Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 13

Classification Consistency and Accuracy Under the Compound Multinomial Model*

Won-Chan Lee[†]

November 2005
Revised April 2007
Revised April 2008

[†] 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

1 Introduction 1

2 The Models 2

2.1 Multinomial Model 2

2.2 Compound Multinomial Model 3

3 Classification Indices 4

3.1 Classification Consistency Indices 4

3.2 Classification Accuracy Indices 5

3.3 A Bias-Correction Procedure 6

4 Illustrative Examples 7

4.1 Data Source 7

4.2 Computer Program 8

5 Results 9

6 Discussion 11

7 References 15

Abstract

This paper describes procedures for estimating single-administration classification consistency and accuracy using the multinomial and compound multinomial models for tests with complex item scoring. Various classification consistency and accuracy indices are discussed. The procedures are illustrated using a real data set obtained from a test consisting of both polytomous and dichotomous items.

1 Introduction

When a measurement procedure is used to make categorical decisions, it is typically recommended that estimates be provided of the consistency of the decisions over two replications of the same measurement procedure (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). *Classification consistency* refers to the level of agreement between the classifications based on two randomly parallel forms of a test (Livingston & Lewis, 1995). Due to the difficulty of obtaining such data from repeated testings, classification consistency typically is estimated based on a single administration of a test using a statistical model to estimate the observed score distribution. *Classification accuracy* refers to the degree to which the classifications based on examinees' observed scores (called observed classifications) agree with those based on examinees' true scores (called true classifications) (Livingston & Lewis, 1995, Lee, Hanson, & Brennan, 2002). Estimating classification accuracy typically involves estimating both the observed score distribution and the true scores (or a distribution).

For a test that consists of dichotomously-scored items, several procedures have been reported in the literature for estimating classification indices based on a single administration of a test. The procedures can be categorized into two types—those that make distributional-form assumptions about true scores and those that do not. The approach using assumptions about the true-score distribution is referred to here as the *distributional* approach, and the one without is referred to as the *individual* approach. Some examples of the distributional approach include Huynh (1976) and Hanson and Brennan (1990). Huynh (1976) assumed a beta distribution for true scores and a binomial distribution for errors. Hanson and Brennan (1990) extended Huynh's approach by using the four-parameter beta distribution for true scores and either a binomial or compound binomial distribution for errors. By contrast, as an individual approach, Subkoviak (1976) does not make any distributional-form assumptions for true scores. The Subkoviak's procedure estimates classification consistency one examinee at a time and then averages over examinees to obtain overall classification consistency for the whole sample group. Subkoviak (1976), however, did not consider classification accuracy. In a sense, this paper extends Subkoviak's approach to complex assessments, and thus falls in the category of the individual approach.

The methods discussed in the previous paragraph assume that a test consists of dichotomously-scored items. There have been some procedures reported in the literature that can deal with tests consisting of polytomously-scored items or mixtures of dichotomous and polytomous items. The procedure suggested by Livingston and Lewis (1995) involves computing "effective test length" to create a substitute test consisting of dichotomous items. The four-parameter beta-binomial model approach (Hanson & Brennan, 1990) is then applied to the substitute test. Classification consistency and accuracy can also be estimated rather easily for any data using a somewhat untenable assumption that observed scores for two forms follow a bivariate normal distribution (Peng &

Subkoviak, 1980; Woodruff & Sawyer, 1989). If a test can be divided into two similar half tests, the procedures proposed by Woodruff and Sawyer (1989) and Breyer and Lewis (1994) can be used for computing classification consistency. Roughly speaking, the split-half procedures estimate classification consistency using the assumptions of a bivariate-normal distribution and the Spearman-Brown formula. Brennan and Wan (2004) developed a bootstrap procedure for estimating classification consistency for complex assessments such as scores by multiple raters, equated scale scores, weighted sum scores of different types of items, etc.

Given some of the ad hoc aspects of the approaches discussed above, the goal of the present paper is to provide procedures that are based on psychometric models that properly represent the characteristics of various types of items and test scores such as (un)weighted sums of polytomous item scores and (un)weighted sums of polytomous and dichotomous item scores. The procedures presented in this paper employ a multinomial model (Lee, 2007) for a test with a single item set, and a compound multinomial model (Lee, 2007) for a test consisting of mixtures of different item sets. Items are categorized into different item sets based on either their number of score points or the number of fixed content categories. For example, a test may consist of three different item sets: (a) a set of dichotomous items with 2 score points (0 and 1); (b) a set of polytomous items with 3 score points (1, 2, and 4); and (c) a set of polytomous items with 4 score points (0, 2, 4, and 6). A more complicated example might have the set of dichotomous items in the previous example nested within different fixed content categories.

