Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 10

A Multinomial Error Model for Tests with Polytomous Items*

Won-Chan Lee†

January 2005

*A previous version of this paper was presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, April 2001. The author is grateful to Bradley Hanson, Robert Brennan, Michael Kolen, and Mary Pommerich for their helpful comments on the paper.

†Send correspondence to Won-Chan Lee, 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).

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

Contents

List of Tables

List of Figures

Abstract

This paper introduces a multinomial error model, which models the raw scores from polytomously scored items. The multinomial error model is implemented in this paper for estimating conditional standard errors of measurement and reliability for both raw and scale scores. A simulation study is presented, which suggests that the multinomial conditional standard errors of measurement for the raw and scale scores are stable estimates. A compound multinomial error model is also presented when the items are stratified according to content categories and/or prespecified numbers of score categories. The applicability of the multinomial and compound multinomial models is illustrated by two real data examples. The first example considers test scores obtained from polytomous items only, and the second example contains test scores from a mixture of dichotomous and polytomous items.

1 Introduction

Several decades ago, Lord (1955, 1957) presented a formula, under the binomial error model, for estimating the conditional standard error of measurement (SEM) for raw (i.e., number correct) scores obtained from dichotomous items. In the terminology of generalizability theory, Lord's SEM is the absolute (Δ–type) error, which concentrates on the difference between examinee observed and universe (or true) scores (Brennan, 2001). It has been proved that the average of Lord's error variances over all examinees in the sample group is equal to the error variance in the KR21 (Kuder & Richardson, 1937) reliability coefficient (Brennan & Kane, 1977). Keats (1957) proposed a correction for Lord's error variance to make the average error variance equal to the error variance incorporated in the KR20 (Kuder & Richardson, 1937) reliability coefficient. The error variance in KR20 is the relative (δ–type) error in generalizability theory, which quantifies the measurement error involved in differences between examinee observed and universe scores relative to the differences between population means for observed and universe scores (Brennan, 2001). When a test consists of polytomous items, Cronbach's (1951) α is often used to compute reliability of raw scores. The error variance in α is the relative error. If α is applied to dichotomous items, the result is identical to KR20.

In many testing programs, raw scores typically are transformed to scale scores for the purposes of score reporting and making decisions about examinees. This paper mainly considers scale scores that are non-linearly transformed from raw scores. It has been widely recognized that conditional SEMs and reliability should be expressed in the metric of scale scores—e.g., Standard 2.2 in the current *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Several procedures have been developed for estimating conditional SEMs (and reliability) for scale scores (Kolen, Hanson, & Brennan, 1992; Kolen, Zeng, & Hanson, 1996; Brennan & Lee, 1997, 1999; Feldt & Qualls, 1998). In particular, Brennan and Lee (1997, 1999) provide formulas, under the binomial error model, to compute conditional SEMs for scale scores, which can be viewed as the scale-score analogue of Lord's SEM. Readers can refer to Lee, Brennan, and Kolen (2000) for reviews of several procedures for computing conditional scale-score SEMs. The procedures cited above have a limitation in that they were developed under the assumption of dichotomously scored items, and do not consider explicitly the case for polytomous items. Wang, Kolen, and Harris (2000) present a polytomous IRT model approach for estimating conditional SEMs and reliability for scale scores with polytomous items.

The primary purposes of the present paper are (1) to introduce a multinomial error model that can be used to estimate conditional SEMs and reliability for both raw and scale scores for tests with polytomous items; (2) to demonstrate the performance of the model in recovering the true conditional SEMs in a simulation study; and (3) to illustrate the application of the multinomial error model to real data. In addition, a compound multinomial error model is

introduced when items are assumed to be sampled from stratified domains (e.g., content domains). It is shown that the compound multinomial error model can be applied effectively to the case of mixed item types, i.e., a test consisting of both dichotomous and polytomous items. When the multinomial error model is applied to raw scores, the resultant conditional SEM can be viewed as a generalization of Lord's SEM to polytomous items. When the multinomial error model is applied to scale scores, the resultant conditional SEM is an extension of Brennan and Lee's (1997, 1999) binomial conditional SEM for scale scores to the polytomous situation.

2 The Multinomial Error Model

Suppose a test contains n polytomous items, and each item is scored as one of k possible score points, $c_1 < c_2 < \cdots < c_k$. It is assumed that a sample of n items has been drawn at random from an undifferentiated universe of such items. Let $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_k\}$ denote the proportions of items in the universe for which an examinee can get scores of $c_1, c_2, \ldots, c_k$, respectively, such that $\pi_1 + \pi_2 + \cdots + \pi_k = 1$. Further, let $X_1, X_2, \ldots, X_k$ be the random variables representing the numbers of items scored $c_1, c_2, \ldots, c_k$, respectively, such that $X_1 + X_2 + \cdots + X_k = n$. The total score Y is represented by the sum of the item scores: $Y = c_1X_1 + c_2X_2 + \cdots + c_kX_k$. Let Ω be the space of the set of points $(x_1, x_2, \ldots, x_k)$. The dependent random variables $X_1, X_2, \ldots, X_k$ follow a multinomial distribution, which is given by

$$f(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1!\, x_2! \cdots x_k!}\, \pi_1^{x_1}\pi_2^{x_2}\cdots\pi_k^{x_k}, \qquad (x_1, x_2, \ldots, x_k) \in \Omega. \quad (1)$$

It should be noted that Equation 1 and all the subsequent equations (except for reliability formulas) are for a single individual having $\boldsymbol{\pi}$.

2.1 Raw Scores

2.1.1 PDF of Raw Scores

The problem here is that of finding the probability density function (PDF) of $Y = c_1X_1 + c_2X_2 + \cdots + c_kX_k$. The PDF of Y is crucial to compute the PDF of scale scores and conditional SEMs presented later. Let Ψ be the space of Y. Since $\Pr(Y = y) = \Pr(c_1X_1 + c_2X_2 + \cdots + c_kX_k = y)$,

$$g(y) = \Pr(Y = y \,|\, \boldsymbol{\pi}) = \sum_{c_1x_1 + c_2x_2 + \cdots + c_kx_k = y} f(x_1, x_2, \ldots, x_k), \qquad y \in \Psi, \quad (2)$$

where the summation is over all the sets of points of Ω such that $c_1x_1 + c_2x_2 + \cdots + c_kx_k = y$. Given the fact that $\sum_\Omega f(x_1, x_2, \ldots, x_k) = 1$, it is also true that $\sum_\Psi g(y) = 1$, because all elements in Ω are completely transformed to elements in Ψ.

The first step to find $g(y)$ is to determine the space of Y, Ψ. Note that $c_1 n \leq y \leq c_k n$, and the increment of y depends on the c values. In many cases, $c_1, c_2, \ldots, c_k$ are equally spaced with a constant a. Then, it follows that $\Psi = \{c_1 n, c_1 n + a, c_1 n + 2a, \ldots, c_k n\}$. If $c_1, c_2, \ldots, c_k$ are unequally spaced, one can go through all possible sets of $x_1, x_2, \ldots, x_k$ to find all the unique values of Y, which can be done via a simple computer program.

The conditional SEMs for the raw scores can be derived in two different ways. The most direct method is to use the distribution of Y, $g(y)$. A much simpler approach, which does not require the use of $g(y)$, involves the usual formula to compute the variance of a linear combination. As presented in the next section, however, the computation of conditional SEMs for scale scores requires $g(y)$, which, in turn, determines the PDF of scale scores, $h(s)$. Note that all the formulas for the raw scores are expressed in the total score metric.

