Center for Advanced Studies in
Measurement and Assessment

CASMA Research Report

Number 27

Classification Consistency and Accuracy for Complex Assessments Using Item Response Theory

Won-Chan Lee[†]

April 2008

[†] Won-Chan Lee is Associate Director, 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
FWeb: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

This paper describes procedures for estimating single-administration classification consistency and accuracy indices using item response theory (IRT). This IRT approach is applied to real test data comprised of mixtures of dichotomous and polytomous items. Several different IRT model mixtures are considered. Comparisons are made between the IRT approach and two non-IRT procedures.

1 Introduction

Classification consistency in the measurement literature refers to the degree to which examinees are classified into the same categories over replications (typically two) of the same measurement procedure. By contrast, *classification accuracy* refers to the extent to which actual classifications using observed cut scores agree with true classifications based on known true cut scores.

If data are available from two repeated measurement procedures, estimating classification consistency is straightforward. However, such data from repeated measurements are rarely available. Consequently, a number of procedures have been suggested in the literature for estimating classification consistency based on a single administration of a test, which typically involves estimating the observed score distribution by imposing a psychometric model on the test data. Some procedures have been developed mainly for dichotomous items (Huynh, 1976; Subkoviak, 1976; Peng & Subkoviak, 1980; Hanson & Brennan, 1990). Later, procedures were suggested for tests with polytomous items or mixtures of dichotomous and polytomous items, including procedures suggested by Breyer and Lewis (1994), Livingston and Lewis (1995), Brennan and Wan (2004), and Lee (2005).

Estimating classification accuracy typically involves estimating the true score distribution as well as the observed score distribution. The procedures discussed in Huynh (1976), Hanson and Brennan (1990), and Livingston and Lewis (1995) employ a family of beta distributions for estimating the true score distribution. Some procedures such as those discussed by Subkoviak (1976) and Brennan and Wan (2004) focus on each individual's classification status without considering the distribution of true scores as a whole, and thus typically are not used for estimating classification accuracy indices. However, Lee (2005) shows that classification accuracy can be computed without estimating the whole true score distribution.

Although there has been a great deal of research on the development of single-administration classification consistency and accuracy indices, only a few procedures have been developed under the framework of IRT. Huynh (1990) probably was the first who considered IRT for estimating classification consistency and accuracy indices for dichotomous items. Schulz, Kolen, and Nicewander (1999) also considered a dichotomous IRT model. Subsequently, Wang, Kolen, and Harris (2000) extended the IRT approach using polytomous IRT models.

In this paper, IRT is considered as a "general" framework for computing single-administration classification consistency and accuracy indices. The meaning of the term "general" is twofold: (1) the procedure can be used for a single IRT model or mixtures of different IRT models, and (2) estimated indices can be computed when cut scores are expressed in different score metrics including IRT proficiency scores (θ), summed raw scores, and transformed scale scores. Based on real data sets consisting of a mixture of dichotomous and polytomous items, the IRT procedure is illustrated for estimating results for summed raw scores using various mixed IRT models.

2 Classification Consistency and Accuracy

It is assumed that a test consisting of both dichotomous and polytomous items is scored by summing all item scores. These summed scores can be further transformed to scale scores for reporting purposes. Note that if a test is scored using θ (sometimes called a pattern scoring), the procedure discussed by Rudner (2001, 2005) can be used for computing classification accuracy indices. Note also that the model and indices presented in this section are generalizations of previous research including Schulz et al. (1999), Wang et al. (2000), and Lee, Hanson, and Brennan (2002). The formulas for the classification indices are presented in forms that are somewhat different from those in previous studies, however.

Let θ and $g(\theta)$ denote, respectively, the latent trait measured by a measurement procedure and its density. The marginal probability of the total summed score X is given by

$$\Pr(X = x) = \int_{-\infty}^{\infty} \Pr(X = x \,|\, \theta) g(\theta) d\theta. \tag{1}$$

Note that $\Pr(X = x \,|\, \theta)$ in Equation 1 is the conditional measurement error distribution, which is also called the conditional summed-score distribution. Due to the IRT assumption of conditional independence, the probability of a summed-score can be expressed by multiplication of probabilities for item responses given θ. Once the probabilities for item responses are obtained for a particular mixture of IRT models, the conditional summed-score distribution typically is modeled using a compound multinomial distribution (Kolen & Brennan, 2004). As a special case, when all items are scored dichotomously, a compound binomial model is used for modeling conditional number-correct score distributions. The compound-multinomial modeling implies that measurement errors are due to replications of a measurement procedure involving parallel forms with identical item parameters (Lee, Brennan, & Kolen, 2000). This issue is discussed further later.