The multinomial and compound multinomial procedures are implemented in this paper to compute well-known classification consistency and accuracy indices. The procedures are illustrated using a real data set obtained from a test consisting of both polytomous and dichotomous items.

2 The Models

The multinomial model is presented first, followed by the compound multinomial model. Note that the presentation of the models is analogous to Lee (2007).

2.1 Multinomial Model

Suppose a test consists of n polytomous items, and each item is scored as one of k possible score points, $c_1, c_2, \ldots, c_k$. Assume that a sample of n items is 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. Note 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 summed score Y is defined as the sum of the item scores: $Y = c_1X_1 + c_2X_2 + \cdots + c_kX_k$. The random variables $X_1, X_2, \ldots, X_k$ follow a multinomial distribution:

$$\Pr(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k \,|\, \boldsymbol{\pi}) = \frac{n!}{x_1!\, x_2! \cdots x_k!}\, \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_k^{x_k}. \tag{1}$$

Note that Equation 1 and all the subsequent equations are for *a single individual* having $\boldsymbol{\pi}$, unless otherwise noted.

Since there are several sets of values of $X_1, X_2, \ldots, X_k$ that lead to a particular value of y, the probability density function (PDF) of Y is obtained as:

$$\Pr(Y = y \,|\, \boldsymbol{\pi}) = \sum_{c_1x_1 + c_2x_2 + \cdots + c_kx_k = y} \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k \,|\, \boldsymbol{\pi}), \tag{2}$$

where the summation is over all the values of $X_1, X_2, \ldots, X_k$ such that $c_1x_1 + c_2x_2 + \cdots + c_kx_k = y$.

2.2 Compound Multinomial Model

Suppose a test contains a set of fixed content categories. Assuming that errors within each content category follow the multinomial distribution and errors are uncorrelated across content categories, the total scores over content categories are distributed as a compound multinomial distribution.[1] The compound multinomial distribution can also be used when a test consists of items that differ in terms of the number of possible score points, k—for example, mixtures of dichotomous ($k = 2$) and polytomous items ($k > 2$).

Suppose a test consists of L sets of items. 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 summed score for item set l is $Y_l = c_{l1}X_{l1} + c_{l2}X_{l2} + \cdots + c_{lk_l}X_{lk_l}$. Total summed scores across all item sets are defined as: $T = \sum_{l=1}^{L} w_l Y_l$, where w_l is the weight for the set $l = 1, 2, \ldots, L$. Under the assumption of uncorrelated errors over L sets of items,

$$\Pr(Y_1 = y_1, \ldots, Y_L = y_L \,|\, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_L) = \prod_{l=1}^{L} \Pr(Y_l = y_l \,|\, \boldsymbol{\pi}_l), \tag{3}$$

where $\Pr(Y_l = y_l \,|\, \boldsymbol{\pi}_l)$ is computed using Equation 2. Then, it follows that the PDF of the total summed scores is:

$$\Pr(T = t \,|\, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_L) = \sum_{y_1, \ldots, y_L : \sum w_l y_l = t} \Pr(Y_1 = y_1, \ldots, Y_L = y_L \,|\, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_L), \tag{4}$$

where $y_1, \ldots, y_L : \sum w_l y_l = t$ indicates that the summation is taken over all possible sets of $y_1, \ldots, y_L$ summed-score values such that the weighted sum of the scores is equal to a total summed score t. Note that when there is only one item set, the compound multinomial model reduces to the multinomial model.

[1] Note that the phrase "compound multinomial model" is sometimes used to characterize statistical models different from that discussed here.

3 Classification Indices

Unless otherwise noted, the presentation of formulas in this section focuses on total score T obtained from mixtures of dichotomous and polytomous items (i.e., $L = 2$) and the compound multinomial model is thus assumed. Recall that the formulas presented in the previous section are for a single examinee. In this section, the subscript p is introduced to indicate that reference is to a single examinee as opposed to an average over all examinees in the sample.

3.1 Classification Consistency Indices

Suppose examinees are classified into H mutually exclusive categories based on a set of $H - 1$ cut scores, $\lambda_1, \lambda_2, \ldots, \lambda_{H-1}$. Let $\lambda_0 = min(T)$ and $\lambda_H = max(T)$. Let $f_{ph} \equiv \Pr(\lambda_{h-1} \leq T < \lambda_h \,|\, \boldsymbol{\pi}_{p1}, \boldsymbol{\pi}_{p2})$ denote the category probability for observed total scores, which is the probability that examinee p is classified into category h, except that for f_{pH} the range includes λ_H. $\Pr(T = t)$ is computed using Equation 4. A consistent classification is made for examinee p if the examinee is classified into the same category on any two random replications of the measurement procedure. Thus, the probability of a consistent classification for examinee p is

$$\phi_p = \sum_{h=1}^{H} f_{ph}^2, \tag{5}$$

which we call an index of *conditional* classification consistency. We call this index “conditional” in the sense that it is conditional on the person, which is analogous to the use of the phrase “conditional” standard error of measurement.

