Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 16

A Strong True-Score Model for Polytomous Items*

Won-Chan Lee
Tianyou Wang
Seonghoon Kim
Robert L. Brennan†

April 2006

---

*A previous version of this paper was presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, April 2006.

†Won-Chan Lee is Research Scientist, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@uiowa.edu). Tianyou Wang is Research Scientist, CASMA, University of Iowa. Seonghoon Kim is Research Associate, ACT, Inc. Robert L. Brennan is E. F. Lindquist Chair in Measurement and Testing and Director, CASMA, University of Iowa.

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

All rights reserved

Contents

1 Introduction **1**

2 The Dirichlet-Multinomial Model **1**

3 Estimation **3**

4 Applications of the Model **4**
4.1 Smoothing Observed-Score Frequency Distribution 5
4.2 Estimating Conditional SEMs and Reliability 5
4.3 Estimating Classification Consistency and Accuracy 7
4.3.1 Classification Consistency 7
4.3.2 Classification Accuracy 8

5 Real Data Examples **9**
5.1 Data Sources 10
5.2 Results 10

6 Discussion **21**

7 References **23**

A Moments of the Beta-Binomial Distribution **25**

List of Tables

1 Summary Statistics for Reading 11
2 Parameter Estimates and Variances of Category Scores for Reading 12
3 Classification Indices for Reading Using the Third Category . . . 13
4 Summary Statistics for Math 15
5 Parameter Estimates and Variances of Category Scores for Math 16
6 Classification Indices for Math Using the First Category 16
7 Summary Statistics for Writing 17
8 Parameter Estimates and Variances of Category Scores for Writing 18
9 Classification Indices for Writing Using the Fifth Category 19
10 Summary Statistics for Essay 19
11 Parameter Estimates and Variances of Category Scores for Essay 20
12 Classification Indices for Essay Using the First Category 21

List of Figures

1 Raw-to-scale score Transformation Functions 26
2 Actual and Fitted Raw-score Distributions for Reading 27
3 Actual and Fitted Scale-score Distributions for Reading 28
4 Conditional SEMs for Reading Using the Third Category 29
5 Number of Item Combinations Leading to Total Raw-score Points 30
6 Actual and Fitted Raw-score Distributions for Math 31
7 Actual and Fitted Scale-score Distributions for Math 32
8 Conditional SEMs for Math Using the First Category 33
9 Actual and Fitted Raw-score Distributions for Writing 34
10 Actual and Fitted Scale-score Distributions for Writing 35
11 Conditional SEMs for Writing Using the Fifth Category 36
12 Actual and Fitted Raw-score Distributions for Essay 37
13 Conditional SEMs for Essay Using the First Category 38

Abstract

This paper presents a strong true-score model for tests that consist of polytomously-scored items. Errors conditional on true score are assumed to be distributed as multinomial, true category-proportion scores are assumed to follow a Dirichlet distribution, and the marginal category scores are the Dirichlet-multinomial distribution. The model is illustrated using real data sets in the following psychometric applications: obtaining a smoothed fitted observed-score distribution; estimating conditional standard errors of measurement and reliability for both raw and transformed scale scores; and computing classification consistency and accuracy indices for raw and scale scores.

1 Introduction

A strong true-score model establishes a mathematical relationship between observed scores and unobservable true and error scores based on "strong" assumptions about the distributions of the true and error scores. For example, Keats and Lord (1962) assumed a two-parameter beta distribution for true scores and a binomial distribution for errors. The resulting marginal observed-score distribution is often called the beta-binomial (or negative hypergeometric) distribution. Lord (1965) used a four-parameter beta distribution for true scores and either a binomial or compound binomial distribution for errors. Lord (1965) also provided a two-term approximation to the compound binomial for practical implementation. Subsequently, Lord (1969) proposed a generalized version of the four-parameter beta distribution.

Strong true-score models have been implemented in various psychometric problems. Among them are test equating (more specifically, smoothing the observed-score distribution, see Kolen & Brennan, 2004, pp. 75–77), estimating psychometric properties such as reliability and conditional standard errors of measurement (Kolen, Hanson, & Brennan, 1992), and estimating classification consistency and accuracy (Huynh, 1976; Hanson & Brennan, 1990; Lee, Hanson, & Brennan, 2002).

The models discussed in the previous paragraph are limited to tests that consist of dichotomously-scored items. A strong true-score model is presented in this paper, which can effectively deal with polytomously-scored items. The Dirichlet-multinomial model presented in the next section can be viewed as a multivariate version of the two-parameter beta-binomial model (Keats & Lord, 1962). Note that the Dirichlet-multinomial distribution was originally derived by Mosimann (1962). The model is presented first, followed by estimation procedures. Illustrations of the model in three psychometric applications then follow.

2 The Dirichlet-Multinomial Model

We shall consider a test that consists of n polytomously-scored items. Each item is scored as one of k possible score points, $c_1, c_2, \ldots, c_k$. It is assumed that n items are drawn at random from an undifferentiated universe of such items. Let $(\pi_1, \pi_2, \ldots, \pi_k)$ denote true proportions of items scored $c_1, c_2, \ldots, c_k$. Note that $\pi_1 + \pi_2 + \cdots + \pi_k = 1$. Let $X_1, X_2, \ldots, X_k$ represent category scores, which indicate the observed numbers of items scored $c_1, c_2, \ldots, c_k$ such that $X_1 + X_2 + \cdots + X_k = n$. It is here assumed that the conditional distribution of the random variables $X_1, X_2, \ldots, X_k$ given $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_k\}$ is a multinomial (Lee, 2005a; Mosimann, 1962):

$$h(x_1, \ldots, x_k \mid \boldsymbol{\pi}) = \frac{n!}{x_1!\, x_2! \cdots x_k!}\, \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_k^{x_k}. \tag{1}$$

Without loss of generality, the π_i and x_i in Equation 1 are person-level variables.

Now we assume that the true category-proportion scores are random variables, $\Pi_1, \Pi_2, \ldots, \Pi_k$, which follow a Dirichlet distribution:

$$g(\pi_1, \ldots, \pi_k; \alpha_1, \ldots, \alpha_k) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \pi_1^{\alpha_1 - 1} \pi_2^{\alpha_2 - 1} \cdots \pi_k^{\alpha_k - 1}, \tag{2}$$

where the α parameters are greater than zero. Equation 2 is called the multivariate β-distribution by Mosimann (1962).

It can be shown that the marginal distribution of $X_1, X_2, \ldots, X_k$ for the entire population is

$$f(x_1, \ldots, x_k) = \frac{n! \left[\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)\right] \left[\prod_{i=1}^{k} \Gamma(x_i + \alpha_i)\right]}{\left[\prod_{i=1}^{k} x_i!\right] \left[\prod_{i=1}^{k} \Gamma(\alpha_i)\right] \left[\Gamma\left(n + \sum_{i=1}^{k} \alpha_i\right)\right]}. \tag{3}$$

Equation 3 is referred to here as the Dirichlet-multinomial distribution. It is sometimes called the compound multinomial, multivariate binomial-beta, or negative multivariate hypergeometric distribution by different authors (see Johnson & Kotz, 1969).

Suppose the total score Y is represented by the sum of the item scores: $Y = c_1 X_1 + c_2 X_2 + \cdots + c_k X_k$. Given the probabilities of occurrence of $X_1, X_2, \ldots, X_k$, the density of Y can be determined. Since there are multiple combinations of values of $X_1, X_2, \ldots, X_k$ that lead to a particular y value, the distribution of Y is:

$$t(y) = \sum_{c_1 x_1 + c_2 x_2 + \cdots + c_k x_k = y} f(x_1, \ldots, x_k), \tag{4}$$

where the summation is over all the values of $X_1, X_2, \ldots, X_k$ such that $c_1 x_1 + c_2 x_2 + \cdots + c_k x_k = y$. The conditional total-score distribution is similarly defined as:

$$t(y \,|\, \boldsymbol{\pi}) = \sum_{c_1 x_1 + c_2 x_2 + \cdots + c_k x_k = y} h(x_1, \ldots, x_k \,|\, \boldsymbol{\pi}). \tag{5}$$

In the multinomial error theory (Lee, 2005a), any test of n items is conceptualized as a random sample from a large pool (or universe) of such items. Each examinee in the population is presumed to be able to get certain proportions of items in the universe scored $c_1, c_2, \ldots, c_k$ designated $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_k\}$. This is much the same as an urn-sampling model, in which there are a large (or infinite) number of balls in a multinomial urn, and the proportions of $c_1, c_2, \ldots, c_k$ balls are $\pi_1, \pi_2, \ldots, \pi_k$. Parallel test forms are conceptualized as samples of n items drawn at random from the urn. This sampling model gives rise to a multinomial distribution. This multinomial distribution for a particular set of π_i values is the conditional distribution of errors due to the form differences. When

the conditional multinomial distribution is integrated over the distribution of Π_i, the resultant marginal category-score distribution is the Dirichlet-multinomial distribution. Both the conditional and marginal category-score distributions can be transformed to the total-score (Y) metric.

3 Estimation

Using the moments of the Dirichlet-multinomial distribution, it can be shown that

$$\mathcal{E}(X_i) = \frac{n\alpha_i}{\sum_{i=1}^{k} \alpha_i}, \tag{6}$$

$$\sigma^2(X_i) = n\mathcal{E}(\Pi_i)[1 - \mathcal{E}(\Pi_i)] \left[\frac{n + \sum_{i=1}^{k} \alpha_i}{1 + \sum_{i=1}^{k} \alpha_i} \right], \tag{7}$$

and

$$\sigma(X_i X_j) = -n\mathcal{E}(\Pi_i)\mathcal{E}(\Pi_j) \left[\frac{n + \sum_{i=1}^{k} \alpha_i}{1 + \sum_{i=1}^{k} \alpha_i} \right], \tag{8}$$

where $\mathcal{E}$ is the expectation operator. It is also known that the variance and covariance for the multinomial variables conditional on $\pi_1, \pi_2, \ldots, \pi_k$ are

$$\sigma^2(X_i \,|\, \boldsymbol{\pi}) = n\pi_i(1 - \pi_i), \tag{9}$$

and

$$\sigma(X_i X_j \,|\, \boldsymbol{\pi}) = -n\pi_i\pi_j. \tag{10}$$

If the parameters of the multinomial distribution, π_i, are fixed to $\mathcal{E}(\Pi_i)$, Equations 7 through 10 give the following result:

$$\boldsymbol{\Sigma}_d = \left[\frac{n + \sum_{i=1}^{k} \alpha_i}{1 + \sum_{i=1}^{k} \alpha_i} \right] \boldsymbol{\Sigma}_m = C\boldsymbol{\Sigma}_m, \tag{11}$$

where $\boldsymbol{\Sigma}_d$ and $\boldsymbol{\Sigma}_m$ are the variance-covariance matrix of the Dirichlet-multinomial and multinomial distribution, respectively. Then, it follows, in terms of determinants, that

$$|\boldsymbol{\Sigma}_d^*| = C^{k-1} \, |\boldsymbol{\Sigma}_m^*|, \tag{12}$$

where $\boldsymbol{\Sigma}_d^*$ and $\boldsymbol{\Sigma}_m^*$ denote the variance-covariance matrices of the first $k - 1$ variables, which is required to avoid having singular matrices.

