbf{u}$ in R^k exists such that $d(\mathbf{x}) - \mathbf{u}'\mathbf{T}(\mathbf{x})$ is constant for $\mathbf{x}$ in $\mathcal{X}$.

It should be noted that O_N and $\hat{E}_N$ have value in measuring model discrepancy when the model assessed is not valid. The difference $O_N - \hat{E}_N$ measures the deviation between an observed average and an average fitted under the model. This difference typically has a limit in large samples. The result depends on the expected log-likelihood per observation,

$$E(\ell_N(\boldsymbol{\omega})) = \sum_{\mathbf{x}\in\mathcal{X}} p(\mathbf{x})\log\pi(\mathbf{x};\boldsymbol{\omega}),\ \ \boldsymbol{\omega}\in\Omega. \tag{19}$$

Given (19), consider the following result.

Theorem 3. In Theorem 1, instead of the assumptions that (3) holds for $\boldsymbol{\omega} = \boldsymbol{\omega}_0$ and $\mathbf{J}(\boldsymbol{\omega}_0)$ is positive definite, only assume that a $\boldsymbol{\omega}_0$ in Ω exists such that

$$E(\ell_N(\boldsymbol{\omega}_0)) \ge E(\ell_N(\boldsymbol{\omega})),\ \ \boldsymbol{\omega}\in\Omega. \tag{20}$$

In addition, assume that any sequence $\boldsymbol{\omega}_t$ in Ω, $t \ge 1$ converges to $\boldsymbol{\omega}_0$ whenever $E(\ell_N(\boldsymbol{\omega}_t))$, $t \ge 1$, converges to $E(\ell_N(\boldsymbol{\omega}_0))$. Then O_N converges with probability 1 to $\sum_{\mathbf{x}\in\mathcal{X}} d(\mathbf{x},\boldsymbol{\omega}_0)p(\mathbf{x})$, and $\hat{E}_N$ converges with probability 1 to $E(\boldsymbol{\omega}_0)$. If (3) holds for $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, then $E(\boldsymbol{\omega}_0) = \sum_{\mathbf{x}\in\mathcal{X}} d(\mathbf{x};\boldsymbol{\omega}_0)p(\mathbf{x})$.

Proof. By Berk (1972), $\hat{\boldsymbol{\omega}}_N$ converges to $\boldsymbol{\omega}_0$ with probability 1. Given the strong law of large numbers, O_N converges with probability 1 to $\sum_{\mathbf{x}\in\mathcal{X}} d(\mathbf{x};\boldsymbol{\omega}_0)p(\mathbf{x})$. Because E is a continuous function, $\hat{E}_N$ converges to $E(\boldsymbol{\omega}_0)$ with probability 1. The final conclusion follows from (6). □

In applications such as IRT, the number of elements in the set $\mathcal{X}$ can be large compared to the sample size N even though N is quite large. This problem is treated in the literature on contingency tables in the context of log-linear models and latent-class models (Haberman 1977a, b, 1988). Although a thorough treatment of the issue is beyond the scope of the current article, several general points can be made based on the available literature. It is certainly true that normal approximations for MLEs and generalized residuals can hold for quite large sets $\mathcal{X}$. Typically the dimension k of the parameter space must be small relative to the sample size N. This criterion is much more important in practice than the size of $\mathcal{X}$. A second requirement is specific to generalized residuals (Haberman 1978, pp. 342–343). The ratio

$$\frac{\sum_{\mathbf{x}\in\mathcal{X}} |e(\mathbf{x},\mathbf{y}(\boldsymbol{\omega}_0);\boldsymbol{\omega}_0)|^3}{N^{1/2}[\sigma(\boldsymbol{\omega}_0)]^3}$$

must be small for a normal approximation to apply to the generalized residuals z_N and z_{LN}.

3. MODELS FOR IRT

IRT is commonly employed by psychometricians to analyze educational tests; however, its use also extends to medical diagnosis and psychiatric epidemiology (Eaton and Bohrnstedt 1989), multiple recapture methods for estimating population sizes (Chao 1987), and systems reliability and population genetics (Holland and Rosenbaum 1986). For recent discussions on IRT models in the statistical literature, see articles by Cohen and Jiang (1999), Scott and Ip (2002), Rice (2004), Baldwin, Bernstein, and Wainer (2009), and Chang and Ying (2009). For a history of the development of IRT, see the article by Bock (1997).

IRT can be regarded as a special case of latent structure analysis. In an IRT model, the random vectors $\mathbf{X}_i$ have elements X_{ij}, $1 \le j \le q$, with integer values from 0 to $r_j - 1$, where the integer r_j is greater than 1. Thus, $\mathcal{X}$ consists of all q-dimensional vectors $\mathbf{x}$ with integer elements x_j, $0 \le x_j \le r_j - 1$, $1 \le j \le q$. In educational testing, r_j typically represents the number of scores for one item of a test, so that $r_j = 2$ may be employed with an item that is scored correct ($X_{ij} = 1$) or not correct ($X_{ij} = 0$). In an IRT model, it is assumed that the X_{ij}, $1 \le j \le q$, are conditionally (locally) independent given a latent random vector $\boldsymbol{\theta}_i$ of dimension $H \ge 1$ with range Θ. The vector $\boldsymbol{\theta}_i$ is often regarded in educational testing as representing examinee proficiency, but its use does not require such an interpretation. The latent vector may be polytomous, in which case Θ is finite, or continuous, in which case Θ has a nonempty interior. If Θ is finite, then the IRT model can be regarded as a latent-class model (Heinen 1996; Haberman 2005). In this article, all models considered have $\Theta = R^H$ and assume that $\boldsymbol{\theta}_i$, $i \ge 1$, has a continuous distribution. Thus,

$$\pi(\mathbf{x};\boldsymbol{\omega}) = \int \pi(\mathbf{x}|\boldsymbol{\theta};\boldsymbol{\omega})g(\boldsymbol{\theta};\boldsymbol{\omega})d\boldsymbol{\theta}, \tag{21}$$

where the conditional independence assumption implies that $\pi(\mathbf{x}|\boldsymbol{\theta};\boldsymbol{\omega})$, $\mathbf{x}$ in $\mathbf{X}$, satisfies the product rule

$$\pi(\mathbf{x}|\boldsymbol{\theta};\boldsymbol{\omega}) = \prod_{j=1}^{q} \pi_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega}). \tag{22}$$

Here for $1 \le j \le q$, $0 \le x + j < r_j$, and $\boldsymbol{\theta}$ in Θ, the $\pi_j(x_j|\boldsymbol{\theta};\cdot)$ are positive continuously differentiable functions on Ω. For $\boldsymbol{\theta}$ in Θ and $1 \le j \le q$,

$$\sum_{x_j=0}^{r_j-1} \pi_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega}) = 1. \tag{23}$$

If (21) and (22) hold for $\boldsymbol{\omega}$ in Ω, then $\pi_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega})$ is the conditional probability that $X_{ij} = x_j$ given $\boldsymbol{\theta}_i = \boldsymbol{\theta}$ for $\boldsymbol{\theta}$ in Θ. One may term $\pi_j(x_j|\cdot;\boldsymbol{\omega})$ the item characteristic function for item j and item score x_j. In addition, it may be assumed that for $1 \le j \le q$, $0 \le x_j < r_j$, and $\boldsymbol{\omega}$ in Ω, $\pi_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega})$ is continuous in $\boldsymbol{\theta}$. If (3) holds for some $\boldsymbol{\omega}_0$ in Ω, then $\pi_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega}_0)$ represents the conditional probability that $X_{ij} = x_j$, $i \ge 1$, given $\boldsymbol{\theta}_i = \boldsymbol{\theta}$. The parameterization

$$p_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega}) = [\gamma_j(\boldsymbol{\theta};\boldsymbol{\omega})]^{-1}\exp[\lambda_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega})] \tag{24}$$

may be used, where $\lambda_j(x_j|\boldsymbol{\theta};\cdot)$ is continuously differentiable for each $\boldsymbol{\theta}$ in Θ, $\lambda_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega})$ is continuous in $\boldsymbol{\theta}$ for each $\boldsymbol{\omega}$ in Ω, and

$$\gamma_j(\boldsymbol{\theta};\boldsymbol{\omega}) = \sum_{x_j=0}^{r_j-1} \exp[\lambda_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega})]. \tag{25}$$

If $r_j = 2$ and $\lambda_j(0|\boldsymbol{\theta};\boldsymbol{\omega}) = 0$, then $\lambda_j(1|\boldsymbol{\theta};\boldsymbol{\omega})$ is the item logit function

$$\log[\pi_j(1|\boldsymbol{\theta};\boldsymbol{\omega})/\pi_j(0|\boldsymbol{\theta};\boldsymbol{\omega})]$$

(Holland 1990). For $0 \le x_j < r_j$ and $\boldsymbol{\theta}$ in Θ, if $\nabla\lambda(x_j|\boldsymbol{\theta};\boldsymbol{\omega})$ is the gradient of $\lambda(x_j|\boldsymbol{\theta};\cdot)$ at $\boldsymbol{\omega}$ and

$$\mathbf{m}_j(\boldsymbol{\theta};\boldsymbol{\omega}) = \sum_{x_j=0}^{r_j-1} \pi_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega})\nabla\lambda_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega}), \tag{26}$$

then $\log \pi_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega})$ has gradient

$$\nabla \log \pi_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega}) = \nabla\lambda(x_j|\boldsymbol{\theta};\boldsymbol{\omega}) - m_j(\boldsymbol{\theta};\boldsymbol{\omega}) \tag{27}$$

at $\boldsymbol{\omega}$.