Strictly speaking, ϕ_p and $1 - \phi_p$ depend on $\boldsymbol{\pi}_{pl}$ $(l = 1, 2)$, which are never known. Often, $\overline{\boldsymbol{x}}_{pl}$ are used as estimates of $\boldsymbol{\pi}_{pl}$ to obtain $\hat{\phi}_p$ and $1 - \hat{\phi}_p$.

An overall index of classification consistency for a group of M examinees can be obtained as

$$\phi = \frac{1}{M} \sum_{p=1}^{M} \phi_p. \tag{6}$$

To obtain an estimate of ϕ, the summation in Equation 6 is taken over estimates of the conditional indices, $\hat{\phi}_p$.

Another well-known index for classification consistency is the Cohen’s (1960) kappa coefficient. The kappa coefficient quantifies classification consistency that exceeds chance levels. The formula for kappa for summed scores is given by

$$\kappa = \frac{\phi - \phi_c}{1 - \phi_c}, \tag{7}$$

where ϕ_c is the expected or chance probability of consistent classification. The chance probability is the sum of products of marginal category probabilities, which are determined, in this paper, as averages (over examinees) of the conditional category probabilities. That is,

$$\phi_c = \sum_{h=1}^{H} \left[\frac{1}{M} \sum_{p=1}^{M} f_{ph} \right]^2 . \tag{8}$$

3.2 Classification Accuracy Indices

Classification consistency is concerned with classifications based on two model-predicted observed score distributions. By contrast, classification accuracy involves a comparison of classifications based on a single estimated observed score distribution and correct classifications based on examinees' "known" true scores. For the sake of simplicity, it will be assumed here that the true cut scores used to determine examinees' true category status are the same as the observed cut scores.

If an examinee's true score is known, then the true category to which the examinee is assigned is determined, and an accurate classification is made for the examinee only if the examinee is classified in the same category based on his or her observed score (obtained under the model). From this perspective, an index of conditional classification accuracy is simply equal to the category probability that is associated with the examinee's true category, as follows:

$$\gamma_p = f_{ph}, \tag{9}$$

where examinee p is in category h on the true classification. The corresponding overall index for a group of M examinees is

$$\gamma = \frac{1}{M} \sum_{p=1}^{M} \gamma_p. \tag{10}$$

Estimates of γ_p and γ can be obtained using $\overline{\boldsymbol{x}}_{pl}$ in place of $\boldsymbol{\pi}_{pl}$. In that case, an examinee's true category status is, in fact, the category into which the examinee is assigned based on the classifications using actual data. In other words, the actual classifications based on the data at hand are treated as definitive true classifications (this issue will be discussed further later). For example, if an examinee's actual observed score (as an estimate of the examinee's true score) falls in the second category, an estimate of γ_p for the examinee is simply the sum of probabilities for all model-predicted observed-score points that belong to the second category.

For purposes of a simple comparison, let us consider a binary classification situation (i.e., pass/fail). In effect, $\hat{\phi}_p$ "counts" both twice-passing *and* twice-failing decisions as consistent classifications based on two hypothetical classifications, whereas $\hat{\gamma}_p$ "counts" only twice-passing *or* twice-failing decisions depending on the examinee's status on the original test based on the actual classification and one hypothetical classification. Note that (a) if $\hat{\gamma}_p > .5$ then $\hat{\gamma}_p > \hat{\phi}_p$; and (b) if $\hat{\gamma}_p < .5$ then $\hat{\gamma}_p < \hat{\phi}_p$. The first condition (i.e., $\hat{\gamma}_p > .5$)

will be more likely to be observed in most cases. For example, if an examinee's actual test score is above the cut score, then the model-predicted probability of the examinee's passing the test ($\hat{\gamma}_p$) is expected to be greater than .5, if the model describes the data adequately. Thus, it is anticipated that the index of classification accuracy be larger than the index of classification consistency.