2.1.2 Conditional Raw-Score SEM

Let e denote the error of measurement. Note that $\sigma^2_{e(Y)} = \sigma^2_Y$ for a particular individual with fixed $\boldsymbol{\pi}$, i.e., variance of observed scores for a person equals error variance for that person. The equality also holds for each X variable. The standard deviation of Y is the conditional raw-score SEM for the examinee with $\boldsymbol{\pi}$, which is given by

$$\sigma_{e(Y)} = \sqrt{\sum_{\Psi} y^2 \, g(y) - \left[\sum_{\Psi} y \, g(y)\right]^2}. \tag{3}$$

An unbiased estimator of $\sigma_{e(Y)}$ for an examinee with $\hat{\pi}_i = x_i^*/n$ $(i = 1, 2, \ldots, k)$, where x_i^* is the observed number of items scored c_i, is

$$\hat{\sigma}_{e(Y)} = \sqrt{\frac{n}{n-1}} \sqrt{\sum_{\Psi} y^2 \, \hat{g}(y) - \left[\sum_{\Psi} y \, \hat{g}(y)\right]^2}, \tag{4}$$

where $\hat{g}(y)$ is computed by replacing π_i with $\hat{\pi}_i$ in Equations 1 and 2.

Another derivation to compute $\sigma_{e(Y)}$ is presented next, which does not require the use of $g(y)$. Since Y is a linear combination of Xs, the conditional error variance for Y can be expressed as

$$\sigma^2_{e(Y)} = \sum_{i=1}^{k} c_i^2 \sigma^2_{e(X_i)} + 2 \sum \sum_{i<j} c_i c_j \sigma_{e(X_i)e(X_j)}, \tag{5}$$

where $\sigma_{e(X_i)e(X_j)}$ represents the covariance of errors. It can be shown that $\sigma_{e(X_i)e(X_j)} = -n\pi_i\pi_j$ (Olkin, Gleser, & Derman, 1980). The marginal distribution of X_i $(i = 1, 2, \ldots, k)$ is binomial, $b(n, \pi_i)$ (Olkin et al., 1980). It follows that $\sigma^2_{e(X_i)} = n\pi_i(1 - \pi_i)$. Thus, the conditional SEM given $\boldsymbol{\pi}$ is given by

$$\sigma_{e(Y)} = \sqrt{n \sum_{i=1}^{k} c_i^2 \pi_i (1 - \pi_i) - 2n \sum \sum_{i<j} c_i c_j \pi_i \pi_j}. \tag{6}$$

Substituting $\hat{\pi}_i = x_i^*/n$ for π_i and multiplying by $\sqrt{n/(n-1)}$, an unbiased estimator of $\sigma_{e(Y)}$ can be obtained as

$$\hat{\sigma}_{e(Y)} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{k} c_i^2 x_i^*(n-x_i^*) - 2\sum\sum_{i<j} c_i c_j x_i^* x_j^*\right]}. \tag{7}$$

Equation 7 is computationally less burdensome than Equation 4, and the two equations will produce identical results except that the value of $\hat{\sigma}_{e(Y)}$ in Equation 4 will be zero whenever any one (or more) of $\hat{\pi}_i$ is zero, which is not the case for Equation 7—this statement is also true for the parameters, i.e., Equations 3 and 6. However, with a minor modification, for instance, using 10^{-10} when $\hat{\pi}_i = 0$ in Equation 4, the results for the two equations will be virtually identical for most practical purposes.

Note that, if expressed in the mean score metric, Equation 7 will produce results identical to Equation 5.32 in Brennan (2001), which is a general formula for computing conditional Δ–type (i.e., absolute) SEMs for polytomous items under generalizability theory. Note also that, as a special case, if items are scored dichotomously, Equation 7 reduces to

$$\hat{\sigma}_{e(Y)} = \sqrt{\frac{y(n-y)}{n-1}}, \tag{8}$$

where y is the number of items correct for the examinee. Equation 8 is Lord's (1955, 1957) SEM estimated under the binomial error model.

2.1.3 Raw-Score Reliability

Reliability is often defined as

$$\rho_{YY'} = 1 - \frac{\bar{\sigma}^2_{e(Y)}}{\sigma^2_Y}, \tag{9}$$

where σ^2_Y is the variance of the raw scores for the population, and $\bar{\sigma}^2_{e(Y)}$ is the overall error variance, or the average conditional error variance over all persons in the population. Equation 9 can be estimated as

$$\hat{\rho}_{YY'} = 1 - \frac{\bar{\hat{\sigma}}^2_{e(Y)}}{\hat{\sigma}^2_Y} = 1 - \frac{(1/m)\sum \hat{\sigma}^2_{e(Y)}}{\sum (Y-\overline{Y})^2/m}, \tag{10}$$

where $\overline{Y}$ is the average raw score and the summations are taken over m examinees in the sample.

In the terminology of generalizability theory (see, for example, Brennan, 2001), the average error variance in the numerator in Equation 10 is absolute error variance, rather than the relative error variance in α. If items are scored dichotomously, Equation 10 is a very close approximation to KR21. Equation 10 should be interpreted with caution. Although it is a well-known, variance-ratio

type of reliability, Equation 10 can be viewed as a non-traditional estimator in that the error variance in the numerator is not the same as that in the denominator. The error variance in the numerator is absolute error variance, whereas the error variance in the denominator is relative error variance. This makes Equation 10 an underestimate of α–type reliability. This is true for all subsequent reliability estimators presented in this paper.

2.2 Scale Scores

2.2.1 PDF of Scale Scores

Suppose there exists a function, $u(Y)$, which transforms the raw scores Y to scale scores S. It does not matter whether $u(Y)$ is linear or non-linear. We shall consider the transformation function $u(Y)$ that can be either one-to-one (i.e., every single point of Y is converted to a unique point of S) or many-to-one (i.e., several points of Y are converted to a single point of S). Let Θ be the space of S. The PDF of S is determined as

$$h(s) = \Pr(S = s \,|\, \boldsymbol{\pi}) = \sum_{y:u(y)=s} g(y), \qquad s \in \Theta, \tag{11}$$

where $y : u(y) = s$ is interpreted as the summation taken over all y values such that $u(y) = s$. If $u(Y)$ is a one-to-one function, $h[u(y)] = g(y)$.

2.2.2 Conditional Scale-Score SEM

Given $h(s)$ for a particular examinee, the conditional scale-score SEM for the examinee is

$$\sigma_{e(S)} = \sqrt{\sum_{\Theta} s^2\, h(s) - \left[\sum_{\Theta} s\, h(s)\right]^2}. \tag{12}$$

An unbiased estimator of the conditional scale-score SEM can be obtained as

$$\hat{\sigma}_{e(S)} = \sqrt{\frac{n}{n-1}} \sqrt{\sum_{\Theta} s^2\, \hat{h}(s) - \left[\sum_{\Theta} s\, \hat{h}(s)\right]^2}, \tag{13}$$

where $\hat{h}(s)$ is computed by using $\hat{\pi}_i = x_i^*/n$ $(i = 1, 2, \ldots, k)$ in Equations 1, 2, and 11. Note that Equation 13 is the polytomous analogue of the Brennan and Lee's (1997, 1999) binomial conditional scale-score SEM for dichotomous items.

2.2.3 Scale-Score Reliability

As presented in Kolen et al. (1992) and Kolen et al. (1996), reliability for scale scores under the assumption of the uncorrelated true and error scale scores can be defined as:

$$\rho_{SS'} = 1 - \frac{\bar{\sigma}^2_{e(S)}}{\sigma^2_S}, \tag{14}$$

where σ_S^2 is the variance of scale scores for the population, and $\bar{\sigma}_{e(S)}^2$ is the overall scale-score error variance, or the average conditional scale-score error variance over all the examinees in the population. Equation 14 can be estimated as

$$\hat{\rho}_{SS'} = 1 - \frac{\bar{\hat{\sigma}}_{e(S)}^2}{\hat{\sigma}_S^2} = 1 - \frac{(1/m)\sum \hat{\sigma}_{e(S)}^2}{\sum (S - \overline{S})^2/m}, \tag{15}$$

where $\overline{S}$ is the average scale score and the summations are taken over m examinees in the sample.

3 The Compound Multinomial Model

The situation considered in this section is one in which a test consists of a set of fixed content categories. Assuming that errors within each content category follow the multinomial distribution and errors are uncorrelated across categories, the total scores over categories are distributed as what might be referred to as a compound multinomial distribution. If the procedure developed in this section is applied to dichotomous items, the raw-score result will equal to Feldt's (1984) SEM, and the scale-score result will be identical to that derived from Brennan and Lee's (1999) compound binomial procedure. Another common situation in which the compound multinomial distribution can be used is when a test is composed of items that differ in terms of the number of score categories, k—for example, a mixture of dichotomous items ($k = 2$) and polytomous items ($k > 2$). It might even be the case that different items with prespecified numbers of score categories are associated with different content categories.