Consider a transformation function $u(X)$, which converts the summed scores X to scale scores S. The conditional scale-score distribution is given by

$$\Pr(S = s \,|\, \theta) = \sum_{x:u(x)=s} \Pr(X = x \,|\, \theta), \tag{2}$$

where $x\!:\!u(x) = s$ indicates that the summation is taken over all x values such that $u(x) = s$. The marginal scale-score distribution, $\Pr(S = s)$, is obtained by integrating the conditional scale-score distribution over the θ distribution.

2.1 Classification Consistency Indices

Let $x_1, x_2, \ldots, x_{K-1}$ denote a set of observed cut scores that are used to classify examinees into K mutually exclusive categories. That is, examinees with an observed score less than x_1 are assigned to the first category; examinees with a score greater than or equal to x_1 and less than x_2 are assigned to the second category; and so on. Given the conditional summed-score distribution and the cut scores, the conditional category probability can be computed by summing conditional summed-score probabilities for all xs that belong to category h, as follows:

$$p_\theta(h) = \sum_{x=x_{(h-1)}}^{x_h - 1} \Pr(X = x \,|\, \theta), \tag{3}$$

where $h = 1, 2, \ldots, K$; the first category includes the minimum possible score; and the last category contains the maximum possible score.

The *conditional classification consistency index*, ϕ_θ, typically is defined as the probability that an examinee having θ is classified into the same category on independent administrations of two parallel forms of a test (Lee et al., 2002). Thus, ϕ_θ can be computed as

$$\phi_\theta = \sum_{h=1}^{K} \left[p_\theta(h)\right]^2. \tag{4}$$

The conditional classification consistency index quantifies classification consistency for different levels of θ. The *marginal classification consistency index*, ϕ, is given by

$$\phi = \int_{-\infty}^{\infty} \phi_\theta \, g(\theta) d\theta. \tag{5}$$

Another well-known index, the κ coefficient (Cohen, 1960), is computed as

$$\kappa = \frac{\phi - \phi_c}{1 - \phi_c}, \tag{6}$$

where ϕ_c is the chance probability. As typically defined (Cohen, 1960; Hunyh, 1976), the chance probability is computed by $\phi_c = \sum_{h=1}^{K} \left[p(h)\right]^2$, where $p(h)$ is the marginal category probability obtained by integrating the conditional category probabilities over the θ distribution.

If the cut scores are expressed on the θ metric, the θ cut scores can be transformed onto the summed-score metric, and the same procedures discussed thus far can be used. Expected summed scores can be obtained from the θ cut scores as

$$E(X \,|\, \theta = \theta^*) = \sum_i \sum_j j \Pr(U_i = j \,|\, \theta = \theta^*), \tag{7}$$

where θ^* is a θ cut score; U_i is a random variable representing responses for item i; $\Pr(U_i = j \,|\, \theta = \theta^*)$ is the conditional probability for score j for item i depending on the IRT model the item is associated with; the first summation is over all items; and the second summation is over all possible scores for item i. Equation 7 transforms all θ cut scores, $\theta_1^*, \theta_2^*, \ldots, \theta_{K-1}^*$, to summed-score cut scores, $x_1, x_2, \ldots, x_{K-1}$.

Now suppose a set of cut scores is expressed in the metric of scale scores that are transformed from summed scores based on $u(X)$. The conditional scale-score distribution is computed from Equation 2. Using the scale-score cut scores and conditional scale-score distribution, the conditional scale-score category probabilities, denoted here as $q_\theta(h)$, can be computed. Then, Equations 4 through 6 can be used to obtain the scale-score results by replacing $p_\theta(h)$ and $p(h)$ by $q_\theta(h)$ and $q(h)$, respectively. Note that the results for the summed and scale scores will be identical if the scale-score cut scores are determined directly from summed-score cut scores based on $u(X)$, and $u(X)$ is a one-to-one function at the cut-score levels. However, if $u(X)$ is a function in which several summed-score values convert to a single scale-score value, the results for the summed and scale scores will not be the same.