3.1 Examples of Item Characteristic Functions

For a few illustrations of item characteristic functions, consider the case of $H = 1$ and $r_j = 2$ for $1 \le j \le q$. Let θ be used instead of $\boldsymbol{\theta}$ because $\boldsymbol{\theta}$ has only one element θ_1. In the two-parameter logistic (2PL) model for $H = 1$, each $r_j = 2$, Ω is the subset of R^{2q} that consists of $2q$-dimensional vectors $\boldsymbol{\omega}$ with elements $\omega_{q+j} > 0$ for $1 \le j \le q$, and

$$\lambda_j(x_j|\theta;\boldsymbol{\omega}) = x_j(\omega_{q+j}\theta + \omega_j) \tag{28}$$

for real θ, $0 \le x_j \le 1$, and $1 \le j \le q$. It is customary to describe $a_j = \omega_{q+j}$ as the item discrimination and $b_j = -\omega_j/\omega_{q+j}$ as the item difficulty (e.g., Hambleton et al. 1991, chap. 1). In the corresponding one-parameter logistic (1PL) model for $H = 1$, Ω is the subset of R^{q+1} that consists of $q + 1$-dimensional vectors $\boldsymbol{\omega}$ with element $\omega_{q+1} > 0$. In this case,

$$\lambda_j(x_j|\theta;\boldsymbol{\omega}) = x_j(\omega_{q+1}\theta + \omega_j) \tag{29}$$

for real θ, $0 \le x_j \le 1$, and $1 \le j \le q$ (Rasch 1960). The item discrimination $a = \omega_{q+1}$ and the item difficulty is $b_j = -\omega_j/\omega_{q+1}$. Thus, the 1PL model corresponds to a 2PL model in which ω_{q+j} is constant for $1 \le j \le q$.

In a three-parameter logistic (3PL) model (Birnbaum 1968), Ω is the subset of R^{3q} that consists of $3q$-dimensional vectors $\boldsymbol{\omega}$ with elements $\omega_{q+j} > 0$ for $1 \le j \le q$, and

$$\begin{aligned}\lambda_j(x_j|\theta;\boldsymbol{\omega}) = x_j \log[&\exp(\omega_{2q+j}) + \exp(\omega_{q+j}\theta + \omega_j) \\ &+ \exp(\omega_{2q+j} + \omega_{q+j}\theta + \omega_j)]\end{aligned} \tag{30}$$

for real θ, $0 \le x_j \le 1$, and $1 \le j \le q$. Here $c_j = \exp(\omega_{2q+j})/[1 + \exp(\omega_{2q+j})]$ is a guessing parameter intended to treat the familiar issue that examinees, especially low-scoring ones, often guess the answer in multiple-choice tests (see, e.g., Lord 1980, p. 17). For item discrimination $a_j = \omega_{q+j}$ and item difficulty $b_j = -\omega_j/\omega_{q+j}$,

$$p_j(1|\theta;\boldsymbol{\omega}) = c_j + (1 - c_j)\frac{\exp[a_j(\theta - b_j)]}{1 + \exp[a_j(\theta - b_j)]} \tag{31}$$

for real θ (Hambleton et al. 1991, chap. 1). For fixed ω_j, $1 \le j \le 2q$, the 3PL model converges to the 2PL model as each ω_{2q+j}, $1 \le j \le q$, approaches $-\infty$. In a 3PL model with a constant guessing parameter, Ω is the subset of R^{2q+1} that consists of $2q + 1$-dimensional vectors $\boldsymbol{\omega}$ with elements $\omega_{q+j} > 0$ for $1 \le j \le q$, and

$$\begin{aligned}\lambda_j(x_j|\theta;\boldsymbol{\omega}) = x_j \log[&\exp(\omega_{2q+1}) + \exp(\omega_{q+j}\theta + \omega_j) \\ &+ \exp(\omega_{2q+1} + \omega_{q+j}\theta + \omega_j)]\end{aligned} \tag{32}$$

for real θ, $0 \le x_j \le 1$, and $1 \le j \le q$. If (30) holds and ω_{2q+j} is constant for $1 \le j \le q$, so that the guessing parameter c_j is constant, then (32) holds. One may also consider a 3PL model with a constant guessing parameter and constant item discrimination. Here Ω is the subset of R^{q+2} that consists of $q + 2$-dimensional vectors $\boldsymbol{\omega}$ with element ω_{q+1} positive. In this case,

$$\begin{aligned}\lambda_j(x_j|\theta;\boldsymbol{\omega}) = x_j \log[&\exp(\omega_{q+2}) + \exp(\omega_{q+1}\theta + \omega_j) \\ &+ \exp(\omega_{q+2} + \omega_{q+1}\theta + \omega_j)]\end{aligned} \tag{33}$$

for real θ. Numerous other examples of models for IRT are available for r_j equal or greater than 2 and for H (Adams, Wilson, and Wang 1997; Muraki 1997; Reckase 1997; Bock and Moustaki 2007).

3.2 Density Functions

The density function $g(\boldsymbol{\theta};\cdot)$, $\boldsymbol{\theta}$ in Θ, is a positive continuously differentiable function on Ω. For $\boldsymbol{\omega}$ in Ω, $g(\cdot;\boldsymbol{\omega})$ is a continuous function with finite integral

$$\int g(\boldsymbol{\theta};\boldsymbol{\omega})d\boldsymbol{\theta} = 1. \tag{34}$$

In many common applications with $H = 1$, $g(\theta;\boldsymbol{\omega})$ is always equal to the density $\phi(\theta)$ at θ of the standard normal distribution. This situation applies to the examples in Section 4. It is quite common for $g(\cdot;\boldsymbol{\omega})$ to be a density function for a normal distribution for all $\boldsymbol{\omega}$ in Ω.

If (3) holds for some $\boldsymbol{\omega}_0$ in Ω, then $g(\cdot;\boldsymbol{\omega}_0)$ represents the probability density function of $\boldsymbol{\theta}_i$, $i \geq 1$. One may select a function $\mu(\boldsymbol{\theta};\boldsymbol{\omega})$ for $\boldsymbol{\theta}$ in Θ and $\boldsymbol{\omega}$ in Ω such that $\mu(\boldsymbol{\theta},\cdot)$ is continuously differentiable for $\boldsymbol{\theta}$ in Θ, $\mu(\cdot;\boldsymbol{\omega})$ is continuous for $\boldsymbol{\omega}$ in Ω, $\exp[\mu(\cdot;\boldsymbol{\omega})]$ is integrable for $\boldsymbol{\omega}$ in Ω,

$$g(\boldsymbol{\theta};\boldsymbol{\omega}) = [\delta(\boldsymbol{\omega})]^{-1}\exp[\mu(\boldsymbol{\theta};\boldsymbol{\omega})], \tag{35}$$

and

$$\delta(\boldsymbol{\omega}) = \int \exp[\mu(\boldsymbol{\theta};\boldsymbol{\omega})]d\boldsymbol{\theta}. \tag{36}$$

In applications in which $\boldsymbol{\theta}_i$ is assumed to be normally distributed, $\mu(\boldsymbol{\theta};\boldsymbol{\omega})$ may be defined to be quadratic in $\boldsymbol{\theta}$ for fixed $\boldsymbol{\omega}$ and linear in $\boldsymbol{\omega}$ for fixed $\boldsymbol{\theta}$ (Ziegler 2011, chap. 2). The gradient of $\log g(\boldsymbol{\theta};\cdot)$ can be expressed in terms of the gradient of $\mu(\boldsymbol{\theta};\boldsymbol{\omega})$ provided that the usual conditions of analysis hold for the interchange of integration and differentiation. Consider the following theorem proven in the Appendix.