Suppose a test consists of L sets of items with different item sets being associated with either different number of score categories or different content categories, or both. Each item set contains n_l ($l = 1, 2, \ldots, L$) items that are scored as one of k_l possible score points. For any set l, let X_{lj} ($j = 1, 2, \ldots, k_l$) be the random variables for the numbers of items scored c_{lj} ($j = 1, 2, \ldots, k_l$) such that $\sum_j X_{lj} = n_l$. The raw score for each set will be obtained as $Y_l = c_{l1}X_{l1} + c_{l2}X_{l2} + \cdots + c_{lk_l}X_{lk_l}$. Let Ξ be the space of the total raw scores across all the sets of items defined as: $T = \sum_{l=1}^{L} w_l Y_l$, where w_l is the weight for the set $l = 1, 2, \ldots, L$. The question here is how to compute the conditional SEM for an examinee with $\boldsymbol{\pi}_l$ ($l = 1, 2, \ldots, L$) and reliability for the total raw and scale scores. It is beyond the scope of this paper to discuss how to determine the weights, and readers can refer to Wainer and Thissen (1993) for issues related to weighting different types of items to create composite scores.

3.1 Formulas for Total Raw Scores

Let us consider the PDF of the total raw scores denoted as $g_t(t)$. In order to get $g_t(t)$, the PDF of the raw scores for each item set, $g_l(y_l)$, is computed first using Equation 2. Under the assumption of uncorrelated errors over L sets of

items,

$$\Pr(Y_1 = y_1, \ldots, Y_L = y_L \,|\, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_L) = \prod_{l=1}^{L} g_l(y_l). \tag{16}$$

It follows that the PDF of the total raw scores is given by

$$g_t(t) = \Pr(T = t \,|\, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_L) = \sum_{y_1,\ldots,y_L : \sum w_l y_l = t} \Pr(Y_1 = y_1, \ldots, Y_L = y_L \,|\, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_L), \qquad t \in \Xi, \tag{17}$$

where $y_1, \ldots, y_L : \sum w_l y_l = t$ indicates that the summation is taken over all possible sets of $y_1, \ldots, y_L$ raw score values such that the weighted sum of the scores is equal to a total raw score t.

Similar to Equation 3, the conditional SEM for an examinee with $\boldsymbol{\pi}_l$ ($l = 1, 2, \ldots, L$) is

$$\sigma_{e(T)} = \sqrt{\sum_{\Xi} t^2 \, g_t(t) - \left[\sum_{\Xi} t \, g_t(t)\right]^2}. \tag{18}$$

An estimator of $\sigma_{e(T)}$ can be expressed as:

$$\hat{\sigma}_{e(T)} = b\sqrt{\sum_{\Xi} t^2 \, \hat{g}_t(t) - \left[\sum_{\Xi} t \, \hat{g}_t(t)\right]^2}, \tag{19}$$

where b is the bias-correction factor, and $\hat{g}_t(t)$ is computed using $\hat{\boldsymbol{\pi}}_l = \{x^*_{l1}/n_l, x^*_{l2}/n_l, \ldots, x^*_{lk_l}/n_l\}$ for $l = 1, 2, \ldots, L$. The formula for computing b is presented later in Equation 22.

The conditional SEMs for the total raw scores can also be computed without using the PDF of the total raw scores, $g_t(t)$. The conditional raw-score SEM for each item set, $\sigma_{e(Y_l)}$, can be estimated using Equation 7. Assuming errors are uncorrelated across different sets of items, the estimated conditional SEM for the total raw scores is given by

$$\hat{\sigma}_{e(T)} = \sqrt{\sum_{l=1}^{L} w_l^2 \hat{\sigma}^2_{e(Y_l)}}. \tag{20}$$

Note that if all items are scored dichotomously Equation 20 is equal to Feldt's (1984) raw-score SEM.

The result in Equation 20 uses the weighted sum of the *unbiased* estimates of the conditional variances for the L item sets. Let

$$\hat{\sigma}^*_{e(T)} = \sqrt{\sum_{l=1}^{L} \frac{n_l - 1}{n_l} w_l^2 \hat{\sigma}^2_{e(Y_l)}}, \tag{21}$$

where the quantity $(n_l - 1)/n_l$ makes the term in the square root a weighted sum of the *biased* estimates of the conditional variances for the item sets. Then, the following bias-correction factor can be used in Equation 19:

$$b = \hat{\sigma}_{e(T)}/\hat{\sigma}^*_{e(T)} = \sqrt{\frac{\sum_{l=1}^{L} w_l^2 \hat{\sigma}^2_{e(Y_l)}}{\sum_{l=1}^{L}[(n_l - 1)/n_l] w_l^2 \hat{\sigma}^2_{e(Y_l)}}}. \tag{22}$$

Note that b is not a constant and changes as a function of $\hat{\boldsymbol{\pi}}_l$ values. Analogous to Equation 10, a reliability for the total raw scores can be estimated using the average total raw score and average conditional error variance over the examinees in the sample.

3.2 Formulas for Scale Scores

We shall consider two different situations here. The first situation involves total raw scores over all items in a test that are converted to scale scores through a transformation function, $S_t = v(T)$. In this case, Equation 11 can be used to obtain the PDF of S_t, $h_t(s_t)$, by substituting $g_t(t)$ computed in Equation 17 for $g(y)$. Let Υ be the space of S_t. The conditional SEM for an examinee with $\boldsymbol{\pi}_l$ $(l = 1, 2, \ldots, L)$ is

$$\sigma_{e(S_t)} = \sqrt{\sum_{\Upsilon} s_t^2 \, h_t(s_t) - \left[\sum_{\Upsilon} s_t \, h_t(s_t)\right]^2}. \tag{23}$$

An estimator of $\sigma_{e(S_t)}$ is

$$\hat{\sigma}_{e(S_t)} = b\sqrt{\sum_{\Upsilon} s_t^2 \, \hat{h}_t(s_t) - \left[\sum_{\Upsilon} s_t \, \hat{h}_t(s_t)\right]^2}, \tag{24}$$

where b is defined in Equation 22, and $\hat{h}_t(s_t)$ is obtained using $\hat{\boldsymbol{\pi}}_l = \{x^*_{l1}/n_l,$ $x^*_{l2}/n_l, \ldots, x^*_{lk_l}/n_l\}$ for $l = 1, 2, \ldots, L$.

Another common situation is when total raw scores for each item set are converted to scale scores, and then composite scale scores are defined as a weighted sum of the subset scale scores: $S_c = \sum_{l=1}^{L} a_l S_l$, where a_l is the weight. Given the PDFs of raw scores, $g_l(y_l)$, the PDF of scale scores for each item set, $h_l(s_l)$, can be computed using Equation 11. Equation 13 will then provide $\hat{\sigma}_{e(S_l)}$ for each item set. Assuming errors are uncorrelated across the item sets,

$$\hat{\sigma}_{e(S_c)} = \sqrt{\sum_{l=1}^{L} a_l^2 \hat{\sigma}^2_{e(S_l)}}. \tag{25}$$

Once the conditional scale-score SEMs are obtained, for either situation, the same logic previously described (i.e., Equation 15) can be used to estimate the scale-score reliability.

4 A Simulation Study

A simulation study was conducted to investigate the performance of the multinomial error model in recovering the true conditional SEMs for raw and scale scores. The somewhat arbitrarily chosen characteristics of the test are as follows: number of items (n) is 10, number of score points (k) is 5, and item scores $\boldsymbol{c} = \{c_1, c_2, c_3, c_4, c_5\} = \{0, 1, 2, 3, 4\}$. A hypothetical conversion table was created, in which the raw scores ranging from 0 to 40 are converted to percentile rank-type scale scores ranging from 1 to 99. The top portion of Figure 1 shows the plot for the hypothetical conversion.