Other types of classification accuracy indices that are often used in the literature include false positive and false negative error rates (Hanson & Brennan, 1990). A false positive (negative) error occurs when an examinee is classified into a category that is higher (lower) than the examinee's true category. Indices of *conditional* false positive and false negative error rates, respectively, are given by

$$\gamma_p^+ = \sum_{h=h^*+1}^{H} f_{ph}, \tag{11}$$

and

$$\gamma_p^- = \sum_{h=1}^{h^*-1} f_{ph}, \tag{12}$$

where h^* represents the examinee's true category. Indices of overall false positive and false negative error rates for the whole group, respectively, are

$$\gamma^+ = \frac{1}{M}\sum_{p=1}^{M} \gamma_p^+, \tag{13}$$

and

$$\gamma^- = \frac{1}{M}\sum_{p=1}^{M} \gamma_p^-. \tag{14}$$

3.3 A Bias-Correction Procedure

Wan, Brennan, and Lee (2007) reported that classification consistency indices estimated based on the compound multinomial model were biased. The potential bias in the compound multinomial classification consistency indices is due to use of the observed proportion scores as estimators of the true proportion scores. Even though an examinee's observed proportion scores are unbiased estimators of the examinee's true proportion scores, the classification consistency indices obtained over examinees can be biased, because the variance of the observed scores (over examinees) is typically larger than the variance of the true scores.

Brennan and Lee (2006b) propose a bias-correction procedure, which provides alternative estimates of the true proportion scores such that the variance of the resulting scores is, in theory, equal to the variance of the true scores. The principal idea of the Brennan and Lee's bias-correction procedure stems from two generally known facts that (a) the true score variance is *smaller* than the observed score variance, but (b) it is *larger* than the variance of Kelley's (1947) regressed score estimates. Brennan and Lee (2006b) suggested to use an "optimally" weighted average of the two types of estimates. The weights for the two types of estimates are "optimal" in the sense that the variance of the resulting estimates is equal to the true score variance.

For the case of the multinomial model, the observed proportion score for score category j for examinee p is denoted $\overline{x}_{pj}$, and the multinomial version of Kelley's regressed score estimate (Brennan & Lee, 2006b) is $(1-\hat{\rho}^2)\hat{\mu}_j + (\hat{\rho}^2)\overline{x}_{pj}$, where $\hat{\rho}^2$ is a reliability estimate for the test and $\hat{\mu}_j$ is the mean (over examinees) observed proportion score for score category j. Brennan and Lee (2006b) show that the optimally weighted estimate for score category j is given by

$$\tilde{\pi}_{pj} = \left(1 - \sqrt{\hat{\rho}^2}\right)\hat{\mu}_j + \left(\sqrt{\hat{\rho}^2}\right)\overline{x}_{pj}. \tag{15}$$

The $\tilde{\pi}_{pj}$ values are substituted for the π_{pj} values in the estimation process. When the compound multinomial model is under consideration, the $\tilde{\pi}_{pj}$ values are computed for each item set separately. Even though the bias-correction procedure, originally, was considered for classification consistency indices, it is employed for both consistency and accuracy indices in this paper.

4 Illustrative Examples

The procedures presented in this paper are illustrated using a real data set. The compound multinomial model is employed to compute classification indices based on the data set that contains mixtures of polytomous and dichotomous items. The multinomial model approach is applied to the two separate data sets, each of which consists of only a single item type (i.e., polytomous or dichotomous items). The results are compared to the those from the Livingston and Lewis (1995) procedure.

4.1 Data Source

Data were from a science achievement test administered by a state government to approximately 4000 10th graders. The science test consists of mixtures of 40 dichotomous items scored 0/1 and 7 polytomous items scored 0-3 (integer values). Pseudo cut scores were set to the summed-score values of 15, 30, and 45, which assign examinees into four mutually exclusive categories. Computation of the classification indices for this data set was carried out using the compound multinomial model.

Summary descriptive statistics for the total summed scores are presented in Table 1. As can be seen, the score distributions tend to be negatively skewed. Note that the reliability and SEM were computed based on the compound multinomial model approach (Lee, 2007) as discussed below.

An estimate of reliability is needed to obtain results for the Livingston and Lewis procedure. It would be advisable that a reliability coefficient that involves absolute error variance (in the terminology of generalizability theory, Brennan, 2001) be employed for the Livingston and Lewis procedure (Wan et al., 2007). This is because (a) the focus, in the context of classification consistency, is

Table 1: Summary Statistics

N	Mean	SD	Skew	Kurt	Rel*	SEM*
4178	35.686	9.569	-0.382	2.608	0.848	3.729

Note. *Reliability and SEM are computed based on the compound multinomial model.

not on the relative standing of each examinee in the sample group, but on the absolute magnitude of each examinee's score with respect to the cut scores; and (b) the model adopted in the Livingston and Lewis procedure to describe the errors in each examinee's test scores (obtained over an infinite number of randomly equivalent forms) is the binomial error model, which is known to be closely related to absolute error variance (Brennan & Lee, 2006a). In this paper, reliability coefficients were computed using the compound multinomial model (Lee, 2007), which involves absolute error variance. The conditional error variance was computed for each examinee, and then the average error variance over all examinees was obtained. One minus the ratio of the average error variance to the observed score variance was taken as an estimate of reliability. Note that this estimate of reliability will be very close to the estimate of the Φ coefficient obtained from multivariate generalizability theory. The Φ coefficient computed based on this data set was .853.