Theorem 4. For $\boldsymbol{\theta}$ in Θ and $\boldsymbol{\omega}$ in Ω, let $\nabla\mu(\boldsymbol{\theta};\boldsymbol{\omega})$ be the gradient of $\mu(\boldsymbol{\theta};\cdot)$ at $\boldsymbol{\omega}$. For $\mathbf{u}$ in R^k, let $|\mathbf{u}| = (\mathbf{u}'\mathbf{u})^{1/2}$. For each $\boldsymbol{\omega}$ in Ω, assume that there is a positive real continuous function κ on Θ and positive real ϵ and ϵ_1 such that $\kappa g(\cdot;\boldsymbol{\omega})$ is integrable and such that $\boldsymbol{\omega}_1$ is in Ω and

$$|\nabla\mu(\boldsymbol{\theta};\boldsymbol{\omega}_1)|\exp[\epsilon|\nabla\mu(\boldsymbol{\theta};\boldsymbol{\omega}_1)|] \leq \kappa(\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta, \tag{37}$$

whenever $|\boldsymbol{\omega}_1 - \boldsymbol{\omega}| \leq \epsilon_1$. Then the integral

$$\mathbf{m}_0(\boldsymbol{\omega}) = \int g(\boldsymbol{\theta};\boldsymbol{\omega})\nabla\mu(\boldsymbol{\theta};\boldsymbol{\omega})d\boldsymbol{\theta} \tag{38}$$

is finite and $\log g(\boldsymbol{\theta};\boldsymbol{\omega})$ has gradient

$$\nabla\log g(\boldsymbol{\theta};\boldsymbol{\omega}) = \nabla\mu(\boldsymbol{\theta};\boldsymbol{\omega}) - \mathbf{m}_0(\boldsymbol{\omega}) \tag{39}$$

at $\boldsymbol{\omega}$.

Theorem 4 applies to all examples considered in this article.

To obtain the gradient of $\log\pi(\mathbf{x};\boldsymbol{\omega})$, $\mathbf{x}$ in $\mathcal{X}$, and the Fisher information $\mathbf{J}(\boldsymbol{\omega})$ for $\boldsymbol{\omega}$ in Ω is straightforward, provided that conditions hold for interchange of differentiation and integration. The following theorem applies. The argument is virtually the same as for Theorem 4, so that no proof is provided.

Theorem 5. Let the conditions of Theorem 4 hold. Let $\mathbf{x}$ be in $\mathcal{X}$. For $\boldsymbol{\theta}$ in Θ, let

$$g(\boldsymbol{\theta}|\mathbf{x};\boldsymbol{\omega}) = \frac{\pi(\mathbf{x}|\boldsymbol{\theta};\boldsymbol{\omega})g(\boldsymbol{\theta};\boldsymbol{\omega})}{\pi(\mathbf{x};\boldsymbol{\omega})}, \tag{40}$$

so that the posterior density of $\boldsymbol{\theta}_i$ given $\mathbf{X}_i = \mathbf{x}$ is $g(\boldsymbol{\theta}|\mathbf{x};\boldsymbol{\omega})$, $\boldsymbol{\theta}$ in Θ, if (3) holds. For $\boldsymbol{\theta}$ in Θ and $\boldsymbol{\omega}$ in Ω, the gradient at $\boldsymbol{\omega}$ of the logarithm of $\pi(\mathbf{x}|\boldsymbol{\theta};\cdot)$ is

$$\nabla\log\pi(\mathbf{x}|\boldsymbol{\theta};\boldsymbol{\omega}) = \sum_{j=1}^{q}[\lambda_j(x_j|\boldsymbol{\theta};\boldsymbol{\omega}) - \mathbf{m}_j(\boldsymbol{\theta};\boldsymbol{\omega})]. \tag{41}$$

For each $\boldsymbol{\omega}$ in Ω, assume that there is a positive continuous function κ on Θ and an $\epsilon > 0$ such that $\kappa g(\cdot;\boldsymbol{\omega})$ is integrable and such that $\boldsymbol{\omega}_1$ is in Ω and

$$\begin{aligned}|\nabla\log\pi(\mathbf{x}|\boldsymbol{\theta};\boldsymbol{\omega}_1) + \nabla\log g(\boldsymbol{\theta};\boldsymbol{\omega}_1)|\exp[\epsilon|\nabla\log\pi(\mathbf{x}|\boldsymbol{\theta};\boldsymbol{\omega}_1)\\ + \nabla\log g(\boldsymbol{\theta};\boldsymbol{\omega}_1)|] \leq \kappa(\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta,\end{aligned} \tag{42}$$

whenever $\boldsymbol{\omega}_1$ is in R^k and $|\boldsymbol{\omega}_1 - \boldsymbol{\omega}| < \epsilon$.

Then at $\boldsymbol{\omega}$, $\log\pi(\mathbf{x};\cdot)$ has gradient

$$\begin{aligned}\nabla\log\pi(\mathbf{x};\boldsymbol{\omega}) = \int g(\boldsymbol{\theta}|\mathbf{x};\boldsymbol{\omega})[\nabla\log\pi(\mathbf{x}|\boldsymbol{\theta};\boldsymbol{\omega})\\ + \nabla\log g(\boldsymbol{\theta};\boldsymbol{\omega})]d\boldsymbol{\theta}.\end{aligned} \tag{43}$$

Because a posterior density is used in (43), adaptive quadrature is typically efficient in numerical computations (Naylor 1982; Haberman, von Davier, and Lee 2008).

Computation of the MLE $\hat{\boldsymbol{\omega}}_N$ has been extensively explored, and a variety of computational algorithms have been implemented (Bock and Aitkin 1981; Muraki 1997; Muraki and Bock 2003; Haberman et al. 2008; Cai, du Toit, and Thissen 2011) based either on the EM algorithm (Dempster, Laird, and Rubin 1977) or on the stabilized Newton-Raphson algorithm (Haberman 1988). Within the literature on IRT, the term maximum marginal likelihood is often encountered to refer to maximum-likelihood estimation of the type considered in this article. Computations of MLEs are routinely performed in large-scale testing programs such as TOEFL, GRE, Praxis, SAT, and Grades 3–8 of the New York State Testing Program. Calculations in this article were performed with MIRT, a computer program with executable code freely available for noncommercial use at *mirt@ets.org*.

3.3 Generalized Residuals in IRT

In the case of IRT models, generalized residuals have received limited attention except in cases in which an IRT model can be shown to be a special case of a log-linear model, as is true in the 1PL model (Tjur 1982; Glas and Verhelst 1989, 1995; Haberman 2004, 2009). Traditional residual analysis (Hambleton et al. 1991, p. 60) for $\pi_j(x_j;\boldsymbol{\theta};\boldsymbol{\omega})$ provided in commercial software such as PARSCALE (Muraki and Bock 2003) is not based on generalized residuals and does not result in residuals with an appropriate standard normal approximation (Haberman, Sinharay, and Chon 2013).

In the case of z_N, the case of a residual for $p(\mathbf{x}_1)$, $\mathbf{x}_1$ in $\mathcal{X}$, has been considered (Reiser 1996). In this case, for $\mathbf{x}$ in $\mathcal{X}$ and $\boldsymbol{\omega}$ in Ω,

$$d(\mathbf{x};\boldsymbol{\omega}) = \begin{cases}1, & \mathbf{x} = \mathbf{x}_1,\\ 0, & \mathbf{x} \neq \mathbf{x}_1,\end{cases} \tag{44}$$

and $E(\boldsymbol{\omega}) = \pi(\mathbf{x}_1;\boldsymbol{\omega})$; however, this case does not lead to a generalized residual with a satisfactory normal approximation if the number $\prod_{j=1}^{q} r_j$ of elements of $\mathcal{X}$ is large. An alternative residual has been considered for the marginal probability that $X_{ij} = x_{1j}$ and $X_{ij_1} = x_{1j_1}$ for $1 \leq j < j_1 \leq q$, $0 \leq x_{1j} < r_j$, and $0 \leq x_{1j_1} < r_{j_1}$. Here for $\mathbf{x}$ in $\mathbf{X}$ and $\boldsymbol{\omega}$ in Ω,

$$d(\mathbf{x};\boldsymbol{\omega}) = \begin{cases}1, & x_j = x_{1j} \text{ and } x_{j_1} = x_{1j_1},\\ 0, & x_j \neq x_{1j} \text{ or } x_{j_1} \neq x_{1j_1}\end{cases} \tag{45}$$

(Reiser 1996). Thus, O_N is the fraction of i, $1 \leq i \leq N$, such that $X_{ij} = x_{1j}$ and $X_{ij_1} = x_{1j_1}$ and

$$E(\boldsymbol{\omega}) = \int p_j(x_{1j}|\boldsymbol{\theta};\boldsymbol{\omega})p_{j_1}(x_{1j_1}|\boldsymbol{\theta};\boldsymbol{\omega})g(\boldsymbol{\theta};\boldsymbol{\omega})d\boldsymbol{\theta}. \tag{46}$$

Residuals of this type are typically employed to detect violations of the conditional independence assumption. Their use is predicated on the fact that in typical models such as the 1PL, 2PL, or 3PL models, the fraction of i, $1 \leq i \leq N$, with $X_{ij} = x_j$, $0 \leq x_j \leq r_j - 1$, $1 \leq j \leq q$, is very close to the estimate $p_j(x_j; \hat{\boldsymbol{\omega}}_N)$.