Let $\hat{\boldsymbol{\Sigma}}_d^*$ and $\hat{\boldsymbol{\Sigma}}_m^*$ be consistent estimators of $\boldsymbol{\Sigma}_d^*$ and $\boldsymbol{\Sigma}_m^*$. Further let $\bar{x}_i$ denote the sample mean observed number of items scored c_i. The constant, C, can be estimated using Equation 12 in terms of estimators as:

$$\hat{C} = \left(\frac{|\hat{\boldsymbol{\Sigma}}_d^*|}{|\hat{\boldsymbol{\Sigma}}_m^*|} \right)^{1/(k-1)}. \tag{13}$$

For $\hat{\boldsymbol{\Sigma}}_m^*$, $\hat{\sigma}^2(X_i \mid \boldsymbol{\pi}) = [\bar{x}_i(n - \bar{x}_i)]/n$, and $\hat{\sigma}(X_i X_j \mid \boldsymbol{\pi}) = -\bar{x}_i \bar{x}_j / n$ can be used. For $\hat{\boldsymbol{\Sigma}}_d^*$, the sample variances, $\hat{\sigma}^2(X_i)$, and covariances, $\hat{\sigma}(X_i X_j)$, are used. Then, from Equation 11, $\sum_{i=1}^k \hat{\alpha}_i$ is computed by

$$\sum_{i=1}^{k} \hat{\alpha}_i = \frac{\hat{C} - n}{1 - \hat{C}}. \tag{14}$$

The individual $\hat{\alpha}_i$ values are then estimated using the first moment of the Dirichlet-multinomial distribution (i.e., Equation 6). Setting $\bar{x}_i$ equal to its expected value gives the following k equations:

$$\bar{x}_i = \frac{n\hat{\alpha}_i}{\sum_{i=1}^{k} \hat{\alpha}_i}. \tag{15}$$

Solving the k equations in Equation 15 provides unique solutions to individual $\hat{\alpha}_i$ parameter estimates.

It is evident from Equation 13 that the constant C is estimated based on $k - 1$ variables. As Mosimann (1962) noted, C can also be estimated using any one of the k variables. Note that Equation 13 can be expressed in terms of estimators as:

$$\hat{\sigma}^2(X_i) = \left[\frac{\bar{x}_i(n - \bar{x}_i)}{n} \right] \hat{C}. \tag{16}$$

It follows that

$$\hat{C} = \frac{n\hat{\sigma}^2(X_i)}{\bar{x}_i(n - \bar{x}_i)}. \tag{17}$$

Equation 17 clearly shows that $\hat{C}$ can be determined from the sample mean and variance of a single variable. Thus, there are $k + 1$ possible sets of estimates of C—i.e., using each of k variables and using all variables simultaneously. Since the effectiveness of these estimates is currently not known, it would be wise to compute all of them and choose the one that gives the best results depending upon the purpose of the data analysis. The Appendix provides moments of the beta-binomial distribution when there are only two score categories.

4 Applications of the Model

This section deals with three potential applications of the Dirichlet-multinomial model in practice: smoothing the observed-score distribution, estimating conditional standard errors of measurement (SEMs) and reliability, and estimating

classification consistency and accuracy indices. All results are computed for both raw and scale scores. The results of several real data examples are presented in a later section of this paper.

4.1 Smoothing Observed-Score Frequency Distribution

The smoothed observed-score distribution can be used for various purposes. One such example would be test score equating (Kolen & Brennan, 2004). Assuming the sample size for the data available is sufficiently large, the α parameters of the model can be estimated as discussed in the previous section. Then, there are at least two different ways of obtaining the marginal score distribution.

First, the estimated parameters, $\hat{\alpha}_i$, are substituted for the parameters in Equation 3 to compute $\hat{f}(x_1, \ldots, x_k)$ for all possible combinations of $x_1, x_2, \ldots, x_k$ values. Then, the total-score distribution, $\hat{t}(y)$, can be computed using Equation 4. Alternatively, once the distribution of Π_i is determined from the estimated α parameters (see Equation 2), a large random sample of true category-proportion scores can be drawn from the distribution. The conditional total-score distribution for each individual is computed using Equations 1 and 5. The marginal fitted total-score distribution is then computed by taking the average of the conditional distributions over all individuals in the sample. The second approach is employed in this paper. With a sufficiently large sample size, the two approaches would produce similar results.

4.2 Estimating Conditional SEMs and Reliability

Reliability and conditional standard errors of measurement (SEMs) are important psychometric properties of test scores. It is shown in this section that the Dirichlet-multinomial model can be used to estimate reliability and conditional SEMs for total (raw) scores and transformed scale scores. Let $\tau = c_1\pi_1 + c_2\pi_2 + \cdots + c_k\pi_k$ denote true score. The marginal distribution of the total scores (i.e., Equation 4), which can be expressed in terms of the conditional total-score distribution and the true-score distribution (Lord, 1965), is given by:

$$t(y) = \Pr(Y = y) = \int_{-\infty}^{\infty} \Pr(Y = y \,|\, \tau) u(\tau) d\tau, \tag{18}$$

where $u(\tau)$ is the distribution of τ, and $\Pr(Y = y \,|\, \tau) = t(y \,|\, \tau)$ is the conditional error distribution, which is the same as Equation 5 except that the conditioning variable τ is a function of individual π_i values. The marginal total-score variance is given by

$$\sigma^2(Y) = \sum_i i^2 \Pr(Y = i) - \left[\sum_i i \Pr(Y = i) \right]^2, \tag{19}$$

where the summations are taken over all y values specified in the space of Y.

Similarly, the conditional error variance (Kolen et al., 1992; Lee, Brennan, & Kolen, 2000) is defined by

$$\sigma^2(Y\,|\,\tau) = \sum_i i^2 \Pr(Y = i\,|\,\tau) - \tau^2. \tag{20}$$

The square root of Equation 20 is the conditional SEM for examinees with the true score τ—more specifically, for examinees with a particular combination of π_i values that lead to the true score τ. The overall error variance for the entire population is

$$\sigma^2(E_Y) = \int_{-\infty}^{\infty} \sigma^2(Y\,|\,\tau) u(\tau) d\tau. \tag{21}$$

A reliability coefficient is commonly defined as

$$\rho^2(Y) = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(E_Y)}. \tag{22}$$

Note that the error variance, $\sigma^2(E_Y)$, estimated under the multinomial model is absolute error variance in generalizability theory (Lee, 2005a). Thus, the reliability coefficient estimated by the Dirichlet-multinomial model (i.e., Equation 22) will be a close approximation to the dependability coefficient Φ (see Brennan, 2001, for more detailed description of the Φ coefficient).

Now suppose that the total scores are transformed to a certain type of scale score S using a monotone non-decreasing transformation function $v(Y)$. The marginal scale-score distribution is determined as

$$w(s) = \Pr(S = s) = \sum_{y:v(y)=s} t(y), \tag{23}$$

where $y : v(y) = s$ indicates that the summation is taken over all y values such that $v(y) = s$. The conditional scale-score distribution, $w(s\,|\,\tau) = \Pr(S = s\,|\,\tau)$, can be determined in a similar manner using $t(y\,|\,\tau)$. The conditional expected scale score is defined as

$$\xi = \mathcal{E}(S\,|\,\tau) = \sum_i i \Pr(S = i\,|\,\tau), \tag{24}$$

where the summations are taken over all s values specified in the space of S.

The conditional scale-score error variance is given by

$$\sigma^2(S\,|\,\tau) = \sum_i i^2 \Pr(S = i\,|\,\tau) - \xi^2. \tag{25}$$

The square root of Equation 25 is the conditional scale-score SEM for examinees with the true scale score ξ. The overall scale-score error variance for the entire population is

$$\sigma^2(E_S) = \int_{-\infty}^{\infty} \sigma^2(S\,|\,\tau) u(\tau) d\tau. \tag{26}$$

A reliability coefficient for scale scores can be defined as

$$\rho^2(S) = \frac{\sigma^2(\xi)}{\sigma^2(\xi) + \sigma^2(E_S)}. \tag{27}$$

Equation 27 can be viewed as a dependability coefficient for scale scores.

4.3 Estimating Classification Consistency and Accuracy

The strong true-score theory models for dichotomous items (Keats & Lord, 1962; Lord, 1965) have been used successfully for estimating classification consistency and accuracy (Huynh, 1976; Hanson, 1991; Hanson & Brennan, 1990; Lee et al., 2002). For tests that consist of polytomously-scored items, only a few procedures have been reported (Woodruff & Sawyer, 1989; Livingston & Lewis, 1995; Brennan & Wan, 2004; Lee, 2005b). The Dirichlet-multinomial model provides a theoretical framework for estimating various classification consistency and accuracy indices for tests composed of polytomous items. The classification consistency indices are estimated in this paper based on two hypothetical administrations of the test.[1]

4.3.1 Classification Consistency

Suppose classification decisions are made based on $H-1$ cutoff scores, $c_1, c_2, \ldots, c_{H-1}$, where H is the number of classification categories. Let $c_0 = min(Y)$ and $c_H = max(Y)$. The conditional bivariate distribution of the total raw scores, Y_1 and Y_2, for two independent administrations is given by

$$\Pr(Y_1 = i, Y_2 = j \,|\, \tau) = \Pr(Y_1 = i \,|\, \tau) \Pr(Y_2 = j \,|\, \tau). \tag{28}$$

The marginal bivariate distribution for Y_1 and Y_2 is given by

$$\Pr(Y_1 = i, Y_2 = j) = \int_{-\infty}^{\infty} \Pr(Y_1 = i \,|\, \tau) \Pr(Y_2 = j \,|\, \tau) u(\tau) d\tau. \tag{29}$$

The probability of consistent classification for each category, p_h $(h = 1, \ldots, H)$, can be computed by

$$\begin{aligned} p_h &= \Pr(c_{(h-1)} \leq Y_1 < c_h,\, c_{(h-1)} \leq Y_2 < c_h) \\ &= \sum_{i=c_{(h-1)}}^{c_h - 1} \sum_{j=c_{(h-1)}}^{c_h - 1} \Pr(Y_1 = i, Y_2 = j). \end{aligned} \tag{30}$$

Note that Equation 30 assumes integer scores for Y_1 and Y_2, which is only for the sake of convenience, and p_H includes the upper limit c_H. The overall

[1] Some researchers consider classification consistency indices that are computed based on the actual data and a hypothetical (i.e., model-predicted) administration of the test (e.g., Livingston & Lewis, 1995; Brennan & Wan, 2004; Lee, 2005b).

classification consistency index ϕ is computed by

$$\phi = \sum_{h=1}^{H} p_h. \tag{31}$$

The probability of consistent classification by chance is

$$\begin{aligned} \phi_c &= \sum_{h=1}^{H} \left[\sum_{i=c_{(h-1)}}^{c_h-1} \Pr(Y_1 = i) \right] \left[\sum_{j=c_{(h-1)}}^{c_h-1} \Pr(Y_2 = j) \right] \\ &= \sum_{h=1}^{H} \left[\sum_{i=c_{(h-1)}}^{c_h-1} \Pr(Y = i) \right]^2, \end{aligned} \tag{32}$$

with an exception that the upper limit of the Hth category in the summations is c_H. Coefficient κ is given by

$$\kappa = \frac{\phi - \phi_c}{1 - \phi_c}. \tag{33}$$

4.3.2 Classification Accuracy

While the classification consistency indices are based on the bivariate distribution of observed scores for two administrations of a test, the classification accuracy indices are based on the bivariate distribution of the observed and true score distributions (Lee et al., 2002). Note that the observed-score distribution is discrete, while the true-score distribution is continuous.