2.2 Classification Accuracy Indices

It is assumed that the conditional category probabilities $p_\theta(h)$ [$q_\theta(h)$ for scale scores] have been computed based on the observed cut scores using Equation 3. Now, suppose a set of *true* cut scores on the summed-score metric, $\tau_1, \tau_2, \ldots, \tau_{K-1}$, determine the true categorical status of each examinee with θ or τ (i.e., expected summed score). If the true categorical status, $\eta\,(= 1, 2, \ldots, K)$, of an examinee is known, the conditional probability of accurate classification is simply

$$\gamma_\theta = p_\theta(\eta), \quad \text{for } \theta \in \eta. \tag{8}$$

Note that the true category η can be determined by comparing the expected summed score for θ computed from Equation 7 with the true cut scores. Equation 8 is referred to here as the *conditional classification accuracy index.* It follows that the *marginal classification accuracy index*, γ, is given by

$$\gamma = \int_{-\infty}^{\infty} \gamma_\theta \, g(\theta) d\theta. \tag{9}$$

Classification accuracy is often evaluated by false positive and false negative error rates (Hanson & Brennan, 1990; Lee et al., 2002). The *conditional false positive error rate* is defined here as the probability that an examinee is classified into a category that is higher than the examinee's true category, which is expressed as

$$\gamma_\theta^+ = \sum_{\eta=\eta^*+1}^{K} p_\theta(\eta), \quad \text{for } \theta \in \eta^*. \tag{10}$$

By contrast, the *conditional false negative error rate* is the probability that an examinee is classified into a category that is lower than the examinee's true category, which is given by

$$\gamma_\theta^- = \sum_{\eta=1}^{\eta^*-1} p_\theta(\eta), \quad \text{for } \theta \in \eta^*. \tag{11}$$

The *marginal false positive and false negative error rates*, γ^+ and γ^-, respectively, are

$$\gamma^+ = \int_{-\infty}^{\infty} \gamma_\theta^+ \, g(\theta) d\theta, \tag{12}$$

and

$$\gamma^- = \int_{-\infty}^{\infty} \gamma_\theta^- \, g(\theta) d\theta. \tag{13}$$

When the true cut scores are set on the θ metric, Equation 7 can be used to find the true cut scores on the summed-score metric, $\tau_1, \tau_2, \ldots, \tau_{K-1}$. Then, the exact same process discussed above can be employed.

For the case in which the true cut scores are set on the scale-score metric, the true categorical status η for an examinee with θ can be determined by computing the expected scale score conditioning on θ, and comparing it with the true scale-score cut scores. The following formula can be used to compute the expected scale score for any θ value including the θ cut scores:

$$E(S \,|\, \theta) = \sum s \Pr(S = s \,|\, \theta), \tag{14}$$

where the summation is carried out over all possible s values, and $\Pr(S = s \,|\, \theta)$ is computed using Equation 2. Once η is determined, $q_\theta(h)$ can be used in place of $p_\theta(h)$ in Equations 8 through 13 to compute the scale-score results.

3 Estimation

Estimating the classification indices formulated in the previous section begins with computing the conditional summed-score distribution $\Pr(X = x \,|\, \theta)$ in Equation 1. Of course, it is assumed that IRT calibration has been done and item parameter estimates are available. Computation of the conditional summed-score distribution typically requires a recursive formula. Lord and Wingersky (1984) provide an algorithm that can be used for dichotomous IRT models. For polytomous IRT models, Hanson (1994) and Thissen, Pommerich, Billeaud, and Williams (1995) provide an extension of the Lord and Wingersky algorithm to polytomous items. The recursive algorithm for polytomous models can also be used for mixed IRT models as implemented in this paper. Kolen and Brennan (2004, p. 219) provide the recursive formulas and an illustrative example.