Two reduced data sets were created, respectively, using only the polytomous and dichotomous part of the test. For the part-test with 7 polytomous items (PT), the summed scores range from 0 to 21, and for the part-test with 40 dichotomous item (DT), the summed scores range 0 to 40. In the examples based on these reduced data sets, a binary classification (i.e., a pass/fail decision) is considered, and the pattern of the classification consistency indices was examined as the location of the cut score changed. Since each part-test consists of a single item type, the multinomial model was used to compute the indices. The summed score frequency distributions for PT and DT are displayed in Figure 1. Notice that the shapes of the frequency distributions for PT and DT summed scores are somewhat different in terms of skewness. As discussed later, the shape of the frequency distribution is closely tied to the pattern of the classification consistency indices as the location of the cut score changes.

4.2 Computer Program

A computer program MULT-CLASS (Lee, 2008) was used for computing the classification indices discussed in this paper. The Livingston and Lewis procedure was computed using the computer program BB-CLASS (Brennan, 2004). MULT-CLASS is, in a sense, an extension of BB-CLASS (Brennan, 2004) to multivariate situations. BB-CLASS is designed for estimating classification consistency and accuracy for a test with dichotomous items using a binomial dis-

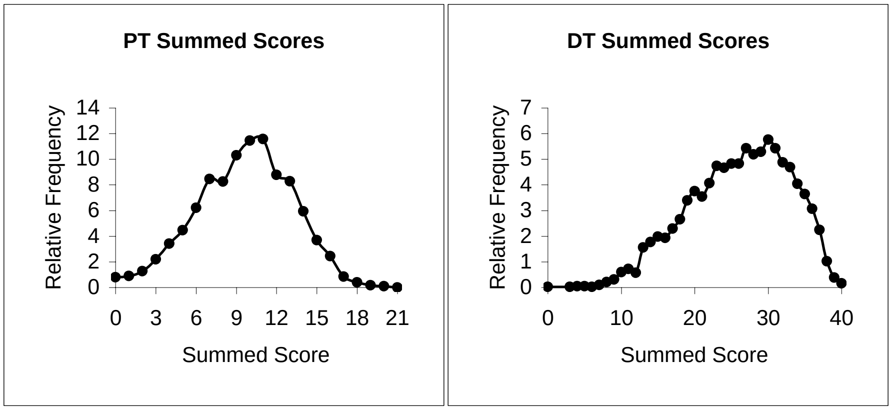

Figure 1: Frequency Distributions for Polytomous-Item Part-Test (PT) and Dichotomous-Item Part-Test (DT)

tribution for errors. (Note, however, that the Livingston and Lewis procedure implemented in BB-CLASS can deal with polytomous items through an ad hoc extension of the beta-binomial model.) By contrast, MULT-CLASS is designed for estimating classification consistency and accuracy for items with any number of score points using a multinomial or compound multinomial distribution for errors. Another difference between MULT-CLASS and BB-CLASS is that BB-CLASS estimates a true score distribution, whereas MULT-CLASS does not. In other words, BB-CLASS provides estimates based on the distributional approach, while MULT-CLASS produces estimates based on the individual approach.

5 Results

Results for the full-length test are summarized in Table 2. Estimates of the overall classification consistency indices (ϕ and κ) and overall classification accuracy indices (γ, γ^+, and γ^-) are displayed for each of the compound multinomial (CM), bias-corrected compound multinomial (CM$_\text{c}$), and Livingston and Lewis (LL) procedures.[2] Focusing on $\hat{\phi}$ and $\hat{\kappa}$, the result for CM is comparable with, but larger than the result for LL. According to Wan et al. (2007), the relative magnitudes of the consistency estimates for CM and LL depend on the location of the cut scores. The bias-corrected compound multinomial procedure tends to produce ϕ and κ estimates that are slightly smaller than those for CM, but still larger than those estimates for LL. In this particular example, the bias-correction procedure does not seem to make substantial corrections.

[2] The κ–type coefficients using the Livingston and Lewis procedure were computed for the comparative purposes only. The original literature for the Livingston and Lewis procedure (i.e., Livingston & Lewis, 1995) did not consider the κ–type coefficient.

Table 2: Estimated Overall Classification Consistency and Accuracy Indices

	$\hat{\phi}$	$\hat{\kappa}$	$\hat{\gamma}$	$\hat{\gamma}^+$	$\hat{\gamma}^-$
CM	0.727	0.564	0.805	0.096	0.099
LL	0.709	0.520	0.793	0.104	0.103
CM_c	0.721	0.536	0.800	0.114	0.086

Note. CM = compound multinomial procedure;
LL = Livingston and Lewis procedure;
CM_c = bias-corrected compound multinomial procedure.