Use of the Louis generalized residuals z_{LN} appears more difficult to find. One exception involves comparison of a 2PL model to a 3PL model (Haberman 2006). Here z_{LN} is used with

$$d(\mathbf{x}; \boldsymbol{\omega}) = \int [p_j(1|\boldsymbol{\theta}; \boldsymbol{\omega})]^{-1}[x_j - p_j(1|\boldsymbol{\theta}; \boldsymbol{\omega})]g(\boldsymbol{\theta}|\mathbf{x}; \boldsymbol{\omega}), \quad (47)$$

$\mathbf{x}$ in $\mathcal{X}$, $\boldsymbol{\omega}$ in Ω, and $1 \leq j \leq q$. Under (3), $d(\mathbf{X}_i; \boldsymbol{\omega})$ has expectation $E(\boldsymbol{\omega}) = 0$.

Some analyses of fit of IRT models have involved use of the sum scores $S_i = \sum_{j=1}^{q} X_{ij}$, $i \geq 1$ to construct tests with approximate chi-square distributions (Orlando and Thissen 2000). The sum scores can also be used to construct generalized residuals. For example, for an integer s from 0 to $\sum_{j=1}^{q}(r_j - 1)$, one may let

$$d(\mathbf{x}; \boldsymbol{\omega}) = \begin{cases} 1, & \sum_{j=1}^{q} x_j = s, \\ 0, & \sum_{j=1}^{q} x_j \neq s, \end{cases} \quad (48)$$

for $\mathbf{x}$ in $\mathcal{X}$ and $\boldsymbol{\omega}$ in Ω. Here O_N is then the fraction of integers i, $1 \leq i \leq N$, such that $S_i = s$, and $\hat{E}_N$ is the MLE $\hat{p}_S(s)$ of the probability that $S_i = s$. Computation of $\hat{E}_N$ involves some complication due to calculation of a sum of independent integer-valued random variables that do not have the same distribution. Appropriate recursive algorithms nonetheless can be found for efficient computation of the conditional probability $p_S(s|\boldsymbol{\theta}; \boldsymbol{\omega})$ that $S_i = s$ given that $\boldsymbol{\theta}_i = \boldsymbol{\theta}$, $\boldsymbol{\omega}$ in Ω, $\boldsymbol{\theta}$ in Θ (Lord and Wingersky 1984; Thissen et al. 1995). It then follows that

$$E(\boldsymbol{\omega}) = \int p_S(s|\boldsymbol{\theta}; \boldsymbol{\omega})g(\boldsymbol{\theta}; \boldsymbol{\omega})d\boldsymbol{\theta}. \quad (49)$$

Alternatively, cumulative distribution functions can be examined for $s < q$ with

$$d(\mathbf{x}; \boldsymbol{\omega}) = \begin{cases} 1, & \sum_{j=1}^{q} x_j \leq s, \\ 0, & \sum_{j=1}^{q} x_j > s, \end{cases} \quad (50)$$

and

$$E(\boldsymbol{\omega}) = \sum_{s_1=0}^{s} \int p_S(s_1|\boldsymbol{\theta}; \boldsymbol{\omega})g(\boldsymbol{\theta}; \boldsymbol{\omega})d\boldsymbol{\theta}. \quad (51)$$

Here O_N is the empirical distribution function at s of the sum scores S_i, $1 \leq i \leq N$.

3.4 Louis Generalized Residuals and IRT

In many cases in IRT, computations are somewhat easier for the same level of accuracy and inferences are more straightforward if the Louis generalized residual corresponding to $d(\mathbf{x}; \boldsymbol{\omega})$, $\mathbf{x}$ in $\mathcal{X}$, $\boldsymbol{\omega}$ in Ω, is replaced by the Louis generalized residual corresponding to

$$d(\mathbf{x}; \boldsymbol{\omega}) - E(\boldsymbol{\omega}|\mathbf{x}), \quad (52)$$

where

$$E(\boldsymbol{\omega}|\mathbf{x}) = \int E_{\boldsymbol{\theta}}(\boldsymbol{\omega}|\boldsymbol{\theta})g(\boldsymbol{\theta}|\mathbf{x}; \boldsymbol{\omega})d\boldsymbol{\omega} \quad (53)$$

and

$$E_{\boldsymbol{\theta}}(\boldsymbol{\omega}|\boldsymbol{\theta}) = \sum_{\mathbf{x} \in \mathcal{X}} d(\mathbf{x}; \boldsymbol{\omega})p(\mathbf{x}|\boldsymbol{\theta}; \boldsymbol{\omega}), \ \boldsymbol{\theta} \in \Theta. \quad (54)$$

Under (3), $d(\mathbf{X}_i; \boldsymbol{\omega}) - E(\boldsymbol{\omega}|\mathbf{X}_i)$ has expectation 0. Thus, O_N is compared to the conditional estimate

$$\hat{E}_{Nc} = N^{-1} \sum_{i=1}^{N} E(\hat{\boldsymbol{\omega}}_N|\mathbf{X}_i) \quad (55)$$

of the expectation $E(\boldsymbol{\omega})$ of O_N under (3).

Comparison of O_N to $\hat{E}_{Nc}$ can be attractive in many applications. For example, this practice has advantages when a 1PL or 2PL model is examined and pairwise generalized residuals are considered as in (45). In this case, the fraction of i, $1 \leq i \leq N$, with $X_{ij} = x_j$, $0 \leq x_j \leq r_j - 1$, $1 \leq j \leq q$, is equal to $p_j(x_j; \hat{\boldsymbol{\omega}}_N)$. It follows that for a given $j < j_1$, all values of z_{LN} obtained for $0 \leq x_{1j} \leq 1$ and $0 \leq x_{1j_1} \leq 1$ are equal except for sign.

Existence of model error does not imply that the error has practical impact on the model application. This issue is commonly noted in the statistical literature. For example, consider the following quotation (Box and Draper 1987, p. 74): “Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.” Research on the size of model error required for practical impact on applications of IRT appears to be relatively limited (Hambleton and Han 2005; Sinharay 2005). In the examples in Section 4, attempts are made to check on practical consequences of model error in applied work.

4. APPLICATIONS

4.1 Application 1: A Test of Basic Skills

This example involves the responses of $N = 8686$ examinees to a separately timed section with $q = 45$ five-option multiple-choice (MC) items in the writing assessment part of a test of basic skills. The first 25 items concern error detection, while the last 20 treat error correction. Experts believed that the test was *speeded*; that is, some examinees had difficulty completing the test within the allotted time. This situation is undesirable unless the test is designed to measure ability to respond quickly, which was not the case in this particular assessment.

In this example, all items are right-scored, so that each X_{ij} is 0 or $r_j - 1 = 1$, where 1 corresponds to a correct answer and 0 corresponds to an incorrect or omitted answer. The 2PL model for $H = 1$ was applied, and Louis generalized residuals z_{LN} were computed with $d(\mathbf{x}; \boldsymbol{\omega})$ defined as in (45) for $x_{1j} = x_{1j_1} = 1$ and $1 \leq j < j_1 \leq q$. The adjustment in (52) was applied. Thus, the fraction of examinees who answered both items j and j_1 correctly was compared to the estimated probability of correct answers for both items. Because an examinee who experiences time pressure by item $j < q$ typically has difficulty providing correct answers to items $j_1 > j$, it is quite common for the local independence assumption to fail in a speeded test (Yen 1993), and the pairwise generalized residuals will tend to be positive for pairs of items near the end of the test. Because $q(q-1)/2 = 990$ pairs of generalized residuals were produced, issues of multiple comparisons obviously arise; however, in this case, the expected number of generalized residuals greater than 5 in magnitude is only about $990\Phi(-5) = 0.0003$.