Differential weights sometimes are associated with different item formats so that the contribution of each item type to the total summed scores can be effectively manipulated. These weights can be incorporated in the recursive formula to obtain the summed-score distribution.

3.1 D and P Methods

Another practical issue is to approximate the integrals associated with the θ distribution for computing marginal results (e.g., Equations 5 and 9). Two different approaches are possible. One approach is to use estimated quadrature points and weights and replace the integrals by summations. For example, the estimated posterior distribution in the Phase II output from the PARSCALE computer program (Muraki & Bock, 2003) can be used. Another alternative would be to use the individual θ estimates. The conditional classification indices are computed first for each examinee and then averaged over all examinees to obtain the marginal results. Obviously, different θ estimates (e.g., EAP vs. MLE) would lead to different results. The former using estimated quadrature points and weights is called the D method and the latter using individual θ estimates the P method in this paper. The conditional classification indices for the D method typically are obtained conditioning on a set of θ quadrature points, while the P method provides estimated conditional classification indices for each individual examinee.

Brennan and Wan (2004) pointed out that single administration classification consistency can be estimated with or without assumptions about the true-score distribution. Lee (2005) called the approach using assumptions about the true-score distribution the distributional approach, and the one without the assumptions the individual approach. A distributional approach uses an estimated true-score distribution for computing marginal classification indices. By contrast, an individual approach estimates classification indices for a single examinee at a time and then averages over examinees to obtain a marginal classification indices (Brennan & Wan, 2004). Some distributional approaches include Huynh (1976) and Hanson and Brennan (1990). Examples of an individual approach are Subkoviak (1976), Brannan and Wan (2004), and Lee (2005).

Note that the IRT D method resembles the distributional approach, while the P method relates to the individual approach. It is illustrated in this paper that both IRT D and P methods can be used to estimate classification consistency and accuracy indices.

4 Method

The IRT procedure was applied to two real data sets, each of which consisted of test scores from a mixture of dichotomous and polytomous items. Various IRT model combinations were employed and results were compared.

4.1 Data Source

The first data set consisted of a multiple-choice test and a constructed-response test. Data for the multiple-choice test were from the Iowa Tests of Basic Skills (ITBS) of Math Problems and Diagrams for Grade 7 (Hoover, Hieronymus, Frisbie, & Dunbar, 1996a), and data for the constructed-response test were from the Constructed-Response Supplement to the Iowa Tests (Hoover, Hieronymus, Frisbie, & Dunbar, 1996b). The same group of examinees took the two tests on two different administrations. These two tests were designed to measure the same overall math problem solving ability. All 35 multiple-choice items had five alternatives, and 18 constructed-response items were rated on either a three-point scale (0-2) or a two-point scale (0-1). The sample size for this data set was 500. A pseudo cut score was set to a summed-score value of 38 for classifying examinees into pass or fail categories.

The second data set was from a science achievement test administered by a state government to approximately 4000 10th graders. The science test consisted of mixtures of 40 dichotomous items scored 0/1 and 7 polytomous items scored 0, 1, 2, or 3. Pseudo cut scores were set to the summed-score values of 15, 30, and 45 for four classification categories. Note that, for both math and science tests, the observed cut scores were treated as true cut scores for computing classification accuracy indices.

4.2 Analysis

Various mixtures of different IRT models were considered. Dichotomous models included the one-parameter logistic (1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL) models. Polytomous models included Masters' (Masters & Wright, 1997) partial credit (PC) model, Muraki's (Muraki, 1997) generalized partial credit (GPC) model, and Samejima's (Samejima, 1997) graded response (GR) model. The following six model mixtures were employed: 1PL+PC, 2PL+GPC, 3PL+GPC, 1PL+GR, 2PL+GR, and 3PL+GR. For all six model combinations, the dichotomous models were considered as special cases of corresponding polytomous models. Note also that, for the 1PL+GR combination, the discrimination parameters for the GR model were fixed at 1.0.

In addition, results from two non-IRT procedures, the Livingston-Lewis (Livingston & Lewis, 1995) and the compound multinomial (Lee, 2005) procedures, were computed and compared to the results from the IRT procedures. Note that a bias-correction method (Brennan & Lee, 2006a) was employed for the compound multinomial procedure. Among several non-IRT procedures are these two that can be used for mixed format tests.