The results for the γ estimates show similar patterns. Namely, CM produces an estimate that is larger than the estimate for LL, and CM_c gives a slightly smaller estimate, which is still larger than the estimate based on LL. However, the difference between the results for CM and LL is smaller than the difference observed for $\hat{\phi}$. The results for the estimated false positive and false negative error rates tend to show slightly larger differences among the three estimation procedures than the results for $\hat{\gamma}$. Note that the sum of $\hat{\gamma}$, $\hat{\gamma}^+$, and $\hat{\gamma}^-$ is necessarily equal to one.

Also notice that $\hat{\gamma}$ is always larger than $\hat{\phi}$. As discussed earlier, if an examinee is classified into a true category for which the model-predicted category probability (i.e., f_{ph}) is higher than the probability for the other categories (as should generally be the case), γ_p in Equation 9 will be larger than ϕ_p in Equation 5. This logic applies to both individual (i.e., CM) and distributional (i.e., LL) approaches.

The conditional classification indices, $\hat{\phi}_p$ and $\hat{\gamma}_p$, estimated using CM are plotted for examinees' actual total scores in Figure 2. The arrows on each plot indicate the location of the cut scores. Note that both $\hat{\phi}_p$ and $\hat{\gamma}_p$ show a wavy pattern along the summed-score values, which is primarily attributable to using multiple cut scores. The conditional estimates tend to show relatively low values near the cut scores, and increase as the score value moves further away from the cut scores. This finding is consistent with previous research (Lee et al., 2002). Unlike the results reported in Lee et al. (2002), in which items are all dichotomously scored, it can be noticed in Figure 2 that multiple values of conditional estimates are sometimes associated with a single total score. This is because different combinations of category proportion scores, when polytomous items are involved, can produce different conditional estimates even though they are associated with the same total score.

The overall classification consistency indices for PT and DT summed scores estimated using the multinomial model approach are plotted in Figure 3. The plot on the top displays the estimates of ϕ, κ, and ϕ_c for PT as a function of different cut scores, and the plot on the bottom is for DT. A few general observations can be made.

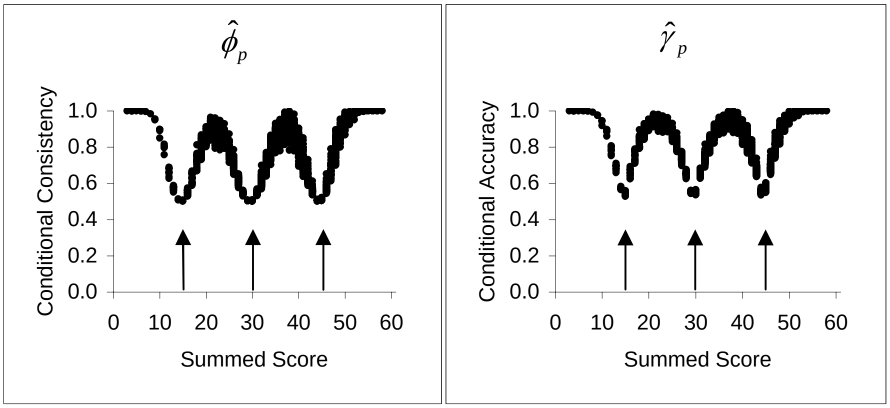

Figure 2: Estimated Conditional Classification Consistency and Accuracy Indices Using the Compound Multinomial Procedure

The relationship between the patterns of the estimated classification consistency indices and the shape of the frequency distributions shown in Figure 3 seems profound. In particular, for both PT and DT, the pattern of the ϕ estimates has an inverse relationship to the summed-score frequency distribution. That is, the higher the relative frequency at the cut score the lower the ϕ estimate becomes. This is not surprising in that the probability of being misclassified will be higher for examinees with scores near the cut score. If the cut score is set at a score value that is associated with a high relative frequency, there will be a large proportion of examinees who are likely to be misclassified, which will result in a relatively small ϕ estimate. It should not be assumed, however, that this is only the case for an individual approach. When the Hanson and Brennan (1990) procedure, which employs the distributional approach, was applied to DT and the same type of analysis was carried out, the results showed exactly the same patterns.

The chance probabilities tend to show a pattern that is similar (with more curvature) to the pattern of the ϕ coefficients, which suggests that a higher ϕ value is attributable to more chance agreement. This somewhat peculiar relationship between ϕ and ϕ_c causes the κ coefficients to show a pattern that is opposite to the pattern of the ϕ estimates, except for extreme cut scores. Wan et al. (2007) report a similar behavioral pattern of the κ coefficient as the location of the cut score changes.