For the data in question, $|z_{LN}|$ exceeded 5 for 66 of the 990 pairs, so that quite strong evidence exists that the model is not correct. Of greater interest is the concentration of large values of $|z_{LN}|$ among items 38–45. Although only 28 pairs of items j and j_1 exist such that $38 \le j < j_1 \le q = 45$, these pairs include 18 of the 66 cases with $|z_{LN}| > 5$. In each of these 18 cases, z_{LN} is greater than 5. In no pair with j and j_1 at least 38 is z_{LN} negative. Thus, the results are quite consistent with a test in which examinees have difficulty finishing the last eight items of the test. The most extreme case is at the very end of the test. For items 44 and 45, z_{LN} is 23.35. The observed O_N is 0.279, while the corresponding $\hat{E}_{Nc}$ is 0.226. For $j = 44$ and $j_1 = 45$, the observed sample correlation of the X_{ij} and X_{ij_1}, $1 \le i \le N$, is 0.341, while the estimated correlation of X_{ij} and X_{ij_1} corresponding to $\hat{E}_{Nc}$ is 0.125. Thus, the deviation from the model is not small, at least for this pair of items. In this particular example, the testing program later decided to reduce the test length from 45 items to 38 items to make the test less speeded. This decision reflected an investigation reported by Boughton, Larkin, and Yamamoto (2004), which used a specialized IRT model called a HYBRID model (Yamamoto 1989) rather than an analysis of generalized residuals.

If one proceeds with the concept that the major concentration of large pairwise interactions involves items $j_1 > 38$, there is reason to compare a test based on the first 38 items relative to a test based on the full 45 items to assess the impact of the large pairwise generalized residuals associated with the items $j_1 > 38$. One could also consider a similar comparison that involved the first 37 items. In this case, one notes problems with pairs with $j > 37$ and $j_1 > j$. One criterion is the sample correlation of the sum score S_i for the test with 45 items and the sum score S_{iu} for the test with just the first u items, where $1 \le i \le N$. For $u = 38$, this sample correlation is 0.971, a value that at first glance appears quite high. This appearance is deceptive, for S_{iu} and S_i share the common summands X_{ij} for $1 \le j \le u$. The correlation of S_i and S_{iu} is

$$\rho(S_i, S_{iu}) = \frac{\operatorname{var}(S_{iu}) + \operatorname{cov}(S_{iu}, S_i - S_{iu})}{\{\operatorname{var}(S_{iu})[\operatorname{var}(S_i - S_{iu}) + 2\operatorname{cov}(S_{iu}, S_i - S_{iu}) + \operatorname{var}(S_{iu})]\}^{1/2}}. \tag{56}$$

If S_{iu} and $S_i - S_{iu}$ are uncorrelated, as is the case if the quantity measured by the final $q - u$ items is unrelated to the quantity measured by the first u items, then the correlation of S_i and S_{iu} reduces to

$$\left[\frac{\operatorname{var}(S_{iu})}{\operatorname{var}(S_{iu}) + \operatorname{var}(S_i - S_{iu})}\right]^{1/2}.$$

If one estimates $\operatorname{var}(S_{iu})$ by the sample variance of the S_i, $1 \le i \le N$, and estimates $\operatorname{var}(S_i - S_{iu})$ by the sample variance of the $S_i - S_{iu}$, $1 \le i \le N$, then for $u = 38$ one obtains an estimate of 0.953 for the case of $S_i - S_{iu}$ and S_{iu} uncorrelated. This estimate is only about 0.018 lower than the sample correlation of S_i and S_{iu}.

4.2 Application 2: A State Test

In an achievement test for a single subject for students in a particular grade in an American state, the data included $N = 45{,}787$ students, each of whom were given $q = 65$ multiple-choice items, and each item had four answer options. In this testing program, reporting of test scores relies on a 1PL model with a standard normal distribution of the real latent variable. The model clearly does not fit the data, for a simple likelihood-ratio chi-square test to compare the 1PL and 2PL models leads to a chi-square statistic of 30,740 on 65 degrees of freedom; however, generalized residuals can provide an indication of specific model problems. As an illustration, the Louis generalized residuals z_{LN} for marginal distributions of the sum score were computed as in (50) by use of the adjustment in (52). In this case, the generalized residuals for the empirical distribution function of the S_i are quite large. For sums from 6 to 13, every z_{LN} is less than -30. As a consequence, it is clear that the data are not consistent with the 1PL model. The difficulty is that for this range of sums, the estimates $\hat{E}_{Nc}$ are somewhat larger than are the observed values O_N. For example, for $s = 9$, the estimated distribution function $\hat{E}_{Nc}$ is 0.010, while the empirical distribution function O_N is only 0.0029. These very small observed O_N for small values of s can suggest that examinees with quite limited proficiency in the tested subject are still able to obtain some correct responses by guessing. Given the 65 items and the four choices, an examinee who selects answers completely at random has an expected score of $65/4 = 16.25$, a value somewhat above 9. This particular inconsistency can be greatly reduced by use of the version of the 3PL model defined as in (33) to have a constant guessing parameter and constant item discrimination. For the sums from 6 to 13, the Louis generalized residuals are reduced to values from -1.18 to -10.1. For example, for $s = 9$, the new value of z_{LN} is -6.00, and $\hat{E}_{Nc}$ is now reduced to 0.0043, a value much closer to $O_N = 0.0029$. The restricted 3PL model just considered still does not fit the data, for a likelihood-ratio test of the 3PL model of (33) again the 3PL model of (32) with a constant guessing parameter yields a chi-square statistic of 18,200 on 64 degrees of freedom. Nonetheless, the generalized residuals associated with the empirical distribution function of the S_i are smaller in magnitude than those for the 1PL model.

The direct impact of the choice of psychometric model can be assessed in terms of impact of models on scoring. For a continuous latent variable, one reasonable approach to scoring via IRT uses the estimated expected a posteriori (EAP) mean

$$\hat{\theta}_i = \int \theta g(\theta | \mathbf{X}_i; \hat{\boldsymbol{\omega}}_N) d\theta \tag{57}$$

for the latent variable θ_i (Bock and Aitkin 1981). One may compare the sample correlations of the EAP measures for the 1PL model and the 3PL model of (33) with constant item discrimination and constant guessing parameter. This sample correlation of 0.989 is quite high; however, the sample standard deviation of the EAP estimates for the 1PL model is 0.951, a value somewhat higher than the sample standard deviation of 0.916 for the 3PL model under study. The root mean squared difference of the two estimates is 0.036, a value that is not negligible given that standard deviations of the estimates are a bit less than 1. In addition, differences in estimates vary substantially with estimated proficiency. The smallest EAP for the 3PL case with constant item discrimination and constant guessing parameter is −2.397. The corresponding EAP for the 1PL case is −3.008. The largest EAP for the 3PL case is 3.162. The corresponding value for the 1PL case is 3.440. The relationship of EAP estimates is weakest for relatively low estimated proficiency. For the subset of examinees with the smallest 1000 values of the EAP for the 3PL model of (33), the sample correlation of the 1PL and 3PL values of the EAP is 0.621. By contrast, for the examinees with the largest 1000 values of the EAP for the model of (33), the sample correlation of EAP estimates is 0.998. Thus, the model choice does appear to have impact, with the effect most pronounced for cases of low estimated proficiency.

It should be emphasized that results for the two cases in this section are not unusual. The methods used in this report have been applied to a variety of other data from commonly used educational tests, and similar results have been obtained. No IRT model appears to agree with any data encountered in educational testing, but the practical impact of model error is quite variable.

5. CONCLUSIONS

Generalized residuals are widely applicable to assessment of models for contingency tables. As we have shown, such applications are readily available for models that are neither log-linear models nor latent-class models. In particular, they are applicable to models in IRT in which the latent variable or latent vector has a continuous distribution. Generalized residuals supplement standard likelihood-ratio chi-square tests by suggesting specific weaknesses in a model, and these weaknesses can be evaluated in terms of the practical consequences of the model failure. The Louis version of generalized residuals and the variation based on (52) provide tools to simplify computations.

Generalized residuals apply not just to marginal probabilities of combinations of item responses. They can be used with distributions of sum scores, expectations of sum scores, products of item scores and sum scores, and numerous other functions of interest. Thus, generalized residuals provide a flexible framework for assessing agreement of models and data. Sinharay (2006) and Sinharay, Johnson, and Stern (2006) suggested a flexible framework for IRT models, but that framework is based on the computationally intensive posterior predictive model checking method.

Although our examples both involve applications of IRT to educational tests, generalized residuals also apply to IRT applications not related to educational testing. For example, they would apply to the IRT model used by Baldwin et al. (2009) to model data on hip fracture or for the IRT applications to health outcomes by Hays, Morales, and Reise (2000).