The two types of items were calibrated simultaneously using PARSCALE (Muraki & Bock, 2003). Some of the options used for running PARSCALE included: the number of quadrature points was set to 101; the convergence criterion was set to .001; and the numbers of EM cycles and Newton steps were set to 200 and 0, respectively. All computations for the classification consistency and accuracy indices were carried out using the computer program IRT-CLASS (Lee & Kolen, 2008). Results for the compound multinomial model approach were computed using MULT-CLASS (Lee, 2008) and results for the Livingston-Lewis procedure were computed using BB-CLASS (Brennan, 2004).

As a preliminary analysis, the unidimensionality assumption was evaluated using principal components analysis (PCA) and Poly-DIMTEST (Li & Stout, 1995). Checking the assumption of unidimensionality was necessary especially for mixed format tests because of the potential presence of format effects. Furthermore, item fit statistics were examined for each item using all six model mixtures. In so doing, extended versions of the Orlando and Thissen's (2000)

$S - X^2$ and $S - G^2$ indices were employed.

5 Results

The results of the analysis are provided in three parts: (1) checking the unidimensionality assumption and the relative degree of fit of the various mixed IRT models to the data, (2) estimated marginal classification consistency and accuracy indices (i.e., $\hat{\phi}$, $\hat{\gamma}$, and $\hat{\kappa}$), and (3) estimated conditional classification consistency and accuracy indices (i.e., $\hat{\phi}_\theta$ and $\hat{\gamma}_\theta$).

5.1 Unidimensionality and Model Fit

Assessing unidimensionality of the data using PCA and Poly-DIMTEST led to somewhat inconsistent conclusions. The analysis of PCA suggested a single dominant dimension for both tests showing that the first component explained about 17% and 16% of the total variance for math and science, respectively, while the second component explained about 3% for both tests. The first three eigenvalues were 8.9, 1.7, and 1.6 for math; and 7.5, 1.4, and 1.3 for science. By contrast, the analysis of Poly-DIMTEST showed a statistically significant ($p < .05$) departure from essential unidimensionality for science. For math, the p-value was .06. Despite this inconsistency, the assessment of unidimensionality suggested that the data might be multidimensional, and care must be exercised in interpreting the results of the classification consistency and accuracy indices.

Table 1 presents the number of misfitting items evaluated by $S - X^2$ and $S - G^2$ for the six model combinations for the two tests. For the math test, the 2PL+GPC, 3PL+GPC, and 3PL+GR combinations showed the best fit having no misfitting items under the α level of .05. The 1PL+GR combination had 21% of the items that were flagged as misfitting. The science test tended to produce substantially more misfitting items than the math test for all model combinations. The 3PL+GPC showed the least number of misfitting items, and 3PL+GR showed the second best fit in terms of the number of misfitting items. Note also that the same model resulted in different identifications of misfitting items depending upon what models it was paired with.

The fit of the models was further evaluated by comparing the actual and model-predicted proportions of examinees for each of the classification categories. The actual proportions were computed by applying the cut scores to the actual data and counting the number of examinees classified into each category. The model-predicted proportions were the estimated marginal category probabilities, $\hat{p}(h)$. Table 2 summarizes the results. The last column in Table 2 shows the average (over all categories) of absolute differences between the actual and estimated proportions. Two observations were evident. First, the actual and estimated proportions for the math test were remarkably similar for all six model combinations, while results for the science test showed substantially worse fits. This was consistent with the results shown in Table 1. A larger number of classification categories for science might have also impacted the results

Table 1: Number of Misfitting Items ($p < .05$)

	$S-G^2$			$S-X^2$		
Model	DI	PI	Total(%)	DI	PI	Total(%)
			Math			
1PL+PC	6	4	10(19)	5	4	9(17)
2PL+GPC	0	0	0(0)	0	0	0(0)
3PL+GPC	0	0	0(0)	0	0	0(0)
1PL+GR	6	5	11(21)	7	5	12(23)
2PL+GR	0	1	1(2)	0	1	1(2)
3PL+GR	0	0	0(0)	0	0	0(0)
			Science			
1PL+PC	34	6	40(85)	34	5	39(83)
2PL+GPC	16	4	20(43)	16	4	20(43)
3PL+GPC	4	2	6(13)	4	2	6(13)
1PL+GR	34	5	39(83)	34	5	39(83)
2PL+GR	20	4	24(51)	21	3	24(51)
3PL+GR	8	3	11(23)	8	3	11(23)