Let $\lambda_1, \lambda_2, \ldots, \lambda_{H-1}$ be true cutoff scores. Also let $\lambda_0 = min(\tau)$ and $\lambda_H = max(\tau)$. The probability of accurate classification for each category, q_h ($h = 1, \ldots, H$), can be computed by

$$\begin{aligned} q_h &= \int_{\lambda_{(h-1)}}^{\lambda_h} \Pr(c_{(h-1)} \leq Y < c_h \,|\, \tau) u(\tau) d\tau \\ &= \int_{\lambda_{(h-1)}}^{\lambda_h} \sum_{i=c_{(h-1)}}^{c_h-1} \Pr(Y = i \,|\, \tau) u(\tau) d\tau, \end{aligned} \tag{34}$$

where the Hth category includes the upper limit c_H. The overall classification accuracy index γ is given by

$$\gamma = \sum_{h=1}^{H} q_h. \tag{35}$$

The false positive and false negative error rates, respectively, for category h are

$$\begin{aligned} q_h^+ &= \int_{\lambda_{(h-1)}}^{\lambda_h} \Pr(c_h \leq Y \leq c_H \,|\, \tau) u(\tau) d\tau \\ &= \int_{\lambda_{(h-1)}}^{\lambda_h} \sum_{i=c_h}^{c_H} \Pr(Y = i \,|\, \tau) u(\tau) d\tau, \end{aligned} \tag{36}$$

and

$$\begin{aligned} q_h^- &= \int_{\lambda_{(h-1)}}^{\lambda_h} \Pr(c_0 \le Y < c_{(h-1)} \,|\, \tau) u(\tau) d\tau \\ &= \int_{\lambda_{(h-1)}}^{\lambda_h} \sum_{i=c_0}^{c_{(h-1)}-1} \Pr(Y = i \,|\, \tau) u(\tau) d\tau. \end{aligned} \tag{37}$$

The overall false positive and false negative error rates are

$$\gamma^+ = \sum_{h=1}^{H} q_h^+, \tag{38}$$

and

$$\gamma^- = \sum_{h=1}^{H} q_h^-. \tag{39}$$

If classification decisions are based on transformed scale scores using a set of scale-score cutoffs, the exact same procedures can be used as for raw scores. $\Pr(Y = i \,|\, \tau)$ and $\Pr(Y = i)$ are replaced with $\Pr(S = i \,|\, \tau)$ and $\Pr(S = i)$. Note that the results for the raw and scale scores will be identical if the scale-score cutoffs are determined directly from the corresponding raw-score cutoffs based on the conversion function, $v(y)$, and the conversion function is one-to-one at the cutoff score points. If several raw-score values including a raw-score cutoff convert to a single scale-score value, the results for the raw and scale scores will not be the same (Lee, 2005b).

5 Real Data Examples

The Dirichlet-multinomial model is illustrated in three application areas discussed in the previous section using four real data sets. As a preliminary analysis, the Dirichlet-multinomial model approach was applied to a test that consists of dichotomously-scored (0 and 1) items. The results were found to be identical to those from the two-parameter beta-binomial model (Hanson & Brennan, 1990) approach, which is a special case of the Dirichlet-multinomial model when the number of score categories is two.

The integration in the equations in the previous section may cause some practical complexities. The approach employed in this paper to evaluating the integrals is based on a simulation procedure. (Note that the choice of quadrature points and weights is treated in the discussion section.) With the estimates of α_i, the Dirichlet distribution is fully defined. A large number of Dirichlet random vectors are generated using the algorithms for the "beta method" as discussed in Narayanan (1990). Then, the integrals are replaced by summations (divided by N) over the random vectors. A value of $N = 10,000$ was used. Note that the sample size used in the generated sample was different from that for the actual data.

5.1 Data Sources

The first data set was obtained from a reading test that consists of 11 polytomous items ($n = 11$), each of which is scored 0, 1, and 2 ($k = 3$). About 2000 examinees took the test ($N = 2223$). Since the original test did not have scale scores, a pseudo conversion table was created so that the raw scores ranging from 0 to 22 convert to 100–400 scale scores. A pseudo cutoff score was established for two achievement levels: 17 in the raw-score metric and 340 in the scale-score metric.

The second data set was from a mathematics test that consists of 17 polytomous items ($n = 17$), each of which is scored 0, 1, and 2 ($k = 3$). The sample size was 2300 ($N = 2300$). A pseudo conversion table was created so that the raw scores ranging from 0 to 34 convert to 100–400 scale scores. Pseudo cutoff scores were established for three achievement levels: 15 and 27 in the raw-score metric; 200 and 360 in the scale-score metric.

The third data set was obtained from a large-scale writing assessment. The data set contain 5000 examinees' ($N = 5000$) responses to 12 items ($n = 12$), each of which is rated on a 0–5 integer scale ($k = 6$). The raw scores ranging from 0 to 60 are converted to integer scale scores ranging from 0 to 5. Pseudo cutoff scores were established for three achievement levels: 33 and 57 in the raw-score metric; 2 and 4 in the scale-score metric.

The fourth data set was from a large-scale licensure test that consists of 10 essay items ($n = 10$), each of which is rated on a 1–5 integer scale ($k = 5$). The sample size was the smallest among the four tests ($N = 650$). Pseudo cutoff scores were established for four achievement levels: 20, 30, and 40 in the raw-score metric. An identical transformation (i.e., scale scores are the same as raw scores) is used for computing scale-score results.

For all four tests, the cutoff scores are also treated as true cutoff scores for computing classification accuracy indices. Figure 1 displays the raw-to-scale score conversion functions for the four tests. Large dots on the conversion lines indicate the cutoff scores. Note that, except for the essay test, the scale-score cutoffs are located on the conversion functions where many raw-score points convert to a single scale score, which would result in different classification results for raw and scale scores.

5.2 Results

Summary statistics of the actual and fitted distributions for raw and scale scores are presented in Table 1 for the reading test. The summary statistics include the first four central moments, overall SEM, true-score variance, and reliability. The second column shows the statistics for the actual data, and the third through sixth columns exhibit the statistics for the fitted distributions with the C parameter estimated using all $k - 1$ categories (i.e., Equation 13) or a single category (i.e., Equation 17). The top portion of the table shows the raw-score results and the bottom portion of the table panel shows the scale-score results. The raw-score estimates of the overall SEM, true-score variance, and reliability

for the actual data are from the output of the GENOVA program (Crick & Brennan, 1983) for a generalizability theory analysis. These three statistics, $\hat{\sigma}(E_Y)$, $\hat{\sigma}^2(\tau)$, and $\hat{\rho}^2(Y)$, are identical, respectively, to the square root of absolute error variance, universe score variance, and the dependability coefficient, which result from using a p x i design.[2] The statistics for the actual data computed based on the traditional approach are not absolute criteria, but provide a reasonable basis for comparison. Note that $\hat{\sigma}(E_S)$, $\hat{\sigma}^2(\xi)$, and $\hat{\rho}^2(S)$, are not estimable (NE) in the generalizability theory analysis for the actual data.

Table 1: Summary Statistics for Reading

	Raw Score				
	Actual	All Cat	Cat 1	Cat 2	Cat 3
N	2223	10000	10000	10000	**10000**
Mean	12.41	12.42	12.41	12.40	**12.22**
SD	4.69	3.82	3.94	3.28	**4.53**
Skew	-.16	-.15	-.16	-.11	**-.15**
Kurt	2.43	2.79	2.77	2.84	**2.65**
$\hat{\sigma}(E_Y)$	2.12	2.28	2.26	2.37	**2.17**
$\hat{\sigma}^2(\tau)$	18.34	9.37	10.42	5.16	**15.82**
$\hat{\rho}^2(Y)$	.80	.64	.67	.48	**.77**
	Scale Score				
	Actual	All Cat	Cat 1	Cat 2	Cat 3
N	2223	10000	10000	10000	**10000**
Mean	277.35	277.84	277.65	277.58	**274.55**
SD	69.31	57.29	58.94	50.02	**66.99**
Skew	-.20	-.19	-.19	-.15	**-.21**
Kurt	2.27	2.56	2.56	2.60	**2.49**
$\hat{\sigma}(E_S)$	NE	34.55	34.12	36.35	**32.59**
$\hat{\sigma}^2(\xi)$	NE	2088.22	2309.88	1181.13	**3425.40**
$\hat{\rho}^2(S)$	NE	.64	.66	.47	**.76**

Table 1 shows that the four sets of estimates produce somewhat different results. The estimates based on the third category appear to provide summary statistics that are closest to the actual ones, in general. Thus, for the reading test, the parameter estimates based on the third category are selected for both

[2]The results reported in GENOVA are in the mean-score metric, while those reported in this paper are in the total-score metric. The mean-score results can be easily transformed to the total-score metric by multiplying the variance terms by n^2.

raw and scale scores to compute conditional SEMs and classification consistency and accuracy indices. Note, however, that the choice of estimates does not alter the results substantially. The statistics from the selected estimation are boldfaced in this and all subsequent tables.

The actual and fitted raw- and scale-score distributions for the reading test are plotted in Figures 2 and 3. Consistent with Table 1, the fitted distribution with the third category appears to provide the best fit to the actual data. The fitted distributions with all categories and the second category do not seem to provide an adequate fit. Note in Figure 2 that the fitted distribution with the third category is not the smoothest one, even if it provides the best fit. If the purpose of using the Dirichlet-multinomial model is to smooth the observed-score distribution, the fitted distribution with the second category might be preferred.[3] However, if the purpose is to compute various psychometric properties such as conditional SEMs and classification indices, the choice would be the estimates from the third category. The scale-score distributions shown in Figure 3 are more irregular than the raw-score distributions due to the characteristic of the many-to-one conversion—i.e., several raw-score points are collapsed into a single scale-score value.

Table 2: Parameter Estimates and Variances of Category Scores for Reading

	Parameter Estimates			
	All Cat	Cat 1	Cat 2	Cat 3
$\hat{\alpha}_1$	1.37	1.19	2.62	**0.69**
$\hat{\alpha}_2$	2.66	2.33	5.11	**1.35**
$\hat{\alpha}_3$	2.16	1.89	4.14	**1.09**

	Variances of Category Scores				
	Actual	All Cat	Cat 1	Cat 2	Cat 3
X_1	4.84	4.49	4.81	3.36	**6.68**
X_2	4.79	6.46	6.92	4.80	**9.20**
X_3	8.54	5.97	6.38	4.44	**8.46**

To further explore the functionality of estimation with different categories, the estimated Dirichlet parameters for the reading test are summarized in Table 2. Also displayed in Table 2 are variances of the marginal category scores based on the actual data and those estimated based on the model (see Equation 7). The selected parameter estimates tend to be smaller than those for the other sets. Note also that the ratio of the $\hat{\alpha}_i$ values is always the same across all estimation sets. For example, $\alpha_1/\alpha_2 = .515$ using all categories or any one of the three categories.

[3] A smoothed-line criterion may not be reasonable, however, if the observed-score distribution is not guaranteed to be smooth when $N \to \infty$.

The means of the category scores for the actual data are always perfectly recovered by any of the category estimations based on Equation 6, and thus not reported in this paper. The perfect recovery of the category means is expected from Equation 15, which shows that each individual α parameter for category i is estimated by equating the parameter to the sample mean of the same category. A category with a larger sample mean, therefore, is always associated with a larger α estimate. Whereas, the variance for each category is closely recovered only if the estimation is based on the same category. For example, the actual variance for the first category is closest to the estimated variance when the parameters are estimated based on the first category, etc. This is due to the fact that the estimation with a particular category uses the first two moments (i.e., mean and variance) of the category (not all categories) as shown in Equation 17. Since the estimation with all categories uses information from all categories simultaneously, the estimated variance for each category is somewhere in the middle of the other variance estimates based on single category estimations. In other words, the estimation with all categories provides estimated variances that are close to the actual ones as a whole, but not closest to any single one compared to those based on single category estimations. Finally, note that the variance of the marginal distribution tends to be estimated better, if estimation is based on the category with the largest variance. For the reading test, the variance for the third category is largest, contributing the most to the total variance, which would lead to a better estimate of variance of the total-test scores (see Table 1). This is not the case, however, for the writing test as discussed later, in which the estimation based on a category associated with excessively large variance could overestimate the variance of the total-test scores.

The estimated classification consistency and accuracy indices for the reading test using the third category are summarized in Table 3. Note that the results for raw and scale scores are slightly different. Although not reported in this paper, when different categories are used for estimation, the largest difference between the estimated ϕ coefficients for raw scores, for example, is .01.