4.1 Simulation Procedure

To generate true proportions of items scored as $0, 1, \ldots, 4$, π_{pi} ($p = 1, 2, \ldots, m$ and $i = 1, 2, \ldots, 5$), it is assumed that the random vectors $\boldsymbol{\pi}_p = \{\pi_{p1}, \ldots, \pi_{pk}\}$ are independent and follow a Dirichlet distribution. (Note that the subscript p is introduced here to designate persons, which was omitted in the previous development.) A Dirichlet distribution with parameters $\beta_1, \beta_1, \ldots, \beta_k$ is defined as (Johnson & Kotz, 1972):

$$t(\pi_{p1}, \pi_{p2}, \ldots, \pi_{pk}) = \frac{\Gamma(\beta_1 + \cdots + \beta_k)}{\Gamma(\beta_1) \cdots \Gamma(\beta_k)} \, \pi_{p1}^{\beta_1 - 1} \pi_{p2}^{\beta_2 - 1} \cdots \pi_{pk}^{\beta_k - 1}. \tag{26}$$

Let γ denote the intraclass correlation between item responses and $\boldsymbol{E}$ denote the expectation operator. It can be shown that $\boldsymbol{E}(\pi_{pi}) = \beta_k \,/\, \sum_{i=1}^{k} \beta_i$ and $\gamma = 1/[(\sum_{i=1}^{k} \beta_i) + 1]$ (Lui, Cumberland, Mayer, & Eckhardt, 1999). Given the values of $\boldsymbol{E}(\pi_{pi})$ and γ, the beta parameters are uniquely determined. For the present simulation study, $\gamma = .4$ and $\boldsymbol{E}(\pi_{pi}) = \{.2, .2, .2, .2, .2\}$ were adopted. Even though the uniform-type pattern of $\boldsymbol{E}(\pi_{pi})$ may not be realistic, it is guaranteed to produce sufficient simulees across the entire true score range. Dirichlet random vectors were generated for 2000 simulees using Jöhnk's method as described in Narayanan (1990, p. 23). For simulee p with $\boldsymbol{\pi}_p$, the raw and scale score distributions, $g(y)$ and $h(s)$, respectively, were computed using Equations 2 and 11. Then, the true raw (τ_p) and scale scores (ξ_p) for simulee p were computed as:

$$\tau_p = n \sum_{i=1}^{k} c_i \pi_{pi}, \tag{27}$$

and

$$\xi_p = \sum_{\Theta} s \, h(s). \tag{28}$$

True conditional raw and scale-score SEMs for the 2000 simulees were computed using Equations 3 and 12. The simulation steps were as follows:

1. A set of random item responses for $n = 10$ items and $m = 2000$ simulees were generated by comparing uniform random deviates with the cumulative π_{pi} values. For example, if a random number fell within an interval

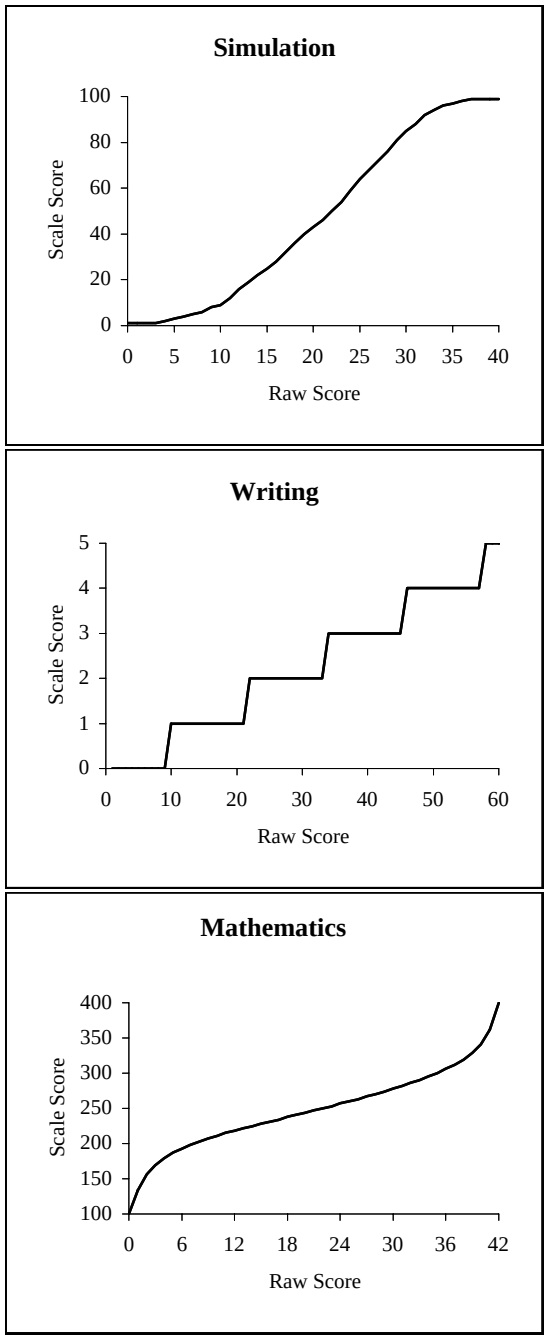

Figure 1: Raw-to-Scale Score Conversions

$(0, \pi_{pi})$, a score of c_1 was assigned, if within $(\pi_{p1}, \pi_{p1}+\pi_{p2})$, a score of c_2 was assigned, and so on.

2. The estimated conditional raw and scale-score SEMs were computed for the 2000 simulees based on the simulated data using the multinomial model via Equations 4 and 13.

3. The above steps were replicated $r = 100$ times.

To evaluate the accuracy of the estimated SEMs, root mean-squared error ($RMSE$) for the conditional SEMs was computed for each simulee as:

$$RMSE_p = \sqrt{\frac{1}{r}\sum_{j=1}^{r}(\hat{e}_{pj} - e_p)^2}, \tag{29}$$

where e_p is the true SEM for simulee p, $\hat{e}_{pj}$ is the estimated SEM for simulee p obtained in replication j. The average $RMSE$ over all simulees was computed as $\sqrt{(1/m)\sum_p RMSE_p^2}$.

4.2 Results

Depicted in Figure 2 are the true and mean estimated conditional SEMs over 100 replications for 2000 simulees. The plots on the left column present the results for the raw scores, and the plots on the right column present the results for the scale scores. The top plots display the true conditional SEMs for the 2000 simulees for the raw and scale scores. The middle plots represent the square root of the mean of the 100 estimated error variances for each of the 2000 simulees. The bottom plots show the fitted true and mean estimated SEMs. The polynomial degree of 4 (P=4) was selected because the R-square value showed a big jump between degrees 3 and 4, and little difference was observed for degrees greater than 4. The same criterion was employed to select polynomial degrees for all subsequent analyses. As discussed later, the use of a fitted SEM with a high degree polynomial, or an arithmetic mean of the conditional error variance considered in the next section on real data examples, is based on the practical necessity that only one value of the conditional SEM be reported at each score point.

In contrast to the well-known fact that the raw-score SEMs conditioning on the true scores for dichotomous items have a smooth inverted-U shape, the true and mean estimated SEMs for polytomous items estimated under the multinomial error model are vertically scattered showing more spread near the mid true score values. The SEMs conditional on a particular true raw score can vary depending upon the configuration of the item scores. For the raw scores near the middle of the score range, there exist many possible combinations of item scores that lead to the same raw score. By contrast, there are relatively few possible combinations of item scores leading to the same raw score that is either very low or high.