6 Discussion

Classification consistency and accuracy are important considerations for a test that involves cut scores to make categorical decisions about examinees. Classification consistency and accuracy become particularly important in the context

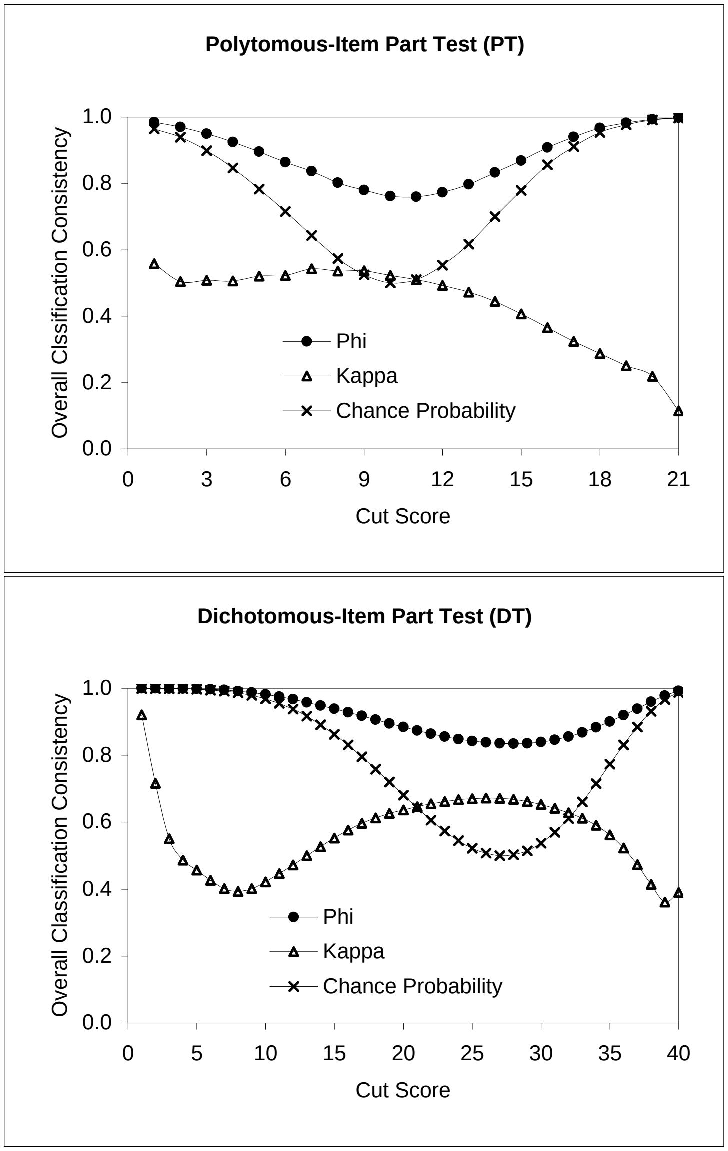

Figure 3: Classification Consistency Using the Multinomial Procedure for Polytomous-Item and Dichotomous-Item Part-Tests

of the No Child Left Behind Act (NCLB, Public Law 107–110). In recent years, many testing programs including licensure and certification programs have adopted a variety of innovative item types in their test forms to measure a broader set of skills. Psychometric properties of more complex assessments still need to be evaluated, and estimating classification consistency and accuracy indices based on a single administration of complex assessments certainly is not a simple matter. This paper provides a methodology for estimating classification indices for tests with complex item scoring using the multinomial and compound multinomial models. When all items in a test are scored dichotomously, the multinomial model procedure (without bias correction) reduces to Subkoviak (1976)'s procedure.

As an individual approach, the compound multinomial procedure starts with computing classification indices for each individual (i.e., conditional indices), and then group-level indices are obtained by taking an average of the conditional results over all individuals in the sample group. Thus, estimation of a true score distribution is not required. By contrast, a distributional approach, such as the Livingston and Lewis procedure, does involve estimating the true score distribution. For a distributional approach, conditional estimates can be computed for a set of discrete true score values (as opposed to single individuals for an individual approach), and the overall or marginal indices are obtained by taking a weighted sum of the conditional estimates over the set of true score values.