APPENDIX: PROOFS OF THEOREMS

Proof of Theorem 1: Let

$$\nabla\ell_N(\boldsymbol{\omega}) = N^{-1}\sum_{i=1}^{N}[\pi(\mathbf{X}_i;\boldsymbol{\omega}]^{-1}\nabla\pi(\mathbf{X}_i;\boldsymbol{\omega}),\ \boldsymbol{\omega}\in\Omega \tag{A.1}$$

be the gradient function of ℓ_N. As $N \to \infty$, the difference

$$N^{1/2}\{\hat{\boldsymbol{\omega}}_N - \boldsymbol{\omega}_0 - [\mathbf{J}(\boldsymbol{\omega}_0)]^{-1}\nabla\ell_N(\boldsymbol{\omega}_0)\} \to_p \mathbf{0}_k, \tag{A.2}$$

where $\mathbf{0}_k$ is the k-dimensional vector with all elements 0 and $\to_p$ denotes convergence in probability (Birch 1964). In addition, the probability is 1 that $\ell_N(\hat{\boldsymbol{\omega}}_N) = \hat{\ell}_N$ for all but a finite number of N and $\hat{\boldsymbol{\omega}}_N$ converges to $\boldsymbol{\omega}$ (Rao 1973, p. 360).

For $\boldsymbol{\omega}$ in Ω, let $\mathbf{y}(\boldsymbol{\omega})$ in R^k be a solution of the equation

$$\mathbf{J}(\boldsymbol{\omega})\mathbf{y}(\boldsymbol{\omega}) = \sum_{\mathbf{x}\in\mathcal{X}} d(\mathbf{x};\boldsymbol{\omega})\nabla\pi(\mathbf{x};\boldsymbol{\omega}), \tag{A.3}$$

so that

$$\sigma^2(\boldsymbol{\omega}) = S(\mathbf{y}(\boldsymbol{\omega});\boldsymbol{\omega}). \tag{A.4}$$

For any $\mathbf{y}$ in R^k, $\mathbf{x}$ in $\mathcal{X}$, and $\boldsymbol{\omega}$ in Ω, let

$$e(\mathbf{x},\mathbf{y};\boldsymbol{\omega}) = d(\mathbf{x};\boldsymbol{\omega}) - E(\boldsymbol{\omega}) - \mathbf{y}'[\nabla\lambda(\mathbf{x};\boldsymbol{\omega}) - \mathbf{m}(\boldsymbol{\omega})]. \tag{A.5}$$

Let

$$\bar{e}_N(\boldsymbol{\omega}) = N^{-1}\sum_{i=1}^{N} e(\mathbf{X}_i,\mathbf{y}(\boldsymbol{\omega});\boldsymbol{\omega}). \tag{A.6}$$

Because

$$\sum_{\mathbf{x}\in\mathbf{X}} \nabla\pi(\mathbf{x},\boldsymbol{\omega}) = \mathbf{0}_k, \tag{A.7}$$

$e(\mathbf{X}_i,\mathbf{y}(\boldsymbol{\omega}_0);\boldsymbol{\omega}_0)$ has expectation 0 for $i \geq 1$, and

$$N^{1/2}[O_N - \hat{E}_N - \bar{e}_N(\boldsymbol{\omega}_0)] \to_p 0 \tag{A.8}$$

(Rao 1973, pp. 391–394). It follows that $N^{1/2}(O_N - \hat{E}_N)$ converges in distribution to a normal random variable with mean 0 and variance $\sigma^2(\boldsymbol{\omega}_0)$.

Clearly $\sigma^2(\boldsymbol{\omega})$ is continuous in $\boldsymbol{\omega}$. Because $\hat{\boldsymbol{\omega}}_N$ converges to $\boldsymbol{\omega}_0$ with probability 1 (Rao 1973, p. 360), s_N converges to $\sigma(\boldsymbol{\omega}_0)$ with probability 1. If $\sigma(\boldsymbol{\omega}_0) > 0$, then z_N, $N \geq 1$, converges in distribution to a standard normal random variable (Rao 1973, p. 122).

Let $f_N(\mathbf{x})$, $\mathbf{x}$ in $\mathcal{X}$, be the fraction of i, $1 \leq i \leq N$, such that $\mathbf{X}_i = \mathbf{x}$. By the strong law of large numbers, $f_N(\mathbf{x})$ converges with probability 1 to $\pi(\mathbf{x};\boldsymbol{\omega}_0)$ as N approaches ∞. By standard arguments from least squares, if

$$\hat{J}_{LN} = N^{-1}\sum_{i=1}^{N}[\hat{\pi}(\mathbf{X}_i)]^{-2}\nabla\pi(\mathbf{X}_i;\hat{\boldsymbol{\omega}}_N)[\nabla\pi(\mathbf{X}_i;\hat{\boldsymbol{\omega}}_N)]' \tag{A.9}$$

and the k-dimensional vector $\hat{\mathbf{y}}_{LN}$ satisfies

$$\hat{J}_{LN}\hat{\mathbf{y}}_{LN} = \sum_{i=1}^{N}\hat{d}_N(\mathbf{X}_i)[\hat{\pi}_N(\mathbf{X}_i)]^{-1}\nabla\pi(\mathbf{X}_i;\hat{\boldsymbol{\omega}}_N), \tag{A.10}$$

then

$$\hat{\sigma}^2_{LN} = S_{LN}(\hat{\mathbf{y}}_{LN}). \tag{A.11}$$

It follows by elementary arguments that $\hat{\sigma}^2_{LN}$ converges to $\sigma^2(\boldsymbol{\omega}_0)$ with probability 1, and z_{LN}, $N \geq 1$, converges in distribution to a standard normal random variable.

Proof of Theorem 2: The basic observation required is that, for $\mathbf{x}$ in $\mathcal{X}$, the gradient of

$$\log\pi(\mathbf{x};\boldsymbol{\omega}) = \lambda(\mathbf{x};\boldsymbol{\omega}) - \log\gamma(\boldsymbol{\omega}) \tag{A.12}$$

at $\boldsymbol{\omega}$ in Ω is

$$\nabla\log\pi(\mathbf{x};\boldsymbol{\omega}) = [\pi(\mathbf{x},\boldsymbol{\omega})]^{-1}\nabla\pi(\mathbf{x};\boldsymbol{\omega}) = \nabla\lambda(\mathbf{x};\boldsymbol{\omega}) - \mathbf{m}(\boldsymbol{\omega}). \tag{A.13}$$

Equation (14) follows from (A.13), while (A.12) leads to (15).

The Fisher information matrix at $\boldsymbol{\omega}$ in Ω is positive definite if, and only if, no $\mathbf{u}$ in R^k, $\mathbf{u} \neq \mathbf{0}_k$, exists such that

$$\mathbf{u}'\nabla\lambda(\mathbf{x};\boldsymbol{\omega}) = \mathbf{u}'\mathbf{m}(\boldsymbol{\omega}),\ \ \mathbf{x} \in \mathcal{X}. \tag{A.14}$$

If (A.14) holds, then $\mathbf{u}'\nabla\lambda(\mathbf{x};\boldsymbol{\omega})$ is constant for $\mathbf{x}$ in $\mathcal{X}$. Conversely, if, for some real h, $\mathbf{u}'\nabla\lambda(\mathbf{x};\boldsymbol{\omega}) = h$ for all $\mathbf{x}$ in $\mathcal{X}$, then $\mathbf{u}'\mathbf{m}(\boldsymbol{\omega}) = h$ and (A.14) holds. Thus, the Fisher information $\mathbf{J}(\boldsymbol{\omega})$ is positive definite if, and only if, no $\mathbf{u}$ in R^k other than $\mathbf{0}_k$ exists such that $\mathbf{u}'\nabla\lambda(\mathbf{x};\boldsymbol{\omega})$ is constant over $\mathbf{x}$ in $\mathcal{X}$.

By (2) and (A.13),

$$\begin{aligned}\mathbf{0}_k &= \sum_{\mathbf{x}\in\mathcal{X}} \nabla\pi(\mathbf{x};\boldsymbol{\omega}) = \sum_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x};\boldsymbol{\omega})[\nabla\lambda(\mathbf{x};\boldsymbol{\omega}) - \mathbf{m}(\boldsymbol{\omega})]\\ &= -\mathbf{m}(\boldsymbol{\omega}) + \sum_{\mathbf{x}\in\mathcal{X}} \pi(\mathbf{x};\boldsymbol{\omega})\nabla\lambda(\mathbf{x};\boldsymbol{\omega}),\end{aligned} \tag{A.15}$$

so that standard arguments from least squares imply that for fixed $\mathbf{y}$ in R^k, $S(a,\mathbf{y})$ is minimized when $a = E(\boldsymbol{\omega}) - \mathbf{y}'\mathbf{m}(\boldsymbol{\omega})$. Thus, $\sigma^2(\boldsymbol{\omega})$ is the minimum of $S(a,\mathbf{y})$ for a in R and $\mathbf{y}$ in R^k. This minimum is positive if, and only if, no $\mathbf{u}$ in R^k exists such that $d(\mathbf{x};\boldsymbol{\omega}) - \mathbf{u}'\nabla\lambda(\mathbf{x};\boldsymbol{\omega})$ is a real constant a for all $\mathbf{x}$ in $\mathcal{X}$.