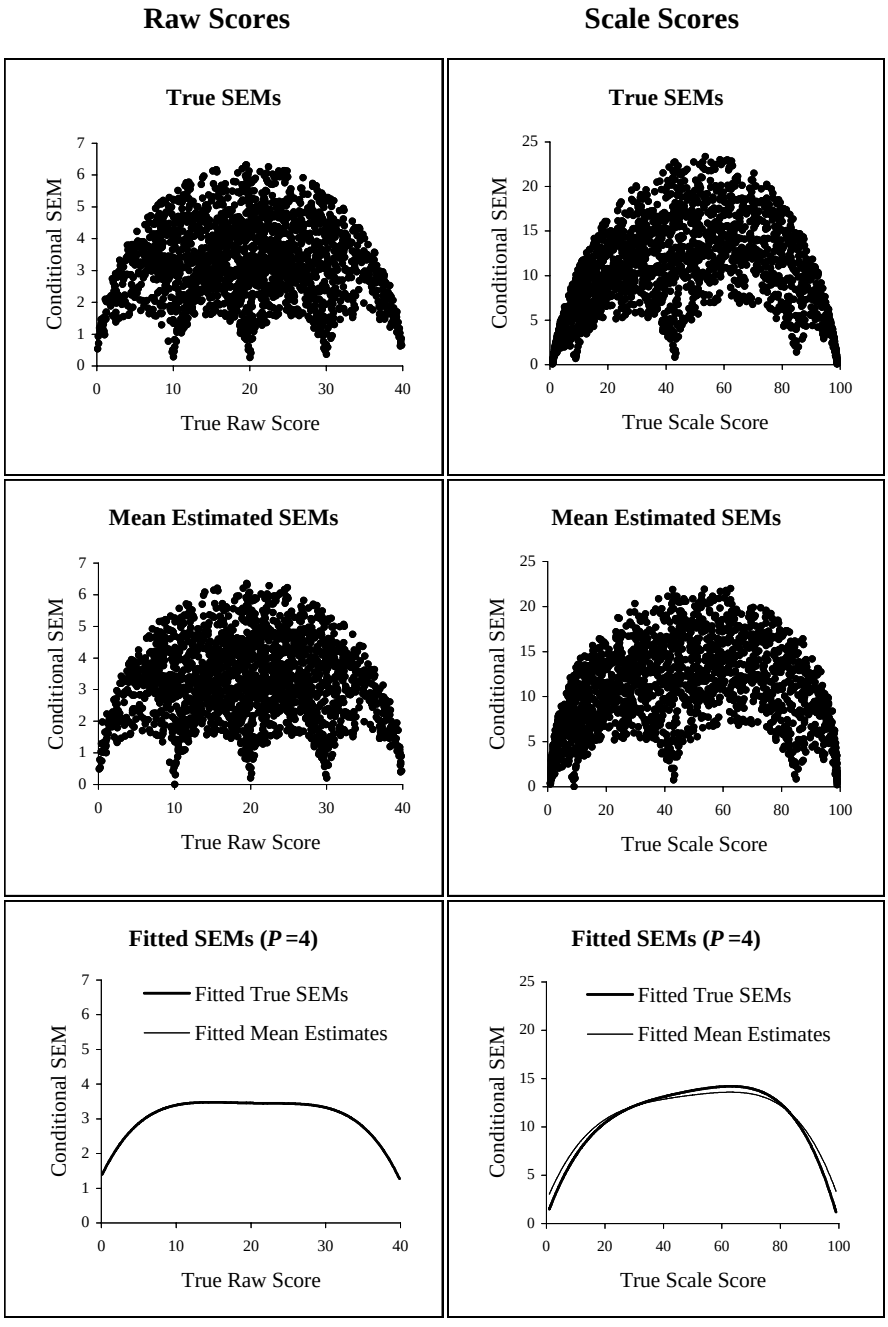

Figure 2: Conditional SEMs for Simulation

The true and mean estimated conditional SEMs for raw scores, in general, appear to be very close, and the fitted SEMs are almost identical. Also notice that the conditional raw-score SEMs show an "umbrella" pattern where the SEMs approach zero at every tenth score point. This is strongly related to the fact that if the scores for all items are the same, the SEMs are necessarily zero. Since 10 items are scored on 0–4 ratings for this simulation study, the only possible raw scores obtainable from identical item scores include 0, 10, 20, 30, and 40. This issue is discussed further later.

The shape of the conditional scale-score SEMs also shows an umbrella pattern, but the scale-score points corresponding to the five raw-score values where the SEMs approach zero are not equally spaced because of a non-linear characteristic of the raw-to-scale score transformation. The transformation function displayed in Figure 1 clearly shows that the scale scores are supposed to be shrunken at both ends and stretched out in the middle. Given the fact that the fitted conditional raw-score SEMs have the shape of a concave-down parabola, one can predict the shape of the conditional scale-score SEMs based on the pattern of the transformation function. It has been found in some previous research (e.g., Brennan & Lee, 1997, 1999; Lee et al., 2000) that the slope of the transformation at each score point is positively correlated with the magnitude of the scale-score SEM at the particular score point. In our example, the transformation function tends to be flat near both ends, which results in small scale-score SEMs. Whereas the fitted true and mean estimated raw-score SEMs do not reveal any noticeable difference, the fitted scale-score SEMs tend to be somewhat overestimated near both ends and slightly underestimated in the middle of the score scale. This finding is consistent with the results reported in Lee et al. (2000). That is, the estimated conditional scale-score SEMs tend to be underestimated at score points where the slope of the transformation is steep, but overestimated at score points where the slope is rather flat. It seems that the systematic bias in the estimated conditional scale-score SEMs is not because of the model itself but because of the non-linear property of the transformation function.

Figure 3 displays the plots of *RMSE* values for the 2000 simulees. For the raw scores, *RMSE* values are, in general, smaller near the middle range of the true score scale than at both tails. The pattern of the *RMSE* values for the scale scores is similar to that for the raw scores in most parts of the score scale, except that the *RMSE* values become very small for extremely low and high true scale scores. The small RMSE values at both extremes of the scale scores appear to be inconsistent with the plots in Figure 2 where relatively large bias is found in the fitted scale-score SEMs at both extremes. The same observation can be made in the true scale score range of 50 through 70 where the bias is relatively large but *RMSE* values are relatively small. The overall *RMSE* consists of two types of errors: bias due to the estimation method and random error over replications. Thus, Figures 2 and 3 together suggest that the region of the scale scores where the bias is large is likely to be associated with small random error. The average *RMSE* values over the 2000 simulees are 1.008 and 4.012 for the raw and scale scores, respectively.

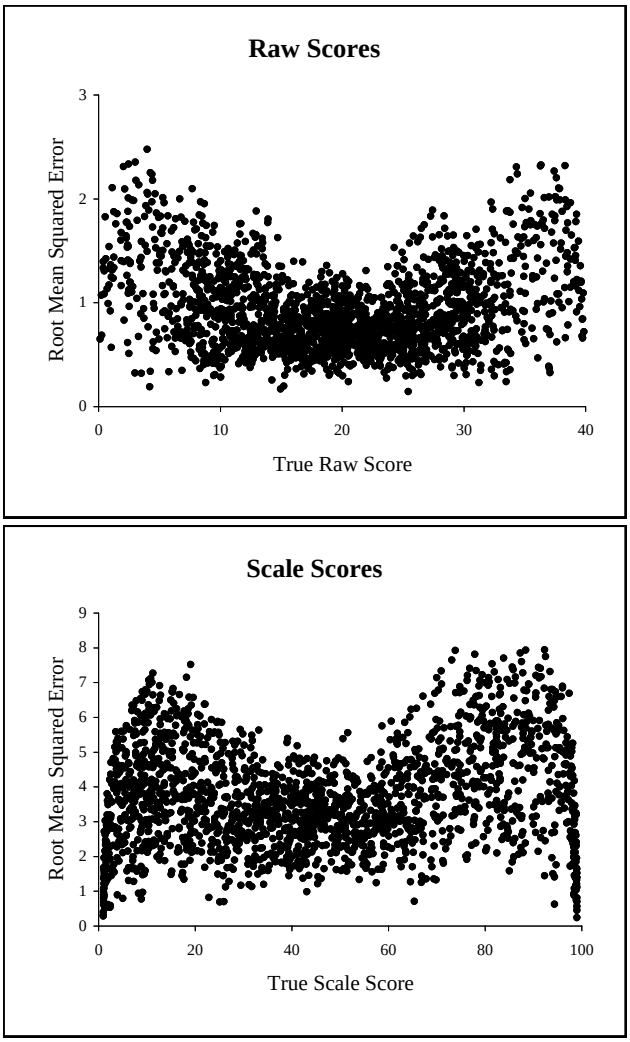

Figure 3: Root Mean Squared Errors of Conditional SEMs for Simulation

Overall, provided the model fits the data, the estimates of the conditional raw-score SEMs based on the multinomial error model appear to be very stable in recovering the true SEMs. Although the bias shown in the estimated conditional scale-score SEMs might be of some concern, the largest bias tends to occur near the true scale scores that are extremely low or high, and the actual magnitude of the bias may not be of great consequence in practice. The maximum difference between the true and the mean estimated scale-score SEMs was 4 in the scale score metric (0–100), which occurred at the true scale score of 98.96; and the maximum difference in terms of the fitted SEMs was 4 at the true scale score of 97.43.