The conceptual definitions of classification consistency (between two observed classifications) and accuracy (between the true and observed classifications) are generally the same for the individual and distributional approaches. Aside from use of different statistical models, variations exist among different procedures in their estimation process. For example, the overall or marginal classification consistency could be estimated either (a) between classifications based on the model-predicted observed score distribution and the actual data or (b) between two model-predicted observed score distributions. Likewise, either the model-predicted observed score distribution or the actual data could be used as observed classifications for computing the marginal classification accuracy indices. The Livingston and Lewis (1995) procedure, as discussed in the original paper, employs the actual data for computing the classification indices. For example, it computes the marginal accuracy indices using the estimated true score distribution for true classifications and the actual data for observed classifications. By contrast, in this paper, the compound multinomial procedure uses the actual data for making true classifications and the model-predicted observed score distribution for making observed classifications. Obviously, the role of the actual classifications is flipped for the two procedures—they can serve as either the true or observed classifications. Although results of the present study do not reveal substantial differences between the two procedures, a useful future study would involve comparisons of various individual and distributional approaches with possible variations in terms of what role the actual classifications play in the estimation process.

Let us consider further the use of observed proportion scores as estimates for the true proportion scores. For the sake of simplicity, suppose items are scored dichotomously. The individual approaches for estimating classification consistency and accuracy including the one presented in this paper do not estimate the distribution of true scores, π, which makes the procedures relatively straightforward both theoretically and computationally. For those individual approaches, classification consistency and accuracy indices typically are computed conditioning on examinees' mean observed scores, $\overline{x}_p$, as estimates for true scores, and the overall indices are computed simply by taking averages of the examinee-level estimates over examinees. For each specific examinee, the individual approaches can provide meaningful, unambiguous, and unbiased classification consistency and accuracy estimates (e.g., ϕ_p and γ_p), partly because $\overline{x}_p$ is an unbiased estimator of π_p. However, when it comes to the overall classification consistency indices which involve an integration (or summation) over examinees, bias is introduced because the variance of mean observed scores (as estimates for true scores) is larger then the variance of true scores. (Note that an unbiased estimate of a parameter does not necessarily guarantee that the variance of the estimates over examinees is equal to the variance of the parameters.) Use of mean observed scores as estimates of true scores for computing *overall* indices is much like using an altered true score distribution with a larger variance. A bias-correction procedure has been developed (Brennan & Lee, 2006b), which suggests using an "optimal" estimate of true score so that the variance (over examinees) of estimates is equal to the variance of true scores. It is recommended that the bias-corrected (compound) multinomial procedure be used in most situations for computing the *overall* classification indices. However, it seems reasonable to use the original (compound) multinomial procedure for computing *conditional* classification estimates for each examinee, because the "optimal" estimate $\tilde{\pi}_p$ is a biased estimator of π_p for a single examinee.

A somewhat comprehensive comparative study has been conducted by Wan et al. (2007). They investigated five procedures for estimating overall classification consistency indices for tests consisting of both dichotomous and polytomous items under various real and simulated testing conditions. The procedures included a normal approximation procedure, the Breyer and Lewis procedure, the Livingston and Lewis procedure, the Brennan and Wan bootstrap procedure, and the compound multinomial procedure. The results of their simulation study showed that the accuracy of the procedures varied across different testing conditions. They also reported that, in general, the normal approximation and Livingston and Lewis procedures produced relatively more accurate classification consistency estimates than the bootstrap and compound multinomial procedures did. However, the bias-corrected compound multinomial and bootstrap procedures yielded much more accurate estimates that were comparable or better than results for the normal approximation and Livingston and Lewis procedures. Note that Wan et al. (2007) did not examine conditional classification consistency indices.

7 References

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. (2004). *BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy (Version 1.0)* (CASMA Research Report No. 9). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from http://www.education.uiowa.edu/casma).

Brennan, R. L., & Lee, W. (2006a). *Some perspectives on KR–21* (CASMA Technical Note No. 2). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from http://www.education.uiowa.edu/casma).

Brennan, R. L., & Lee, W. (2006b). *Correcting for bias in single-administration decision consistency indexes* (CASMA Research Report No. 18). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from http://www.education.uiowa.edu/casma).

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 http://www.education.uiowa.edu/casma).

Breyer, F. J., & Lewis, C. (1994). *Pass-fail reliability for tests with cut scores: A simplified method* (ETS Research Report No. 94–39). Princeton, NJ: Educational Testing Service.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

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.

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge: Harvard University Press.

Lee, W. (2007). Multinomial and compound multinomial error models for tests with complex item scoring. *Applied Psychological Measurement*, *31*, 255–274.

Lee, W. (2008). *MULT-CLASS: A computer program for multinomial and compound-multinomial classification consistency and accuracy (Version 3.0)*. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from http://www.education.uiowa.edu/casma).

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.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Subkoviak, M. J. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, *13*, 265–276.

Wan, L., Brennan, R. L., & Lee, W. (2007). *Estimating classification consistency for complex assessments* (CASMA Research Report No. 22). Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available from http://www.education.uiowa.edu/casma).

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