Table 3: Classification Indices for Reading Using the Third Category

	Raw Score	**Scale Score**
Consistency		
ϕ	.86	.84
ϕ_c	.70	.63
κ	.54	.56
Accuracy		
γ	.90	.87
γ^+	.08	.12
γ^-	.02	.02

Depicted in Figure 4 are estimated conditional SEMs for the reading test using the third category for estimation. The top two plots display conditional raw- and scale-score SEMs for 10,000 examinees, and the bottom two plots show the

fitted conditional SEMs using a high degree (p=4) polynomial regression (more correctly, the square root of the fitted conditional error variances). It is well known that the conditional raw-score SEMs for dichotomous items estimated from the binomial error model produce a smooth concave-down quadratic line. By contrast, the estimated conditional SEMs for polytomous items estimated under the multinomial error model show a scattered pattern, which is consistent with the findings in a previous study (i.e., Lee, 2005a). The conditional SEMs given a true score can vary depending upon the configuration of the item scores.

It is worth noting that the contour line of the conditional SEMs is a concave-down quadratic curve much like the dichotomous case, and that the width of the vertically scattered SEMs is larger in the middle. Since different combinations of item scores give different SEM values for a particular true score, the pattern of widths of the vertically scattered SEMs seems to be related to the pattern of the numbers of possible item combinations that lead to a particular raw-score point as displayed in Figure 5. The relative number of all possible item combinations is larger to achieve total raw-score values located in the middle than in both extremes.

In Figure 4, the conditional SEMs at true-score values near 0, 11, and 22 tend to approach zero. This is due to the fact that, under the multinomial error model, the estimated SEM is necessarily zero when an examinee gets the same scores for all items. Since there are 11 items scored 0, 1, and 2 for the reading test, all 0's, all 1's, and all 2's lead to a total raw score of 0, 11, and 22, respectively. Therefore, the pattern of conditional SEMs for polytomous items will be almost always an umbrella shape with the number of umbrella ribs equal to the number of score categories.

The conditional scale-score SEMs show a pattern similar to that of the raw-score SEMs, which is mainly due to an approximate linear pattern of the raw-to-scale score transformation. From the perspectives of the delta method (Agresti, 1990), the magnitude of the approximated scale-score SEMs is proportionally related to the degree of slope of the transformation (see also Lee, 2005a; Brennan & Lee, 1999).

The fitted polynomial conditional SEMs can be used when it is required to report a single SEM value at any given true score. Although the choice of the polynomial degree requires a subjective judgment, evaluating R-square values for several different polynomial degrees can be helpful. The fitted conditional SEMs tend to be flat in the middle, as opposed to the bell-shaped SEMs for dichotomous items. This is not the case for the math test, however, as discussed next.

The summary statistics for the math test are displayed in Table 4. The results reported in Table 4 and the fitted distributions shown in Figures 6 and 7 together suggest that the estimation with the first category be selected. Note that the fitted raw-score distribution with the first category in Figure 6 is rather a bumpy function.[4] Depending upon the purpose of data analyses, estimation

[4]Currently, the reason for a bumpy fitted distribution from estimation with a particular category is not fully understood. More investigation is needed.

Table 4: Summary Statistics for Math

	Raw Score				
	Actual	All Cat	Cat 1	Cat 2	Cat 3
N	2300	10000	**10000**	10000	10000
Mean	12.40	12.42	**12.41**	12.45	12.43
SD	5.98	4.82	**6.17**	4.38	5.10
Skew	.19	.25	**.36**	.21	.28
Kurt	2.52	2.83	**2.70**	2.86	2.80
$\hat{\sigma}(E_Y)$	3.31	3.42	**3.29**	3.46	3.40
$\hat{\sigma}^2(\tau)$	28.55	11.54	**27.31**	7.21	14.48
$\hat{\rho}^2(Y)$	.72	.50	**.72**	.38	.56
	Scale Score				
	Actual	All Cat	Cat 1	Cat 2	Cat 3
N	2300	10000	**10000**	10000	10000
Mean	190.77	187.87	**191.14**	186.78	188.64
SD	65.98	55.25	**68.22**	50.59	58.07
Skew	.71	.87	**.79**	.89	.84
Kurt	2.58	3.12	**2.68**	3.34	3.00
$\hat{\sigma}(E_S)$	NE	39.90	**37.43**	40.49	39.53
$\hat{\sigma}^2(\xi)$	NE	1460.34	**3252.19**	919.75	1809.86
$\hat{\rho}^2(S)$	NE	.48	**.70**	.36	.54

with a different category could be chosen.

The parameter estimates and variances of category scores for the math test are shown in Table 5. As for the reading test, the selected parameter estimates based on the first category are smallest. The variance for each category for the actual data is almost perfectly recovered by estimation with the same corresponding category. Note also that the actual variance of the first category is largest, and the selected estimation is based on the first category as well.

The estimates of classification indices for the math test are tabulated in Table 6. All other things being equal, the classification consistency and accuracy indices (ϕ, κ, and γ) will be lower when more classification levels are involved, which is clearly the case for the reading and math tests. The math test has three classification categories, whereas the reading test has only two, and thus the math test tends to provide relatively lower consistency and accuracy.

The conditional SEMs for the math test using the first category are depicted in Figure 8. The shape of the conditional SEMs for the math test are very

Table 5: Parameter Estimates and Variances of Category Scores for Math

	Parameter Estimates			
	All Cat	Cat 1	Cat 2	Cat 3
$\hat{\alpha}_1$	9.40	**3.60**	15.08	7.25
$\hat{\alpha}_2$	3.47	**1.33**	5.57	2.68
$\hat{\alpha}_3$	4.65	**1.78**	7.46	3.59
	Variances of Category Scores			
	Actual	All Cat	Cat 1	Cat 2	Cat 3
X_1	13.00	7.88	**13.00**	6.55	8.89
X_2	4.19	5.03	**8.24**	4.19	5.66
X_3	6.97	6.19	**10.23**	5.16	7.00

Table 6: Classification Indices for Math Using the First Category

	Raw Score	**Scale Score**
Consistency		
ϕ	.76	.75
ϕ_c	.53	.50
κ	.48	.49
Accuracy		
γ	.82	.81
γ^+	.12	.12
γ^-	.06	.07

similar to that for the reading test, except that, for the math test, the umbrella ribs are shorter and sufficient data do not exist at the right extreme. The sparse data at the right extreme is a direct consequence of the estimated true-score distribution, which turns out to be skewed to the right. If an infinitely large number of Dirichlet random vectors would had been used, the umbrella ribs would get longer approaching zero and the sparseness of the data would diminish. Even with a larger sample size, the fitted SEM functions may not change substantially because the relative densities of SEMs across the entire true-score range may not change.

The shape of the fitted SEMs is similar to the quadratic function of the dichotomous-item case. Recall that the fitted SEMs for the reading test showed a rather flat pattern. The difference between the two tests stems from the differences in the estimated true-score distributions (or, π_i parameter estimates). Compared to the raw-score results, the scale-score results show relatively larger conditional SEMs in the middle due to the large slope of the transformation in the middle as shown in Figure 1. Note also that the fitted scale-score SEM

function drops down to zero around the true scale score of 360, which appears to be caused by the sparseness of data at the right extreme. The same observation is made for the fitted raw-score SEMs for the two tests discussed next.

The summary statistics for the writing test are presented in Table 7. Since items in the writing test have six score categories, there are seven possible sets of estimates. For this test, choosing a set of estimates is a difficult task. In fact, most estimates seem to differ substantially from the statistics based on the actual data. In particular, compared to the generalizability theory results, the estimated overall SEMs are too large and the estimated true-score variances are too small, which, in turn, results in too small reliability estimates. The relatively inferior estimates for the writing test seem to be due to the "large" number of score categories ($k = 6$). Careful evaluation of both the summary statistics and fitted distributions exhibited in Figures 9 and 10 led to the choice of the estimates using the fifth category to compute conditional SEMs and classification indices.

Table 7: Summary Statistics for Writing

	Raw Score							
	Actual	All Cat	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5	Cat 6
N	5000	10000	10000	10000	10000	10000	**10000**	10000
Mean	37.74	34.39	34.59	33.33	29.22	34.42	**33.67**	35.25
SD	8.10	7.35	7.11	8.37	11.68	7.33	**8.04**	5.90
Skew	-.43	-.33	-.31	-.29	-.19	-.33	**-.32**	-.29
Kurt	2.77	3.61	3.60	3.46	2.82	3.61	**3.50**	3.62
$\hat{\sigma}(E_Y)$	2.15	3.88	3.87	3.95	3.95	3.87	**3.93**	3.76
$\hat{\sigma}^2(\tau)$	61.81	39.02	35.58	54.36	120.90	38.68	**49.18**	20.68
$\hat{\rho}^2(Y)$	.93	.72	.70	.78	.89	.72	**.76**	.59
	Scale Score							
	Actual	All Cat	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5	Cat 6
N	5000	10000	10000	10000	10000	10000	**10000**	10000
Mean	2.86	2.58	2.60	2.49	2.16	2.58	**2.52**	2.65
SD	.70	.68	.66	.76	1.01	.68	**.73**	.57
Skew	-.17	-.29	-.29	-.24	-.15	-.29	**-.27**	-.37
Kurt	2.94	3.25	3.22	3.25	2.75	3.25	**3.26**	3.01
$\hat{\sigma}(E_S)$	NE	.43	.43	.43	.43	.43	**.43**	.43
$\hat{\sigma}^2(\xi)$	NE	.28	.25	.38	.83	.60	**.35**	.15
$\hat{\rho}^2(S)$	NE	.60	.58	.67	.82	.55	**.65**	.45

Implicit in Table 7, the fitted distributions, especially for raw scores, do not seem to provide adequate fits. Except for the third category, the fitted distributions are very similar to each other. Note that the results from using the 6th category are not shown in the figures.

Table 8: Parameter Estimates and Variances of Category Scores for Writing

	Parameter Estimates						
	All Cat	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5	Cat 6
$\hat{\alpha}_1$	0.02	0.02	0.01	0.00	0.02	**0.01**	0.03
$\hat{\alpha}_2$	0.04	0.04	0.03	0.01	0.04	**0.03**	0.07
$\hat{\alpha}_3$	0.84	0.92	0.57	0.22	0.85	**0.64**	1.50
$\hat{\alpha}_4$	1.59	1.74	1.09	0.42	1.61	**1.22**	2.85
$\hat{\alpha}_5$	1.03	1.13	0.71	0.27	1.04	**0.79**	1.85
$\hat{\alpha}_6$	0.24	0.26	0.16	0.06	0.24	**0.18**	0.42
	Variances of Category Scores						
	Actual	All Cat	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5	Cat 6
X_1	0.15	2.17	1.98	3.25	8.96	2.14	**2.89**	1.19
X_2	0.52	2.55	2.35	3.73	8.80	2.53	**3.29**	1.47
X_3	13.57	6.40	6.03	8.16	14.24	6.34	**7.51**	4.76
X_4	9.65	9.33	8.89	11.22	15.36	9.27	**10.64**	6.92
X_5	9.17	6.70	6.47	7.62	9.23	6.68	**7.34**	5.27
X_6	1.72	3.00	2.81	3.93	6.77	2.98	**3.58**	1.91

Table 8 displays the parameter estimates and variances of category scores for the writing test. All the parameter estimates for the writing test tend to be very small. Except for the estimates based on the third category that are relatively too big, the estimates based on other categories appear to be similar to each other, which is consistent with the observation made in Figure 9 in terms of the fitted distributions. The estimated variance for each category is close to the actual one when estimated based on the same corresponding category, in general. However, the estimated variances for X_1 and X_2 based on the first and second category estimation, respectively, substantially differ from the actual ones. This appears to stem from the lack of frequency data for the first two categories. A larger sample size would be required in order to have sufficient frequencies for all category scores. Doubled with the data sparseness for lower categories, the unduly large variance for the third category seems to be the cause of the relatively poor fits of the marginal distribution. Even with the estimates based on the third category does not produce the best fitted marginal distribution. Recall that for the reading and math tests, the estimates based

on the category associated with the largest variance produced the best fitted distribution.