5 Real Data Examples

In this section, two real data examples are presented to illustrate the application of the multinomial and compound multinomial error models for computing conditional SEMs and reliability for the raw and scale scores. The first example includes test data composed of polytomous items only, while the second data set contains mixed item types.

5.1 Example 1: The Multinomial Model

Data were obtained from the Work Keys Writing Assessment (ACT, Inc., 1998). A random sample of 5000 examinees' test scores was selected from a pool containing test scores for examinees who took the tests in 1997–1998. The Writing Assessment consists of 6 items. Each item is rated on a 0–5 scale by two independent raters so that each examinee receives 12 scores. The raw scores ranging from 0 to 60 are converted to integer level (i.e., scale) scores ranging from 0 to 5. Although the actual number of items is six, the analysis was conducted based on 12 ratings as if there were 12 items. The second plot in Figure 1 shows the actual conversion function for the test. The most unique feature of this conversion function is that there are only a few scale score points and several raw scores are converted to a single scale score.

Table 1 summarizes the results of the analysis. To simplify the presentation of results, the raw scores are arbitrarily combined into five intervals with a width of 10 raw-score points. The average SEM is the square root of the average error variances. The number of examinees falling in each interval is also provided. The average SEMs for both raw and scale scores do not seem to demonstrate any obvious pattern even if they do not vary a lot along the score scales. The overall average SEMs for the raw- and scale-score SEMs are 2.152 and .305, respectively.

Also presented in Table 1 are the test score variance, average error variance, and reliability estimated under the multinomial error model for the raw and scale scores. Note that the reliability estimate for the scale scores is lower than the reliability estimate for the raw scores suggesting that the use of fewer score points can reduce the reliability of a test (see Kolen et al., 1992). Recall that

Table 1: Analysis Summary for Writing

Raw Scores			Scale Scores		
Raw Interval	N	Avg SEM	Scale Score	N	Avg SEM
0–10	10	2.595	0	5	.414
11–20	65	2.601	1	90	.383
21–30	1012	1.791	2	1335	.263
31–40	1832	2.208	3	2751	.310
41–50	1909	2.263	4	816	.344
51–60	172	1.997	5	3	.282
$\hat{\sigma}^2_Y =$	65.616		$\hat{\sigma}^2_S =$	.494	
Avg $\hat{\sigma}^2_{e(Y)} =$	4.630		Avg $\hat{\sigma}^2_{e(S)} =$	.093	
$\hat{\rho}_{YY'} =$	.929		$\hat{\rho}_{SS'} =$	.811	

the reliability estimates reported in Table 1 use absolute error variances, and thus are expected to be lower than other reliability estimates using relative error variances such as α. The value of α for the raw scores using the same data is .942. The scale-score reliability can not be estimated using α.

Figure 4 contains the plots of the conditional SEMs for the 5000 examinees. The top two plots depict the conditional error variances for the raw and scale scores. The middle two plots are the square root of the average error variances at each score point, which provide basically the same information as Table 1 except that Table 1 is based on only a few intervals. The bottom two plots represent the square root of the fitted conditional error variances. The conditional error variances for the raw scores show a rough umbrella shape with several equally spaced score points where the error variances approach zero. As discussed earlier in the simulation study section, all identical ratings will result in raw scores of 0, 12, 24, 36, 48, and 60 at which the conditional error variances are necessarily zero. The conditional error variances for the scale scores do not show the umbrella pattern because of the many-to-one character of the conversion.

The fitted and average conditional SEMs, in general, appear to be similar in patterns and sizes. For practical purposes, the choice of either method would be a matter of preference, although the fitted SEMs have the desirable property of smoothness. For the scale scores in this particular example, the fitted and average SEMs do not differ greatly because there are only six scale-score points. With six score points, the fitted SEMs will be exactly the same as the average SEMs if a polynomial degree of 5 is used.