A similar argument applies to $S_{LN}(a,\mathbf{y})$ if $\ell_N(\hat{\boldsymbol{\omega}}_N) = \hat{\ell}_N$. In this case, $\nabla\ell_N(\hat{\boldsymbol{\omega}}_N) = \mathbf{0}_k$, so that (A.1) and (A.13) imply that

$$\begin{aligned}\mathbf{0}_k &= N^{-1}\sum_{i=1}^{N} \nabla \log\pi(\mathbf{X}_i;\boldsymbol{\omega}) = N^{-1}\sum_{i=1}^{N}[\nabla\lambda(\mathbf{x};\hat{\boldsymbol{\omega}}_N) - \mathbf{m}(\hat{\boldsymbol{\omega}}_N)]\\ &= -\mathbf{m}(\hat{\boldsymbol{\omega}}_N) + N^{-1}\sum_{i=1}^{N}\nabla\lambda(\mathbf{X}_i;\hat{\boldsymbol{\omega}}_N).\end{aligned} \tag{A.16}$$

For fixed $\mathbf{y}$ in R^k, $S_{LN}(a,\mathbf{y})$ is minimized by $a = \hat{O}_N - \mathbf{y}'\mathbf{m}(\hat{\boldsymbol{\omega}}_N)$. Thus, $\hat{\sigma}^2_{LN}$ is the minimum of $S_{LN}(a,\mathbf{y})$ for a real and $\mathbf{y}$ in R^k. This minimum is positive if, and only if, no $\mathbf{u}$ in R^k exists such that $\hat{d}_N(\mathbf{X}_i) - \mathbf{u}'\nabla\lambda(\mathbf{x};\hat{\boldsymbol{\omega}}_N)$ is a real constant a for $1 \le i \le N$.

Proof of Theorem 4: Consider $\boldsymbol{\omega}$ in Ω. For $\boldsymbol{\theta}$ in Θ, let $\tau(\boldsymbol{\theta})$ be the supremum of $|\nabla\mu(\boldsymbol{\theta};\boldsymbol{\omega}_1)|$ for $\boldsymbol{\omega}_1$ in R^k such that $|\boldsymbol{\omega}_1 - \boldsymbol{\omega}| \le \epsilon_1$. Because $\mu(\boldsymbol{\theta};\cdot)$ is continuously differentiable, $\tau(\boldsymbol{\theta})$ is finite.

The difference

$$\delta(\boldsymbol{\omega}_1) - \delta(\boldsymbol{\omega}) = \int \Delta(\boldsymbol{\theta};\boldsymbol{\omega}_1,\boldsymbol{\omega})g(\boldsymbol{\theta};\boldsymbol{\omega})d\boldsymbol{\theta}, \tag{A.17}$$

where, for any $\boldsymbol{\theta}$ in Θ,

$$\begin{aligned}\Delta(\boldsymbol{\theta};\boldsymbol{\omega}_1,\boldsymbol{\omega}) &= \frac{g(\boldsymbol{\theta};\boldsymbol{\omega}_1) - g(\boldsymbol{\theta};\boldsymbol{\omega})}{g(\boldsymbol{\theta};\boldsymbol{\omega})}\\ &= \exp[\mu(\boldsymbol{\theta};\boldsymbol{\omega}_1) - \mu(\boldsymbol{\theta};\boldsymbol{\omega})] - 1.\end{aligned} \tag{A.18}$$

By the mean-value theorem,

$$\mu(\boldsymbol{\theta};\boldsymbol{\omega}_1) - \mu(\boldsymbol{\theta};\boldsymbol{\omega}) = (\boldsymbol{\omega}_1 - \boldsymbol{\omega})'\nabla(\boldsymbol{\theta};\boldsymbol{\omega}_2) \tag{A.19}$$

for some $\boldsymbol{\omega}_2 = \alpha\boldsymbol{\omega}_1 + (1-\alpha)\boldsymbol{\omega}$ and α in $(0,1)$. By the Cauchy-Schwarz inequality,

$$|\mu(\boldsymbol{\theta};\boldsymbol{\omega}_1) - \mu(\boldsymbol{\theta};\boldsymbol{\omega})| \le \tau(\boldsymbol{\theta})|\boldsymbol{\omega}_1 - \boldsymbol{\omega}|. \tag{A.20}$$

Let $f(y) = y\exp(y)$ for real y. Application of Taylor's theorem to the exponential function shows that

$$\Delta(\boldsymbol{\theta};\boldsymbol{\omega}_1,\boldsymbol{\omega}) = f((\boldsymbol{\omega}_1 - \boldsymbol{\omega})'\nabla(\boldsymbol{\theta};\boldsymbol{\omega}_2)). \tag{A.21}$$

Because $|f(y) - y| \le |y|\exp(|y|)$ for y real, if $|\boldsymbol{\omega}_1 - \boldsymbol{\omega}|$ is also less than ϵ, then

$$\frac{|\Delta(\boldsymbol{\theta};\boldsymbol{\omega}_1,\boldsymbol{\omega}) - (\boldsymbol{\omega}_1 - \boldsymbol{\omega})'\nabla(\boldsymbol{\theta};\boldsymbol{\omega}_2)|}{|\boldsymbol{\omega}_1 - \boldsymbol{\omega}|} \le \tau(\boldsymbol{\theta})\exp[\epsilon\tau(\boldsymbol{\theta})] \le \kappa(\boldsymbol{\theta}). \tag{A.22}$$

The dominated convergence theorem then leads to the result that $\mathbf{m}_0(\boldsymbol{\omega})$ is finite and

$$\frac{|\delta(\boldsymbol{\omega}_1) - \delta(\boldsymbol{\omega}) - (\boldsymbol{\omega}_1 - \boldsymbol{\omega})'\mathbf{m}_0(\boldsymbol{\omega})|}{|\boldsymbol{\omega}_1 - \boldsymbol{\omega}|}$$

converges to 0 when $\boldsymbol{\omega} \neq \boldsymbol{\omega}$ and $\boldsymbol{\omega}_1$ converges to $\boldsymbol{\omega}$. Thus, the gradient of δ at $\boldsymbol{\omega}$ is $\mathbf{m}_0(\boldsymbol{\omega})$. The conclusion now follows by straightforward differentiation.

[Received March 2011. Revised July 2013.]

REFERENCES

Adams, R. J., Wilson, M. R., and Wang, W. C. (1997), "The Multidimensional Random Coefficients Multinomial Logit Model," *Applied Psychological Measurement*, 21, 1–23. [1438]

Armitage, P. (1955), "Tests for Linear Trends in Proportions and Frequencies," *Biometrics*, 11, 375–386. [1435]

Baldwin, P., Bernstein, J., and Wainer, H. (2009), "Hip Psychometrics," *Statistics in Medicine*, 28, 2277–2292. [1437,1442]

Berk, R. H. (1972), "Consistency and Asymptotic Normality of MLE's for Exponential Models," *The Annals of Mathematical Statistics*, 43, 193–204. [1437]

Birch, M. W. (1964), "A New Proof of the Pearson-Fisher Theorem," *Annals of Mathematical Statistics*, 35, 817–824. [1442]

Birnbaum, A. (1968), "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability," in *Statistical Theories of Mental Test Scores*, eds. F. M. Lord and M. R. Novick, Reading, MA: Addison-Wesley, pp. 397–479. [1435,1438]

Bock, R. D. (1997), "A Brief History of Item Response Theory," *Educational Measurement: Issues and Practice*, 16, 21–33. [1437]

Bock, R. D., and Aitkin, M. (1981), "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm," *Psychometrika*, 46, 443–459. [1439,1441]

Bock, R. D., and Moustaki, I. (2007), "Item Response Theory in a General Framework," in *Handbook of Statistics* (Vol. 26), eds. C. R. Rao and S. Sinharay, Amsterdam: North-Holland, pp. 469–513. [1438]

Boughton, K., Larkin, K., and Yamamoto, K. (2004), "Modeling Differential Speededness Using a Hybrid Psychometric Approach," paper presented at the annual meeting of the American Educational Research Association, San Diego, CA. [1441]

Box, G. E. P., and Draper, N. R. (1987), *Empirical Model-Building and Response Surfaces*, New York: Wiley. [1440]

Cai, L., du Toit, S. H. C., and Thissen, D. (2011), *IRTPRO: Flexible, Multidimensional, Multiple Categorical IRT Modeling*, Chicago, IL: Scientific Software International. [1439]