Table 9 summarizes the estimated classification indices for the writing test using the fifth category. One interesting observation is that the scale-score ϕ is larger than the raw-score ϕ. It is often stated that reliability for scale scores is lower than that for raw scores when the number of the scale-score points is less than that of the raw-score points (Kolen et al., 1992), which is also confirmed with the current example as shown in Table 7. However, the ϕ coefficient for scale scores is not necessarily lower than that for raw scores. The lower value of κ and the higher value of ϕ_c for scale scores are consistent with the convention though.

Table 9: Classification Indices for Writing Using the Fifth Category

	Raw Score	**Scale Score**
Consistency		
ϕ	.79	.87
ϕ_c	.51	.77
κ	.56	.46
Accuracy		
γ	.84	.82
γ^+	.09	.17
γ^-	.07	.01

The conditional SEMs for the writing test using the fifth category are plotted in Figure 11. As discussed previously, the conditional SEMs show an umbrella shape with six ribs approaching zero. The sparseness of data is found in both ends of the score scale, but more sparseness in the right end. The fitted scale-score SEM function tends to be very flat.

Table 10: Summary Statistics for Essay

	Raw Score						
	Actual	All Cat	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5
N	650	10000	**10000**	10000	10000	10000	10000
Mean	32.05	32.01	**31.38**	32.02	32.06	32.02	32.04
SD	6.51	4.58	**5.55**	4.66	3.87	4.58	5.04
Skew	-.39	-.11	**-.12**	-.11	-.07	-.11	-.12
Kurt	2.70	2.99	**2.97**	3.00	2.93	2.99	3.01
$\hat{\sigma}(E_Y)$	3.35	3.69	**3.66**	3.67	3.77	3.69	3.61
$\hat{\sigma}^2(\tau)$	32.01	7.37	**17.37**	8.25	.76	7.37	12.40
$\hat{\rho}^2(Y)$	.74	.35	**.56**	.38	.05	.35	.49

The summary statistics for the essay test are contained in Table 10, and the actual and fitted raw-score distributions are shown in Figure 12. Note that only raw-score results are provided for the essay test. Except for the mean statistic, the other summary statistics based on the estimation procedures are somewhat different from those based on the actual data. This again appears to be associated with the number of score categories, which is five in this example. The underestimate of true-score variance seems to be the most salient problem. The fitted distributions are all smooth, and the one with the first category appears to provide a relatively better fit. Thus, estimation with the first category is used for further computation.

Table 11: Parameter Estimates and Variances of Category Scores for Essay

	Parameter Estimates					
	All Cat	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5
$\hat{\alpha}_1$	1.89	**0.72**	1.68	19.64	1.89	1.10
$\hat{\alpha}_2$	3.07	**1.16**	2.72	31.83	3.07	1.78
$\hat{\alpha}_3$	5.69	**2.15**	5.04	59.00	5.69	3.30
$\hat{\alpha}_4$	4.81	**1.82**	4.26	49.85	4.81	2.79
$\hat{\alpha}_5$	2.91	**1.10**	2.58	30.19	2.91	1.69
	Variances of Category Scores					
	Actual	All Cat	Cat 1	Cat 2	Cat 3	Cat 4	Cat 5
X_1	1.97	**1.36**	2.37	1.41	0.97	1.36	1.63
X_2	2.12	**2.05**	2.91	2.13	1.45	2.05	2.47
X_3	2.24	**3.13**	4.52	3.25	2.24	3.13	3.80
X_4	2.83	**2.83**	4.03	2.94	2.02	2.83	3.42
X_5	2.36	**1.94**	2.75	2.02	1.40	1.94	2.35

Table 11 presents the parameter estimates and variance of category scores for the essay test. For the essay test, almost the same observations can be made as for the other tests. The parameter estimates for the essay test using all categories versus the fourth category appear the same, but they are equal only up to the second decimal place. The estimated variance for the first category based on the estimation with the first category is dissimilar to the corresponding actual variance, which is, as discussed previously, due to sparse frequency data for the first category. Estimation based on the third category tends to yield estimates that are most dissimilar to those based on different categories, which is reflected in the fitted distributions in Figure 12.

The classification results for the essay test using the first category are summarized in Table 12. Since the number of classification categories is largest among the four tests ($H = 4$), the consistency and accuracy for the essay test

Table 12: Classification Indices for Essay Using the First Category

	Raw Score
Consistency	
ϕ	.63
ϕ_c	.46
κ	.32
Accuracy	
γ	.16
γ^+	.00
γ^-	.83

tend to be lower.

Depiceted in Figure 13 are the conditional SEMs for the essay test using the first category. For this example, the conditional SEMs do not show a clear umbrella shape anymore. It might be that the actual data do not contain many examinees with the same item scores for all ten items, which supposedly would influence the estimation procedure. The fitted SEM function tends to be flat in the middle, and there is clear evidence of data sparseness at both extremes.

6 Discussion

Due to the current popularity of various innovative item types, which typically are scored polytomously, there is a high demand for theory and methodologies that can effectively deal with multiple item-score categories. The present paper provides a strong true-score model that can be used with test scores based on polytomously-scored items. The Dirichlet-multinomial model can be viewed as a multivariate extension of Keats and Lord's (1962) two-parameter beta-binomial model. If items have only two score categories, the Dirichlet-multinomial model reduces to the two-parameter beta-binomial model. Under the strong-true score theory presented in this paper, errors conditional on true score are assumed to be distributed as multinomial, true category-proportion scores are assumed to be a Dirichlet distribution, and the marginal category scores are the Dirichlet-multinomial distribution.

The Dirichlet-multinomial model was applied to four different real data sets in three applications, which are likely to be the areas in which the model might be used routinely: (1)obtaining a smoothed fitted observed-score distribution, (2) estimating conditional standard errors of measurement for both raw and transformed scale scores, and (3) estimating classification consistency and accuracy indices for raw and scale scores. All statistics were successfully estimated for all four tests. The computation time using a random sample of 10,000 Dirichlet deviates was less than five minutes for all cases. It was found that estimation can be poor when the number of score categories for each item is large, say, greater than five. For the reading and math tests that consist of

items with three score categories, estimation seemed to work fairly well. It is likely that a larger sample size is required as the number of categories increases to gain accurate parameter estimates. A comprehensive simulation study would be necessary to study this.

It was noted that the means of all category scores are always perfectly recovered by estimation based on all categories or any one of them. It was also demonstrated that estimation based on a single category tended to fit variance of the corresponding category scores only, while estimation based on all categories tended to fit, on average, variances for all category scores. The use of all categories simultaneously in estimation appears to be most appealing, at least theoretically. However, the real data analyses employed in this paper revealed that various statistics, such as the moments for actual data, fitted marginal distributions, and variance components estimated using a generalizability theory analysis, would need to be examined in choosing a category for estimation.

The procedure employed in this paper for estimating the true-score distribution involved drawing a large sample of Dirichlet variables. There exist several different algorithms for generating Dirichlet deviates. Following the recommendations listed in Narayanan (1990), a generating procedure called the "beta method" was used in this paper, which is known to be efficient in terms of computation time when the α parameters are larger than 1.0 and the number of score categories is large. As another alternative to the random sampling approach, quadrature points and weights (Press, Teukolsky, Vetterling, & Flannery, 1992) could be used. For example, estimation might proceed as follows: (1) estimate α_i parameters to compute $\hat{g}(\pi_1, \ldots, \pi_k)$; (2) determine, say 10, equally-spaced quadrature points for each of the π_i parameters; (3) go through 10^k combinations of π_i values to find those combinations that satisfy the condition $\pi_1+\pi_2+\cdots+\pi_k = 1$; (4) for each π_i combination that satisfies the condition, compute statistics conditional on π_i as well as the density of the combination using $\hat{g}(\pi_1, \ldots, \pi_k)$; and (5) compute marginal statistics by replacing integrals with summations.

In estimating reliability, the formula employed the overall error variance in the numerator and the sum of the error and true-score variances in the denominator. Both of the error and true-score variances were estimated based on the model. The resultant reliability estimate is a close approximation to the dependability coefficient. Alternatively, the sample variance could be used in the denominator. One problem with the use of the sample variance, however, is that the sources of errors in the numerator and denominator will be different.

The fitted observed-score distributions reported in this paper were computed based on a Dirichlet random sample. The marginal observed-score distribution can also be computed using Equation 3 by replacing α_i parameters with estimates. When a sufficiently large sample size is used, these two approaches should produce similar results. Advantages of using a random sample for computing the fitted observed-score distribution include that it is computationally less burdensome, especially when the number of score categories is large, and that it is consistent with the other statistics computed based on the model.

7 References

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

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L., & Lee, W. (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. *Educational and Psychological Measurement*, *59*, 5–24.

Brennan, R. L., & Wan, L. (2004). *Bootstrap procedures for estimating decision consistency for single-administration complex assessments* (CASMA Research Report No. 7). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from www.education.uiowa.edu/casma).

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system* (American College Testing Technical Bulletin No. 43). Iowa City, IA: ACT, Inc.

Hanson, B. A. (1991). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes* (Research Report 91–5). Iowa City, IA: ACT, Inc.

Hanson, B. A., & Brennan, R. L. (1990). An investigation of classification consistency indexes estimated under alternative strong true score models. *Journal of Educational Measurement*, *27*, 345–359.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*, 253–264.

Johnson, N. L., & Kotz, S. (1969). *Distributions in statistics: Discrete distributions*. New York: John Wiley & Sons.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution of mental test scores. *Psychometrika*, *27*, 59–72.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, *29*, 285–307.

Lee, W. (2005a). *A multinomial error model for tests with polytomous items* (CASMA Research Report No. 10). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from www.education.uiowa.edu/casma).

Lee, W. (2005b). *Classification consistency under the compound multinomial model* (CASMA Research Report No. 13). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from www.education.uiowa.edu/casma).

Lee, W., Brennan, R. L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: A simulation study. *Journal of Educational Measurement*, *37*, 1–20.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement*, *26*, 412–432.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32*, 179–197.

Lord, F. M. (1965). A strong true score theory with applications. *Psychometrika*, *30*, 239–270.

Lord, F. M. (1969). Estimating true-score distributions in psychological testing (An empirical Bayes estimation problem). *Psychometrika*, *34*, 259–299.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Massachusetts: Addison-Wesley Publishing Company.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate -distribution, and correlation among proportions. *Biometrika*, *49*, 65–82.

Narayanan, A. (1990). Computer generation of Dirichlet random vectors. *Journal of Statistical Computation and Simulation*, *36*, 19–30.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in Fortran* (2nd ed.). New York: Cambridge University Press.

Woodruff, D. J., & Sawyer, R. L. (1989). Estimating measures of pass-fail reliability from parallel half-tests. *Applied Psychological Measurement*, *13*, 33–43.

A Moments of the Beta-Binomial Distribution

When $k = 2$ (0 and 1 with 1 being the second category), the Dirichlet-multinomial model reduces to the two-parameter beta-binomial model with parameters α and β. From Equations 6 and 7, the moments of the two-parameter beta-binomial model for the number-correct category are given by

$$\mathcal{E}(X) = \frac{n\beta}{\alpha+\beta}, \tag{40}$$

and

$$\sigma^2(X) = \left[\frac{n+\alpha+\beta}{1+\alpha+\beta}\right] n\mathcal{E}(\mathit{\Pi})[1-\mathcal{E}(\mathit{\Pi})]. \tag{41}$$

The first two moments of the beta distribution are

$$\mathcal{E}(\Pi) = \frac{\beta}{\alpha+\beta}, \tag{42}$$

and

$$\sigma^2(\Pi) = \frac{\alpha\beta}{(1+\alpha+\beta)(\alpha+\beta)^2}. \tag{43}$$

Thus, Equation 41 can be rewritten as

$$\begin{aligned}\sigma^2(X) &= \left[\frac{n+\alpha+\beta}{1+\alpha+\beta}\right]\left[\left(\frac{n\beta}{\alpha+\beta}\right)\left(1-\frac{\beta}{\alpha+\beta}\right)\right] \\ &= \frac{n\alpha\beta(n+\alpha+\beta)}{(1+\alpha+\beta)(\alpha+\beta)^2}.\end{aligned} \tag{44}$$

From Equation 43, the true-score variance in the total-score metric is given by

$$\sigma^2(\tau) = \frac{n^2\alpha\beta}{(1+\alpha+\beta)(\alpha+\beta)^2}. \tag{45}$$

From Equation 41, an estimator of $\sigma^2(X)$ is

$$\hat{\sigma}^2(X) = \left[\frac{n+\alpha+\beta}{1+\alpha+\beta}\right]\left[\frac{\bar{x}(n-\bar{x})}{n}\right], \tag{46}$$

where $\bar{x}$ is the mean total score over all examinees in the sample. The term, $\bar{x}(n-\bar{x})/n$, in Equation 46 is the biased error variance for an examinee with total score of $\bar{x}$ as an estimate of $\mathcal{E}(\Pi)$.