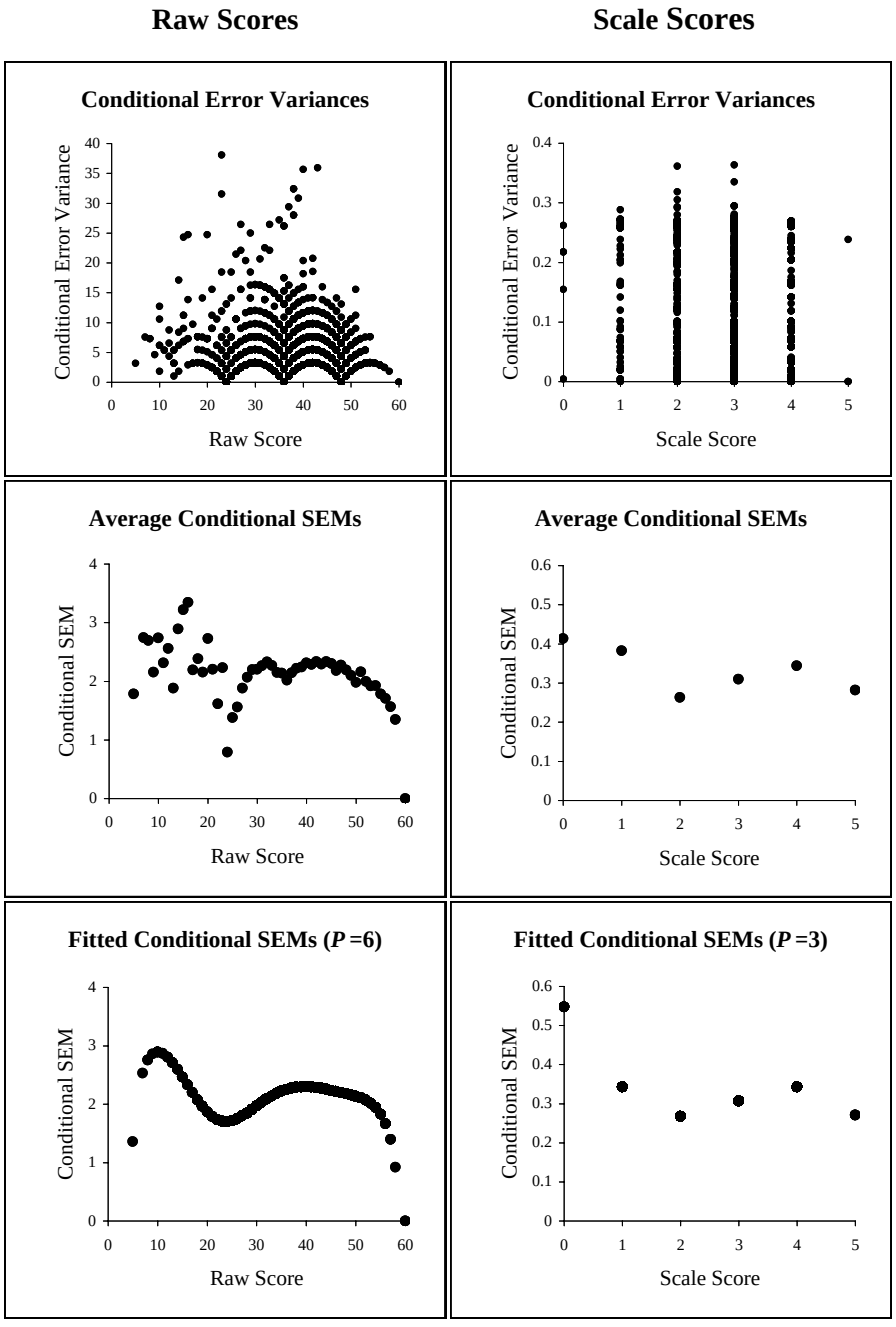

Figure 4: Conditional SEMs for Writing

5.2 Example 2: The Compound Multinomial Model

The data set used to illustrate the compound multinomial model for mixed item types contains 4044 examinees' test scores on a math achievement test administered by a state government to 10th graders in Fall, 2000. The math test consists of a mixture of eight polytomous items that are scored using four integer values, 0–3, and 18 multiple choice items. The total raw scores are defined as the sum of the item scores (i.e., $w_1 = w_2 = 1$). The total raw scores ranging from 0 to 42 are transformed to scale scores ranging from 100 to 400. The bottom plot of Figure 1 exhibits the conversion function for the math test. Notice that the shape of the conversion function for the math test is an inverted-S shape, which is the opposite shape of the transformation used in the simulation study. The difference in the shape of the conversion function affects the pattern of the scale-score SEMs as discussed later.

A summary of the results is presented in Table 2. Both the raw and scale score values are arbitrarily grouped into four intervals, and the number of examinees and average SEM associated with each interval are provided. Unlike the results for the writing test, the average SEMs based on the math data tend to show some particular patterns. The average raw-score SEMs tend to be larger near the middle range of the score scale. By contrast, the average scale-score SEMs are larger near the ranges of very low and high scores. These patterns can be more clearly observed in the plots of the individual conditional SEMs presented later. The reliability estimate of the scale scores is slightly lower than that of the total raw scores. The overall average conditional SEMs are 3.231 and 15.371 for the total raw and scale scores, respectively.

Table 2: Analysis Summary for Math

Raw Scores			Scale Scores		
Raw Interval	N	Avg SEM	SS Interval	N	Avg SEM
0–10	534	2.497	100–200	345	20.691
11–20	863	3.645	201–250	1290	12.552
21–30	1360	3.615	251–300	1904	13.942
31–42	1287	2.739	301–400	505	21.549
$\hat{\sigma}^2_T =$	103.867		$\hat{\sigma}^2_{S_t} =$	1759.235	
Avg $\hat{\sigma}^2_{e(T)} =$	10.440		Avg $\hat{\sigma}^2_{e(S_t)} =$	236.283	
$\hat{\rho}_{TT'} =$	.899		$\hat{\rho}_{S_tS'_t} =$	.866	
Dichotomous Items ($n_1 = 18$)			Polytomous Items ($n_2 = 8$)		
$\hat{\sigma}^2_{Y_1} =$	26.658		$\hat{\sigma}^2_{Y_2} =$	31.239	
Avg $\hat{\sigma}^2_{e(Y_1)} =$	3.182		Avg $\hat{\sigma}^2_{e(Y_2)} =$	7.257	
$\hat{\rho}_{Y_1Y'_1} =$	.881		$\hat{\rho}_{Y_2Y'_2} =$	.768	

The bottom panel of Table 2 displays the raw-score results of two separate analyses for the two different item types. Note that the reliability estimate obtained from 18 dichotomous items is only .02 lower than the total raw-score reliability estimate. This conforms to the frequent criticism that the reliability of the composite scores formed by combining the multiple choice items and constructed response items does not necessarily exceed the reliability of only one part of the test (Wainer & Thissen, 1993). Note also that the sum of the average error variances for the two parts is equal to the average error variance for the total raw scores.

The conditional SEMs for all individual examinees are plotted in Figure 5. The layout of Figure 5 is the same as Figure 4. It is particularly important to note that the patterns of the conditional error variances for the raw and scale scores are reversed, which is due to the non-linear pattern of the raw-to-scale score conversion (see Figure 1), in which the slope of the conversion is steep at both tails and flat in the middle of the score scale. The degree of slope has enormous impact on the magnitude of the resulting scale-score SEMs (Brennan & Lee, 1999). The results for the math test can be compared to those for the simulation study. The conditional scale-score SEMs for the two cases, in general, tend to show reversed patterns because of the reversed patterns of the conversion functions.

The fitted and average conditional SEMs seem to coincide very well. In this case, either approach would serve well in practice. There are some examinees with zero and perfect scores in this data set and their conditional SEMs for both raw and scale scores are necessarily zero. (Notice the average SEMs are zero at the raw scores of 0 and 42 and scale scores of 100 and 400.) In particular, the polynomial fit to the data for the conditional scale-score error variances is greatly affected by the zero SEM values at both ends of the scale although the number of those examinees is small. Thus, the zero and perfect scorers were deleted from the data set in fitting the polynomial because the general trend of the fit is seriously distorted by them.

6 Summary and Discussion

Estimates of reliability and conditional SEMs are often used to evaluate the psychometric properties of test scores. Since most test scores that are reported to examinees are scale scores transformed from raw scores, more appropriate and informative estimates of reliability and conditional SEMs would be expressed in the metric of the scale scores. Moreover, due to the current popularity of the performance assessments, for which items typically are scored polytomously, there has been a high demand for procedures for estimating reliability and conditional SEMs for test scores based on polytomous items. In this paper, a multinomial error model is introduced that can be used to estimate conditional SEMs and reliability for (a) any test scores from either dichotomous or polytomous items; and (b) any types of scale scores including raw scores. The multinomial error model can be viewed as a generalization of the two previous developments: the