Chang, H., and Ying, Z. (2009), "Nonlinear Sequential Designs for Logistic Item Response Theory Models With Applications to Computerized Adaptive Tests," *The Annals of Statistics*, 37, 1466–1488. [1437]

Chao, A. (1987), "Estimating the Population Size for Capture-Recapture Data With Unequal Catchability," *Biometrics*, 43, 783–791. [1437]

Cochran, W. G. (1954), "Some Methods for Strengthening the Common χ^2 Tests," *Biometrics*, 10, 417–451. [1435]

——— (1955), "A Test of a Linear Function of the Deviations Between Observed and Expected Numbers," *Journal of the American Statistical Association*, 50, 377–397. [1435]

Cohen, J., and Jiang, T. (1999), "Comparison of Partially Measured Latent Traits Across Nominal Subgroups," *Journal of the American Statistical Association*, 94, 1035–1044. [1437]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society,* Series B, 39, 1–38. [1439]

Eaton, W. W., and Bohrnstedt, G. W. (eds.) (1989), "Latent Variable Models for Dichotomous Outcomes: Analysis of Data From Epidemiological Catchment Area Program," *Sociological Methods and Research*, 18, 3–182. [1437]

Glas, C. A. W., and Verhelst, N. D. (1989), "Extensions of the Partial Credit Model," *Psychometrika*, 54, 635–659. [1435,1439]

——— (1995), "Testing the Rasch Model," in *Rasch Models: Foundations, Recent Developments, and Applications*, eds. G. H. Fischer and I. W. Molenaar, New York: Springer, pp. 69–95. [1435,1439]

Haberman, S. J. (1976), "Generalized Residuals for Log-Linear Models," in *Proceedings of the Ninth International Biometrics Conference,* vol. 1, pp. 104–172. [1435]

——— (1977a), "Log-Linear Models and Frequency Tables With Small Expected Cell Counts," *The Annals of Statistics*, 5, 1148–1169. [1437]

——— (1977b), "Maximum Likelihood Estimates in Exponential Response Models," *The Annals of Statistics*, 5, 815–841. [1437]

——— (1978), *Analysis of Qualitative Data, Volume I: Introductory Topics*, New York: Academic Press. [1435,1437]

——— (1979), *Analysis of Qualitative Data, Volume II: New Developments*, New York: Academic Press. [1435]

——— (1988), "A Stabilized Newton-Raphson Algorithm for Log-Linear Models for Frequency Tables Derived by Indirect Observation," *Sociological Methodology*, 18, 193–211. [1437,1439]

——— (2004), "Joint and Conditional Maximum Likelihood Estimation for the Rasch Model for Binary Responses," Research Rep. No. RR-04-20, Princeton, NJ: ETS. [1435,1439]

——— (2005), "Latent-Class Item Response Models," Research Rep. No. RR-05-28, Princeton, NJ: ETS. [1437]

——— (2006), "An Elementary Test of the Normal 2PL Model Against the Normal 3PL Alternative," ETS Research Rep. No. RR-06-14, Princeton, NJ: ETS. [1440]

——— (2009), "Use of Generalized Residuals to Examine Goodness of Fit of Item Response Models," ETS Research Report No. RR-09-15, Princeton, NJ: ETS. [1439]

Haberman, S. J., Sinharay, S., and Chon, K. H. (2013), "Assessing Item Fit for Unidimensional Item Response Theory Models Using Residuals From Estimated Item Response Functions," *Psychometrika*, 78, 417–440. [1439]

Haberman, S. J., von Davier, M., and Lee, Y. (2008), "Comparison of Multidimensional Item Response Models: Multivariate Normal Ability Distributions Versus Multivariate Polytomous Distributions," ETS Research Rep. No. RR-08-45, Princeton, NJ: ETS. [1439]

Hambleton, R. K., and Han, N. (2005), "Assessing the Fit of IRT Models to Educational and Psychological Test Data: A Five Step Plan and Several Graphical Displays," in *Advances in Health Outcomes Research Methods, Measurement, Statistical Analysis, and Clinical Applications*, eds. W. R. Lenderking and D. Revicki, Washington, DC: Degnon Associates, pp. 57–78. [1440]

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991), *Fundamentals of Item Response Theory*, Newbury Park, CA: Sage. [1435,1438,1439]

Hays, R. D., Morales, L. S., and Reise, S. P. (2000), "Item Response Theory and Health Outcomes Measurement in the 21st Century," *Medical Care*, 38, II28–II42. [1442]

Heinen, T. (1996), *Latent Class and Discrete Latent Trait Models*, Thousand Oaks, CA: Sage. [1437]

Holland, P. W. (1990), "The Dutch Identity: A New Tool for the Study of Item Response Models," *Psychometrika*, 55, 5–18. [1438]

Holland, P. W., and Rosenbaum, P. (1986), "Conditional Association and Unidimensionality in Monotone Latent Variable Models," *The Annals of Statistics*, 14, 1523–1543. [1435,1437]

Junker, B. W. (1993), "Conditional Association, Essential Independence and Monotone Unidimensional Item Response Models," *The Annals of Statistics*, 21, 1359–1378. [1435]

Lord, F. M. (1980), *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale, NJ: Erlbaum. [1435,1438]

Lord, F. M., and Wingersky, M. S. (1984), "Comparison of IRT True-Score and Equipercentile Observed-Score 'Equatings'," *Applied Psychological Measurement*, 8, 453–461. [1440]

Louis, T. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 44, 226–233. [1436]

Mantel, N., and Haenszel, W. (1959), "Statistical Aspects of the Analysis of Data From Retrospective Studies," *Journal of the National Cancer Institute*, 22, 719–748. [1435]

Muraki, E. (1997), "A Generalized Partial Credit Model," in *Handbook of Modern Item Response Theory*, eds. W. J. van der Linden and R.K. Hambleton, New York: Springer-Verlag, chap. 9. [1438,1439]

Muraki, E., and Bock, R. D. (2003), *PARSCALE 4: IRT Item Analysis and Test Scoring for Rating-Scale Data*, Chicago, IL: Scientific Software. [1439]

Naylor, A. F. M., and Smith, J. C. (1982), "Applications of a Method for the Efficient Computation of Posterior Distributions," *Applied Statistics*, 31, 214–225. [1439]

Orlando, M., and Thissen, D. (2000), "Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models," *Applied Psychological Measurement*, 24, 50–64. [1440]

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), New York: Wiley. [1435,1442]

Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen: Danish Institute for Educational Research. [1435,1438]

Reckase, M. D. (1997), "The Past and Future of Multidimensional Item Response Theory," *Applied Psychological Measurement*, 21, 25–36. [1438]

Reiser, M. (1996), "Analysis of Residual for the Multinomial Item Response Model," *Psychometrika*, 61, 509–528. [1436,1439]

Rice, K. M. (2004), "Equivalence Between Conditional and Mixture Approaches to the Rasch Model and Matched Case-Control Studies, With Applications," *Journal of the American Statistical Association*, 99, 510–522. [1437]

Scott, S. L., and Ip, E. H. (2002), "Empirical Bayes and Item Clustering Effects in a Latent Variable Hierarchical Model: A Case Study From the National Assessment of Educational Progress," *Journal of the American Statistical Association*, 97, 409–419. [1437]

Sinharay, S. (2005), "Assessing Fit of Unidimensional Item Response Theory Models Using a Bayesian Approach," *Journal of Educational Measurement*, 42, 375–394. [1440]

——— (2006), "Model Diagnostics for Bayesian Networks," *Journal of Educational and Behavioral Statistics*, 31, 1–33. [1442]

Sinharay, S., Johnson, M. S., and Stern, H. S. (2006), "Posterior Predictive Assessment of Item Response Theory Models," *Applied Psychological Measurement*, 30, 298–321. [1442]

Thissen, D., Pommerich, M., Billeaud, K., and Williams, V. (1995), "Item Response Theory for Scores on Tests Including Polytomous Items With Ordered Responses," *Applied Psychological Measurement*, 19, 39–49. [1440]

Tjur, T. (1982), "A Connection Between Rasch's Item Analysis Model and a Multiplicative Poisson Model," *Scandinavian Journal of Statistics*, 9, 23–30. [1435,1439]

Yamamoto, K. (1989), "A Hybrid Model of IRT and Latent Class Models," ETS Research Rep. No. RR-89-41, Princeton, NJ: ETS. [1441]

Yates, F. (1948), "The Analysis of Contingency Tables With Groupings Based on Quantitative Characters," *Biometrika*, 35, 176–181. [1435]

Yen, W. (1993), "Scaling Performance Assessments: Strategies for Managing Local Item Dependence," *Journal of Educational Measurement*, 30, 187–213. [1440]

Ziegler, A. (2011), *Generalized Estimating Equations*, New York: Springer. [1439]