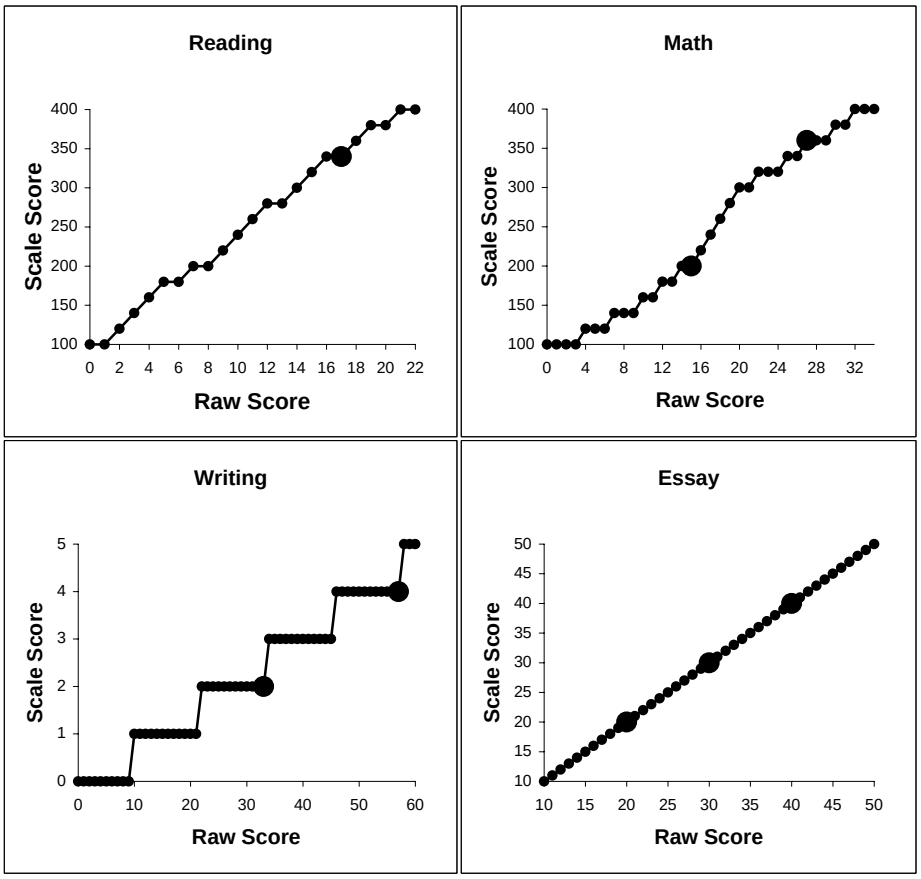

Figure 1: Raw-to-scale score Transformation Functions

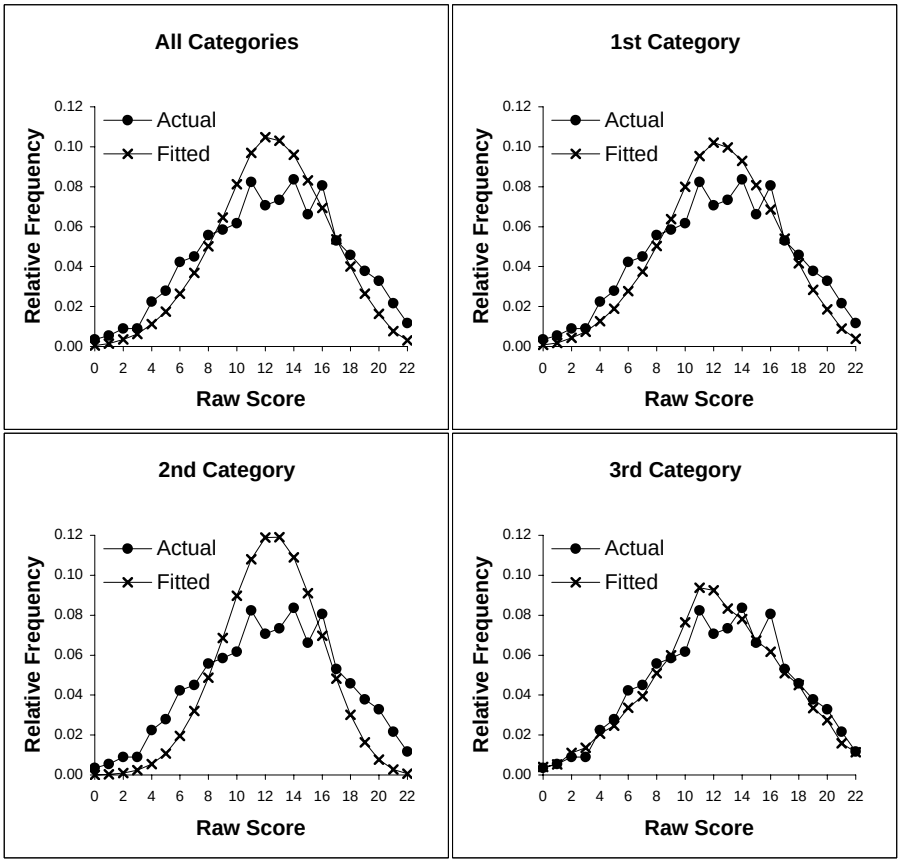

Figure 2: Actual and Fitted Raw-score Distributions for Reading

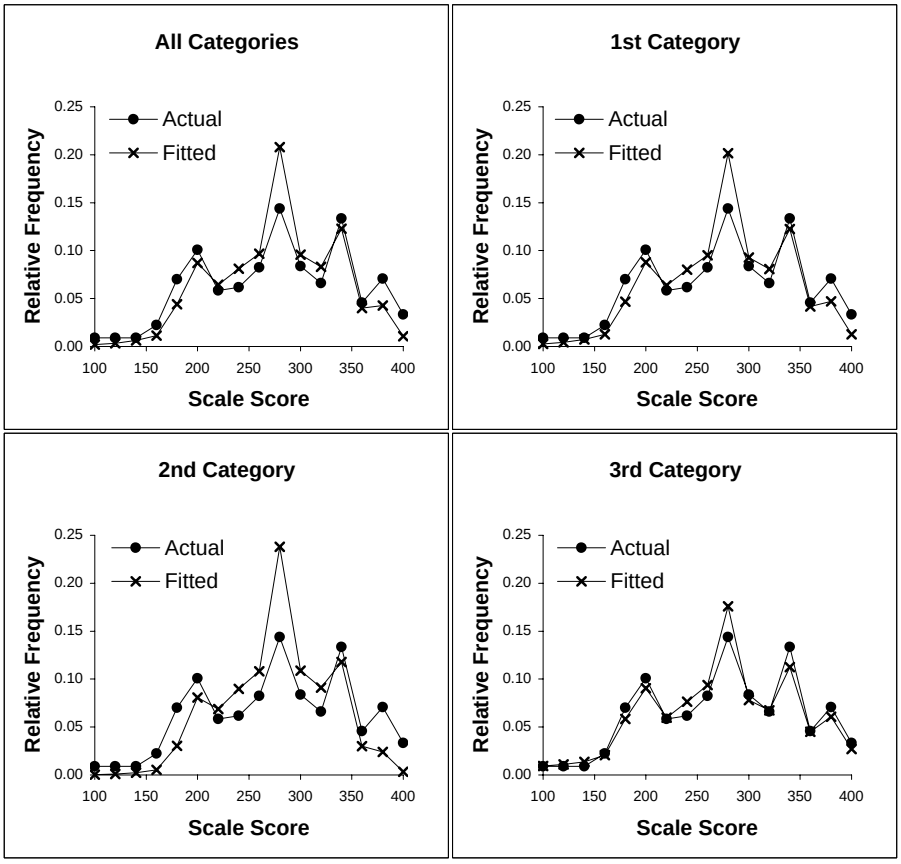

Figure 3: Actual and Fitted Scale-score Distributions for Reading

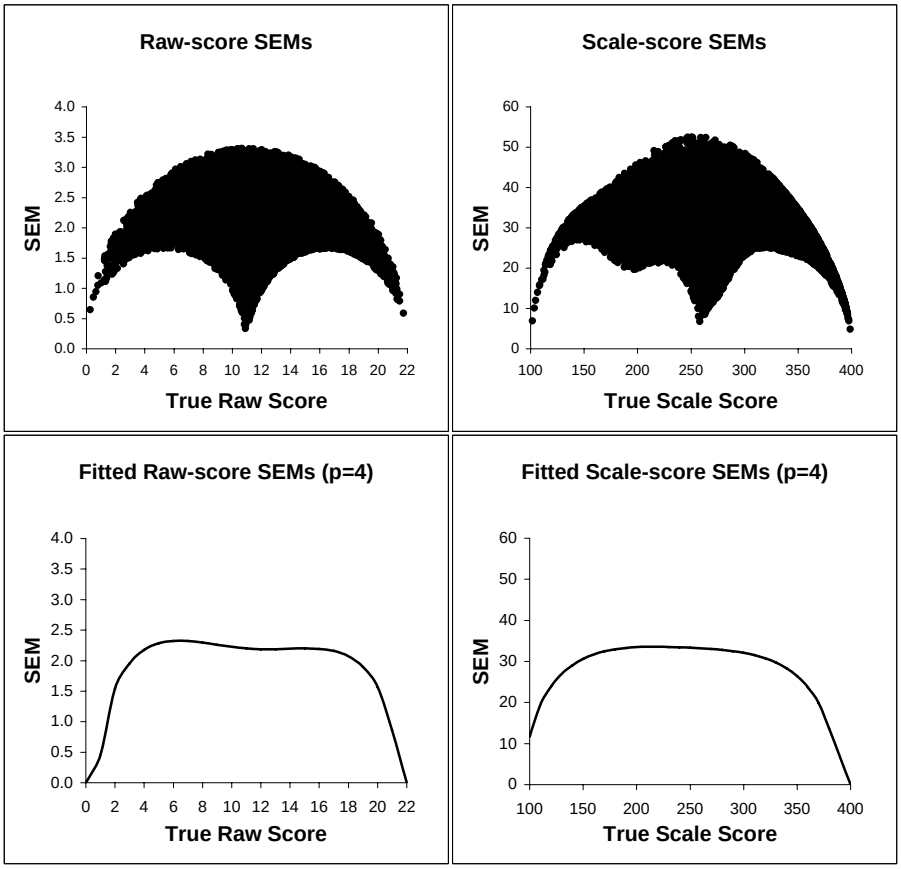

Figure 4: Conditional SEMs for Reading Using the Third Category

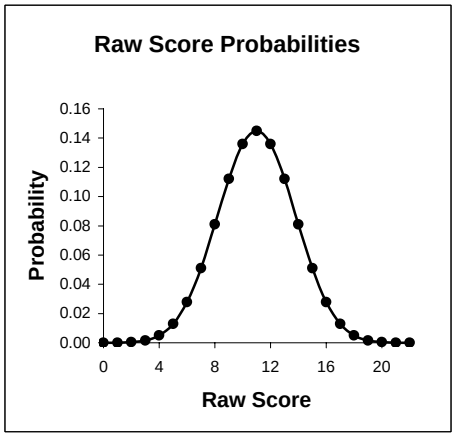