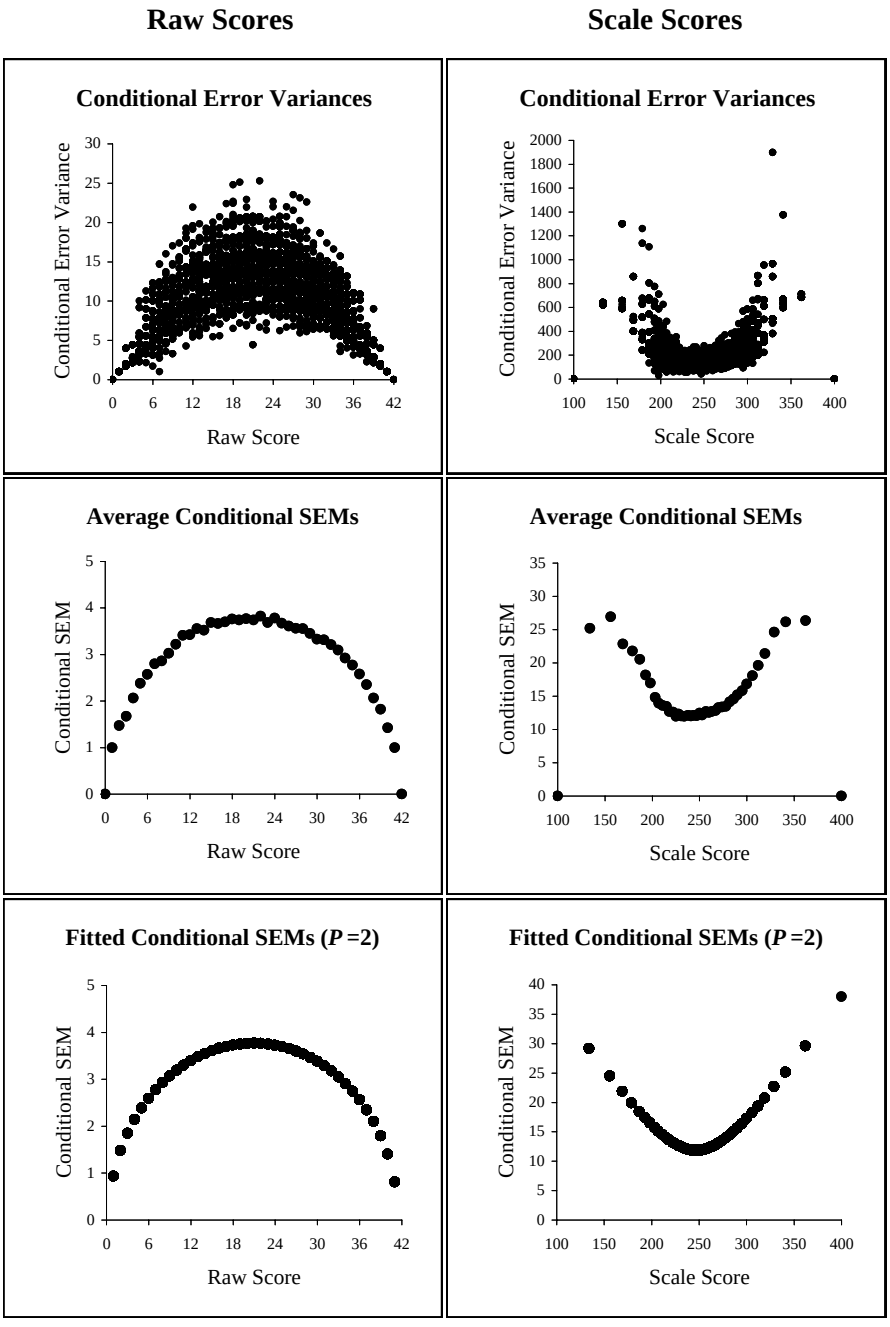

Figure 5: Conditional SEMs for Math

binomial error model by Lord (1955, 1957) for computing conditional raw-score SEMs for dichotomous items, and Brennan and Lee (1997, 1999) who extended the binomial error model and provided procedures for estimating conditional scale-score SEMs.

Conceptualizing the multinomial error model involves the assumption of randomly parallel forms. That is, polytomous items with a prespecified number of score categories in a test form are conceived of as a random sample from an undifferentiated domain of such items. Such random samples with the same number of items are considered to be randomly parallel forms. It is important to note, however, that the multinomial error model presented in this paper does not attempt to model a situation in which the items included on a test form are sampled at random from a domain of items when items in the domain can have a different number of score categories. In this case, the total number of score points for one examinee would differ from that for another examinee, and a more complicated model would be needed.

A compound multinomial model was also introduced, which models test scores from a test consisting of a set of fixed strata such as content specifications or score categories, i.e., a mixture of items with different prespecified multinomial dimensions. A real data example showed that the compound multinomial error model can be applied to the data with a mixture of dichotomous and polytomous items. The only additional assumption involved in the procedure for the mixed item situation was that the errors are uncorrelated across different item types.

The simulation study revealed that the multinomial conditional raw-score SEMs were stable estimates of the true SEMs, provided that the model fits the data. The estimated conditional scale-score SEMs showed slight systematic bias primarily due to the non-linear characteristic of the raw-to-scale score conversion. That is, the bias tended to be negatively correlated with the degree of slope in the conversion along the score scale. The magnitude of the bias, however, may not have a great consequence for practical purposes. A more comprehensive simulation study might be necessary.

The effect of the raw-to-scale score transformations on the pattern of the estimated conditional scale-score SEMs was similar to the findings in other studies using dichotomous items (Feldt & Qualls, 1998; Brennan & Lee, 1999; Lee et al., 2000). The severity of the transformation function along the score scale tended to determine the magnitude of the conditional scale-score SEMs, in general. The degree of slope in the transformation is positively related to the size of the conditional scale-score SEMs along the score scale.

Unlike dichotomous items, the conditional raw- and scale-score SEMs for polytomous items tend to be scattered due to their dependency on the patterns of item scores. For example, Equations 4 or 7 provide an unambiguous estimate of the conditional SEM for a particular examinee. However, different combinations of $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k$ would provide different estimated conditional SEMs even though the raw score is the same. If it is the case that only one value of conditional SEM should be reported at each raw score point, there seem to be at least two different ways to do so. Obviously, one can go through all possible

combinations of $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_k$ that give a particular raw score value, and then take the square root of the average conditional error variance over all the combinations. This approach does not require any examinee data, and thus the test developer can create the report table for the conditional SEMs without any data collection process. As the number of items increases, however, the computation can be very extensive. Also, incorporating all possible combinations of $\hat{\pi}_i$ values may not be supported by data—some combinations may never occur in reality. Another approach would be to use the actual examinee data. Simply, one can compute the conditional error variance for each individual in the sample, and then take the square root of the average conditional error variance over all examinees with the same raw score. Due to the sample dependency of this approach, it would be necessary to use a large representative sample of the population. One problem of this approach might be that even with a reasonably large sample size, data may not include all possible raw score points. Rather than using the arithmetic means of the error variances at each raw score point, it is also possible to use a fitted polynomial regression model. The latter method might be preferable in practice because it can be used even when there are missing data at some raw-score points, and it provides a smooth function of the conditional SEMs. One disadvantage of the latter method is the subjectivity in choosing the degree of the polynomial. The illustrations provided in this paper seemed to support use of either fitted SEMs or the average SEMs over the examinees in the sample. The same approaches can be applied to the cases for scale scores and compound multinomial data.

Another practical complexity for the multinomial error model is that the estimated SEM (for both raw and scale scores) for an examinee with the same item scores for all items (i.e., one of $\hat{\pi}_i$ is one and the rest of $\hat{\pi}_i$ are zeros) will be zero. A similar problem occurs for the binomial error model with only two possible score points (i.e., 0 or 1), in which a zero estimated SEM is assigned to examinees with zero or perfect scores. Recognizing that reporting zero estimated SEMs is not preferable nor practical, Lee et al. (2000) employed an adjustment for the binomial error model. Although a similar adjustment could have been made, no particular adjustment was made for the application of the multinomial error model in the presence paper because any type of adjustment involves some degree of subjectiveness and depends on the situations the users might encounter in practice.

Lee, Brennan, and Kolen (2002, in press) discuss and compare several procedures under the binomial error model to form confidence intervals for true scores expressed in the metric of any transformed scale scores using test scores from dichotomous items. Extending the Lee et al. study, a possible future study might involve developing interval estimation procedures implementing the multinomial or compound multinomial error model for complex assessments with polytomous or mixed item types.

7 References

ACT, Inc. (1998). *Characteristics of the Work Keys assessments*. Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

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

Brennan, R. L., & Kane, M. T. (1977). Signal/noise ratios for domain-referenced tets. *Psychometrika*, *42*, 609-625.

Brennan, R. L., & Lee, W. (1997). *Conditional standard errors of measurement for scale scores using binomial and compound binomial assumptions*. (Iowa Testing Programs Occasional Paper No. 41). Iowa City, IA: University of Iowa.

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.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Feldt, L. S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement*, *44*, 883-891.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: American Council on Education and Macmillan.

Feldt, L. S., & Qualls, A. L. (1998). Approximating scale score standard error of measurement from the raw score standard error. *Applied Measurement in Education*, *11*, 159-177.

Johnson, N. L., & Kotz (1972). *Distributions in statistics: Continuous multivariate distributions*. New York: John Wiley & Sons.

Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika*, *22*, 29-41.

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.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*, 129-140.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151-160.

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., Brennan, R. L., & Kolen, M. J. (2002). *Interval estimation for true scores under various scale transformations.* (ACT Research Report 2002-5). Iowa City, IA: ACT, Inc.

Lee, W., Brennan, R. L., & Kolen, M. J. (in press). Interval estimation for true raw and scale scores under the binomial error model. *Journal of Educational and Behavioral Statistics.*

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, *15*, 325-336.

Lord, F. M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement*, *17*, 510-521.

Lui, K.-J., Cumberland, W. G., Mayer, J. A., & Eckhardt, L. (1999). Interval estimation for the intraclass correlation in Dirichlet-multinomial data. *Psychometrika*, *64*, 355-369.

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

Olkin, I., Gleser, L. J., & Derman, C. (1980). *Probability models and applications.* New York: Macmillan.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*, 103-118.

Wang, T., Kolen, M. J., & Harris, D. J. (2000). Psychometric properties of scale scores and performance levels for performance assessments using polytomous IRT. *Journal of Educational measurement*, *37*, 141-162.