Note. DI=dichotomous items; PI=polytomous items.

presented in Table 2. Second, the 2PL model combinations tended to provide a better fit than the 3PL model combinations. This result was inconsistent with the results for the item-fit statistics shown in Table 1, in which the 3PL model combinations tended to produce the least number of misfitting items. This issue is discussed later in the discussion section.

5.2 Marginal Classification Indices

The results for the estimated marginal classification consistency indices are presented in Table 3. In general, the results for the six model mixtures did not seem to differ substantially. The math test showed smaller differences than the science test across different model combinations. The values for $\hat{\phi}$ and $\hat{\kappa}$ were larger for math than for science, which seemed to be attributable to the smaller number of classification categories and better fits to the data for the math test. Also notice that results for D and P methods were very similar to each other–the maximum difference for $\hat{\phi}$ between the two methods was only .006 computed under the 2PL+GPC combination for math. The differences for $\hat{\kappa}$ values between the two methods tended to be larger than those for $\hat{\phi}$–the maximum difference was .015 for 2PL+GR of the math test.

Table 3 also presents the results for two non-IRT procedures (i.e., com-

Table 2: Actual and Estimated Proportions

Model	1st Cat.	2nd Cat.	3rd Cat.	4th Cat.	Avg Abs Diff
			Math		
Actual	.644	.356	—	—	—
1PL+PC	.641	.359	—	—	.003
2PL+GPC	.643	.357	—	—	.001
3PL+GPC	.649	.351	—	—	.005
1PL+GR	.642	.358	—	—	.002
2PL+GR	.643	.357	—	—	.001
3PL+GR	.649	.351	—	—	.005
			Science		
Actual	.016	.238	.551	.195	—
1PL+PC	.025	.246	.512	.217	.020
2PL+GPC	.015	.245	.543	.196	.004
3PL+GPC	.011	.257	.534	.198	.011
1PL+GR	.023	.246	.520	.211	.016
2PL+GR	.021	.268	.546	.166	.017
3PL+GR	.015	.280	.535	.169	.021

pound multinomial and Livingston-Lewis). Note that results for the compound multinomial procedure (an individual approach) are presented under the column of the P method, and results for the Livingston-Lewis procedure (a distributional approach) are under the column of the D method. As discussed earlier, this was because of the similarity of the D and P methods, respectively, to the distributional and individual approaches. An estimate of reliability was needed to obtain results for the Livingston-Lewis procedure. In this paper, reliability coefficients were computed using the compound multinomial model (Lee, 2007), which involves absolute error variance in the terminology of generalizability theory (Brennan, 2001). Note that, in the context of classification consistency, the focus is on the absolute location of each examinee's score relative to the cut scores, and the binomial error model used for modeling errors for the Livingston-Lewis procedure is closely related to absolute error variance (Brennan & Lee, 2006b).

The values of $\hat{\phi}$ and $\hat{\kappa}$ for the two non-IRT procedures were noticeably smaller than those for the IRT procedures, with the results for the compound multinomial procedure being less so. This discrepancy between the IRT and non-IRT procedures is largely attributable to the difference in the assumptions of the models. The assumption of the binomial and multinomial models for errors associated with the Livingston-Lewis and compound multinomial procedures, respectively, states implicitly that the test forms administered to examinees hy-