Figure 5: Number of Item Combinations Leading to Total Raw-score Points

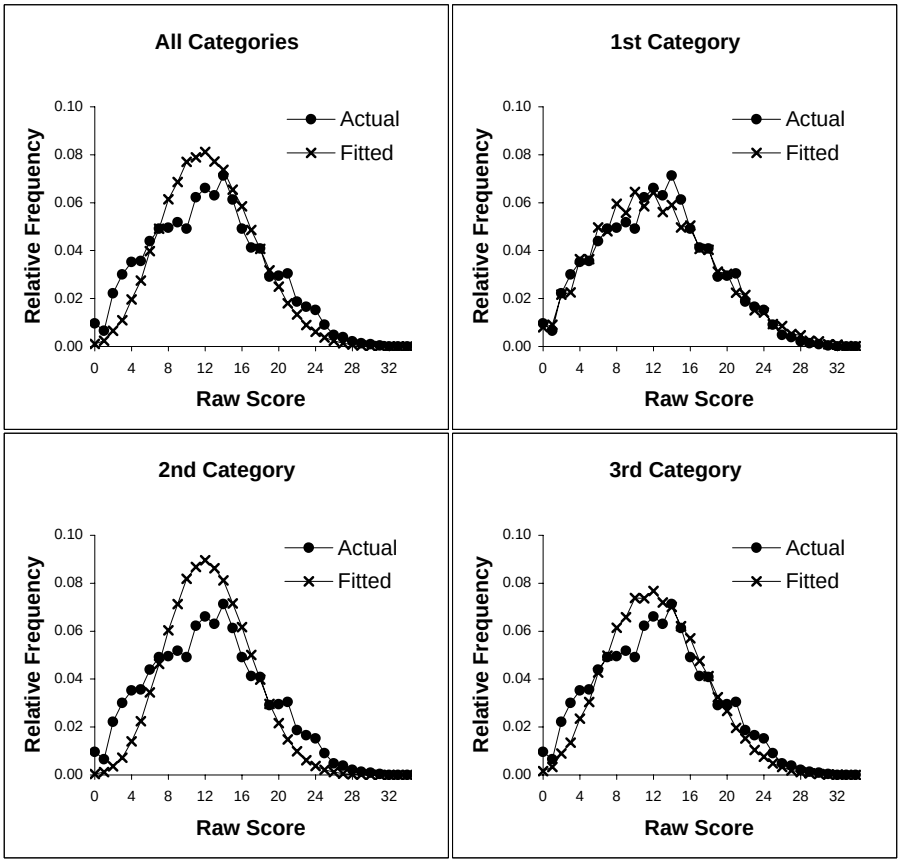

Figure 6: Actual and Fitted Raw-score Distributions for Math

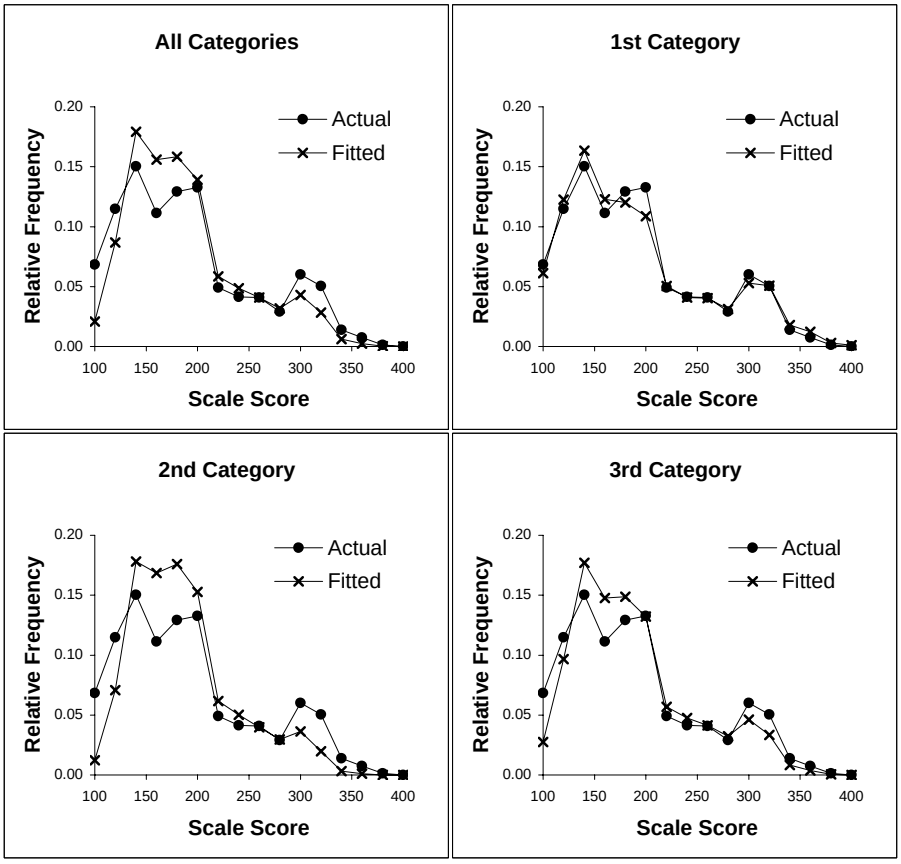

Figure 7: Actual and Fitted Scale-score Distributions for Math

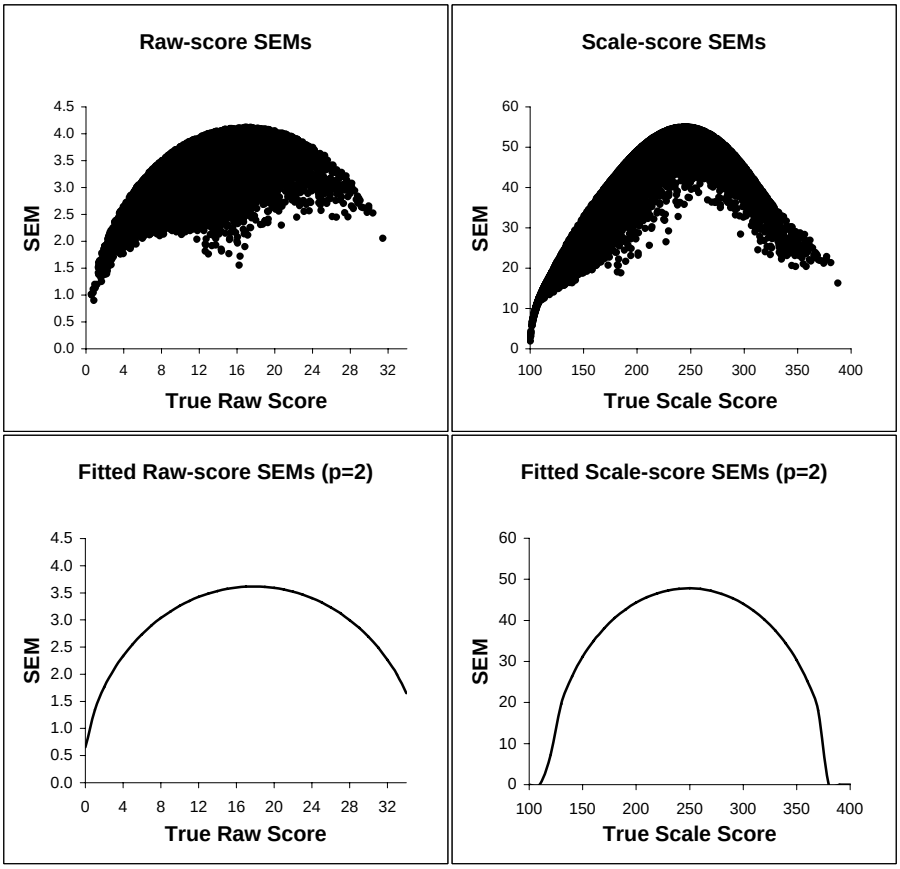

Figure 8: Conditional SEMs for Math Using the First Category

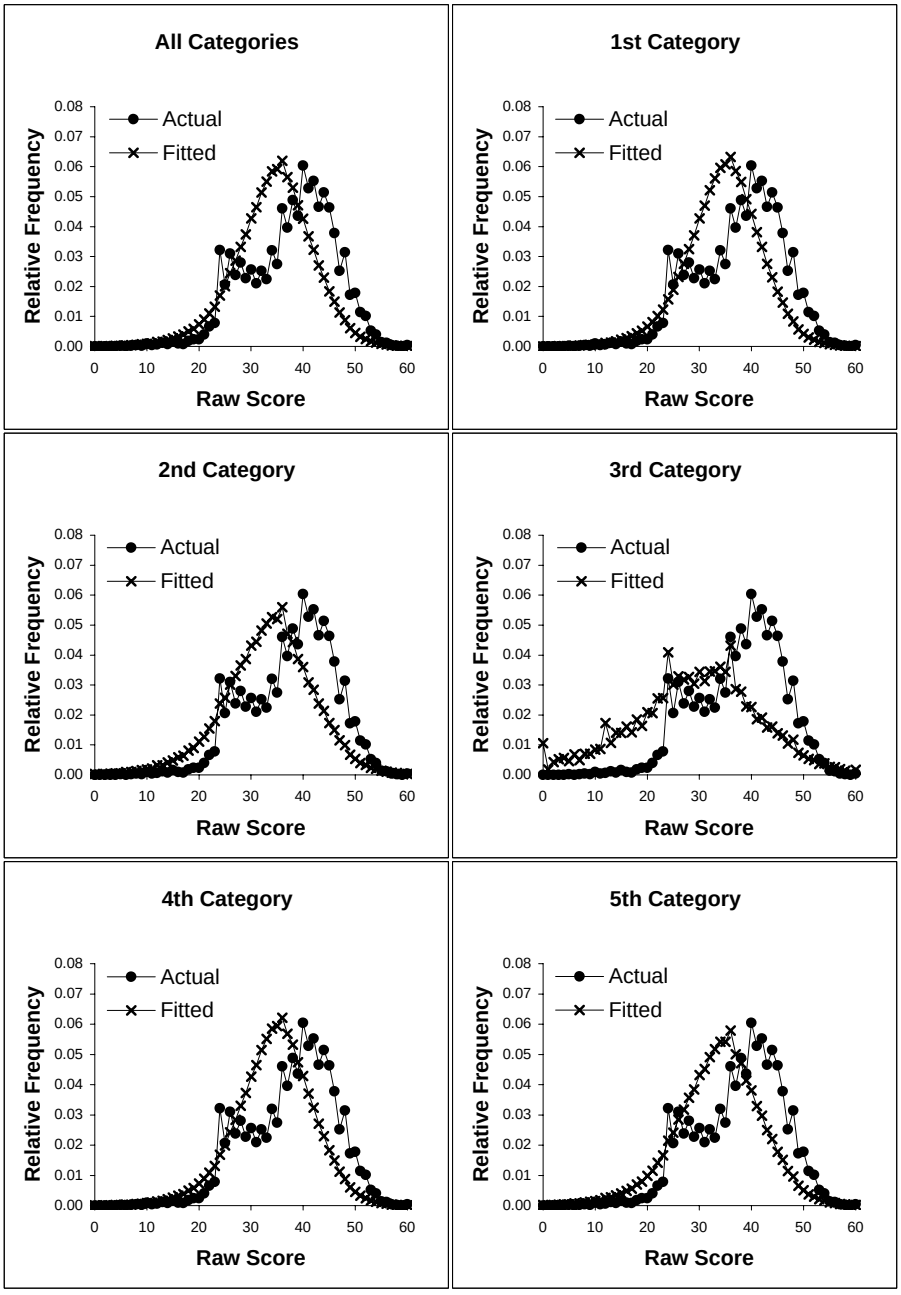

Figure 9: Actual and Fitted Raw-score Distributions for Writing

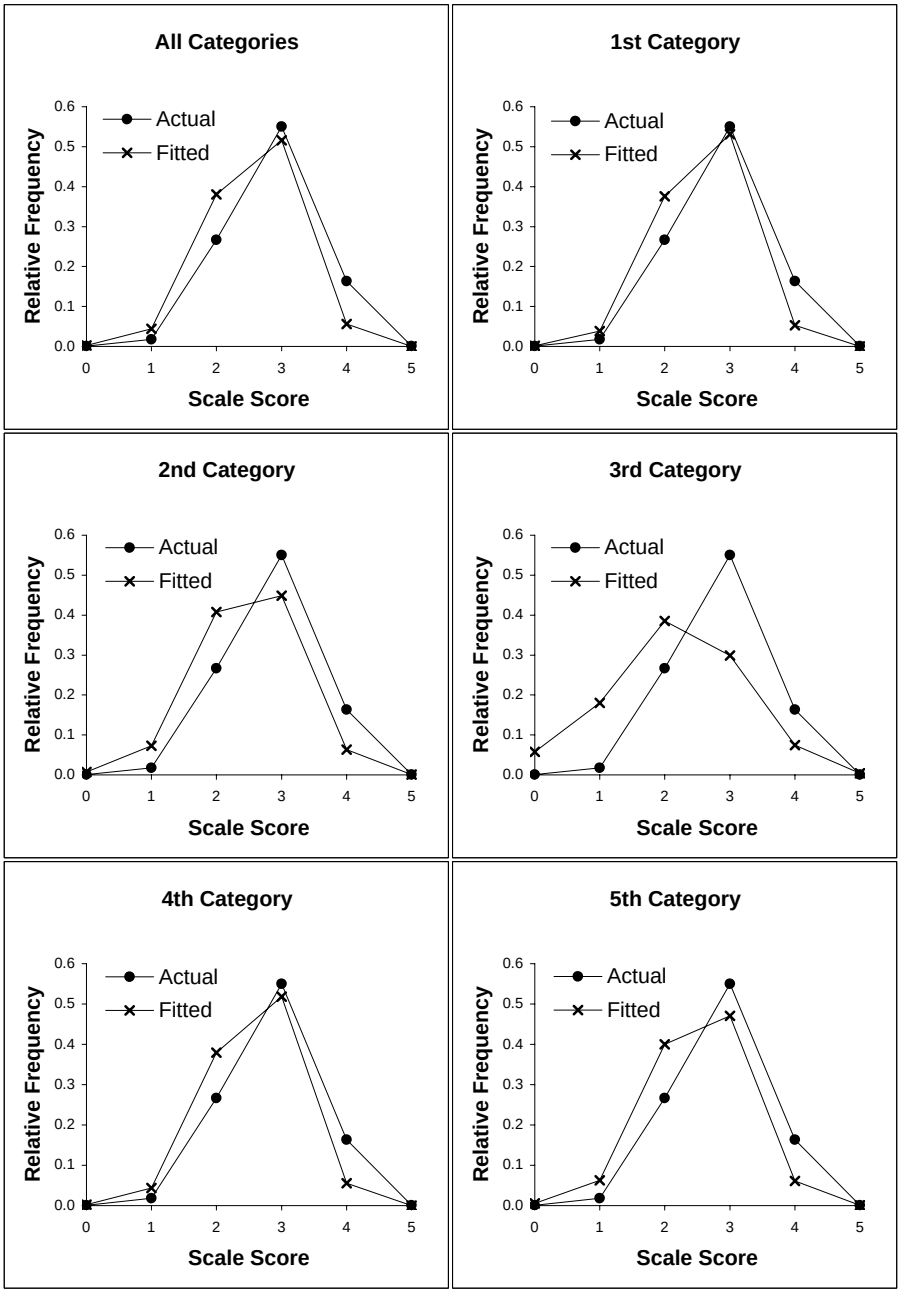

Figure 10: Actual and Fitted Scale-score Distributions for Writing

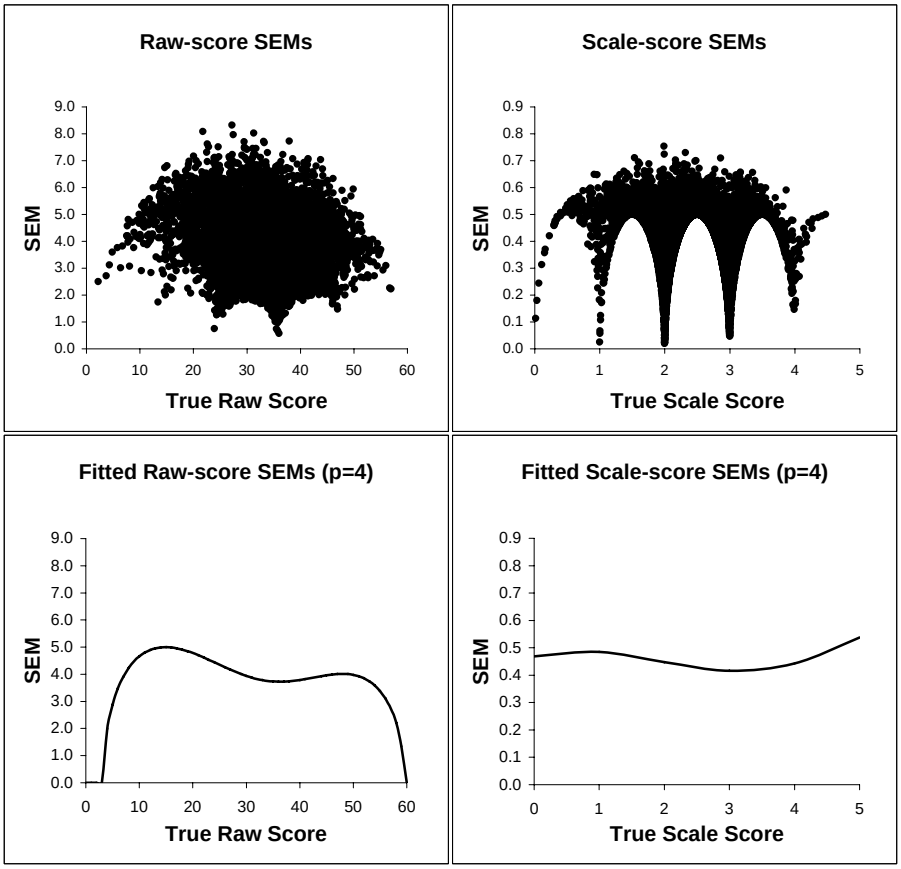

Figure 11: Conditional SEMs for Writing Using the Fifth Category

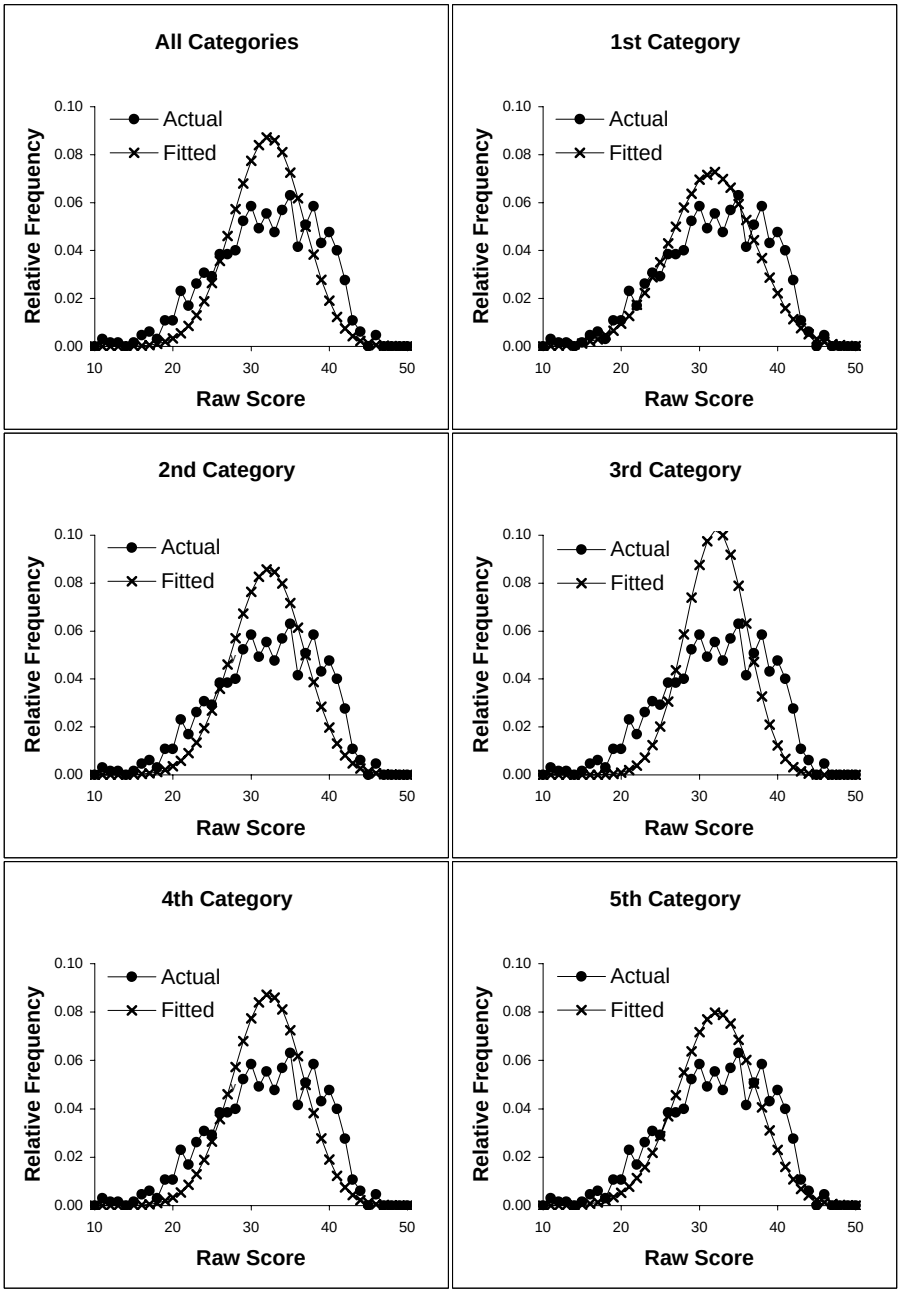

Figure 12: Actual and Fitted Raw-score Distributions for Essay

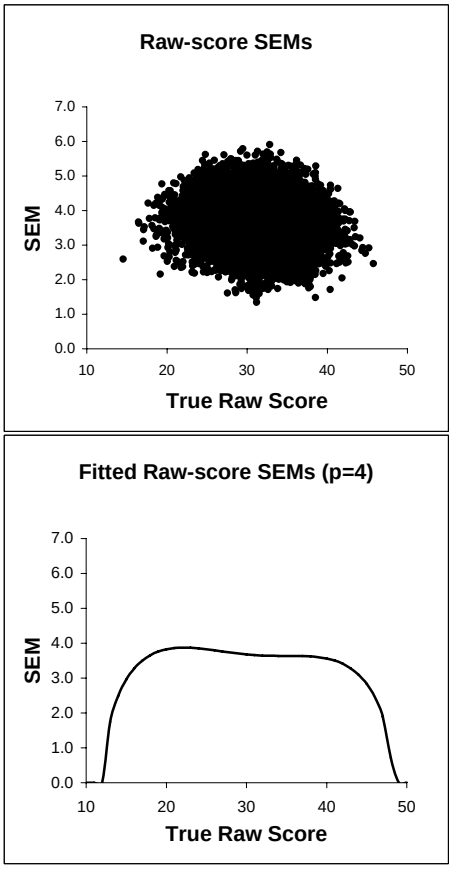

Figure 13: Conditional SEMs for Essay Using the First Category