Table 3: Marginal Classification Consistency

	D Method			P Method		
Model	$\hat{\phi}$	$\hat{\kappa}$	$\hat{\phi}_c$	$\hat{\phi}$	$\hat{\kappa}$	$\hat{\phi}_c$
			Math			
1PL+PC	.881	.741	.540	.881	.741	.539
2PL+GPC	.879	.736	.541	.885	.750	.538
3PL+GPC	.881	.738	.544	.884	.749	.539
1PL+GR	.879	.736	.540	.879	.736	.540
2PL+GR	.879	.736	.541	.885	.751	.538
3PL+GR	.880	.737	.544	.884	.749	.539
CM	—	—	—	.862	.697	.545
LL	.855	.675	.554	—	—	—
			Science			
1PL+PC	.753	.609	.370	.749	.596	.378
2PL+GPC	.745	.579	.394	.744	.590	.376
3PL+GPC	.749	.588	.391	.746	.593	.376
1PL+GR	.748	.596	.376	.743	.584	.382
2PL+GR	.734	.558	.397	.733	.571	.378
3PL+GR	.737	.565	.394	.735	.573	.379
CM	—	—	—	.721	.534	.401
LL	.708	.515	.398	—	—	—

Note. CM=compound multinomial; LL=Livingston-Lewis.

pothetically over an infinite number of replications are *randomly parallel.* By contrast, the IRT procedures assume that the distribution for errors is obtained over an infinite number of replications of administering test forms having exactly identical item parameters, and in this sense, the test forms are said to be *strictly parallel.* The procedures with the weaker assumption about error scores should (and do) produce larger estimates for the classification consistency indices, $\hat{\phi}$ and $\hat{\kappa}$.

Results for the estimated marginal classification accuracy indices for the six IRT model combinations and the two non-IRT procedures are summarized in Table 4. The values of $\hat{\gamma}$ for the six model combinations did not seem to vary a lot, especially for math. Larger differences were found for the two error rates. The maximum differences of $\hat{\gamma}$, $\hat{\gamma}^+$, and $\hat{\gamma}^-$ among the IRT model combinations, respectively, were .014, .03, and .023, all of which were found for science. All procedures, except 1PL+GR for science, produced values of $\hat{\gamma}^+$ that were larger than the values of $\hat{\gamma}^-$. As for the marginal classification consistency in Table 3, results for the D and P methods were very similar. The values of $\hat{\gamma}$ were larger

Table 4: Marginal Classification Accuracy

	D Method			P Method		
Model	$\hat{\gamma}$	$\hat{\gamma}^+$	$\hat{\gamma}^-$	$\hat{\gamma}$	$\hat{\gamma}^+$	$\hat{\gamma}^-$
			Math			
1PL+PC	.916	.050	.034	.915	.045	.041
2PL+GPC	.914	.047	.038	.917	.051	.032
3PL+GPC	.914	.054	.032	.917	.051	.033
1PL+GR	.914	.049	.037	.912	.056	.033
2PL+GR	.914	.048	.038	.916	.054	.030
3PL+GR	.914	.054	.032	.917	.051	.033
CM	—	—	—	.900	.063	.037
LL	.897	.057	.046	—	—	—
			Science			
1PL+PC	.822	.109	.069	.816	.110	.075
2PL+GPC	.814	.117	.068	.815	.109	.075
3PL+GPC	.820	.105	.075	.817	.108	.075
1PL+GR	.822	.087	.091	.815	.108	.077
2PL+GR	.808	.116	.075	.807	.116	.077
3PL+GR	.812	.110	.079	.809	.112	.079
CM	—	—	—	.800	.114	.086
LL	.792	.109	.098	—	—	—

Note. CM=compound multinomial; LL=Livingston-Lewis.

for math than for science.

As discussed earlier, the estimated marginal classification accuracy index $\hat{\gamma}$ for the non-IRT procedures was smaller than that for any of the IRT procedures because of the differences in the assumptions about errors. Since the three indices $\hat{\gamma}$, $\hat{\gamma}^+$, and $\hat{\gamma}^-$ necessarily sum to one, the sums of the two error rates for the non-IRT procedures was larger than those for the IRT procedures.

5.3 Conditional Classification Indices

Figure 1 displays the estimated conditional classification index ($\hat{\phi}_\theta$) for different levels of the expected summed scores using the 2PL+GPC model combination for both tests. This particular model combination, 2PL+GPC, was chosen because it produced estimated category proportions that were most similar to the actual proportions. The expected summed scores were computed for 101 θ quadrature points using Equation 7. The arrows on the plot indicate the location of the cut scores for the two tests. Notice the wavy pattern of $\hat{\phi}_\theta$ values, in which the lowest values of $\hat{\phi}_\theta$ were always associated with expected summed scores

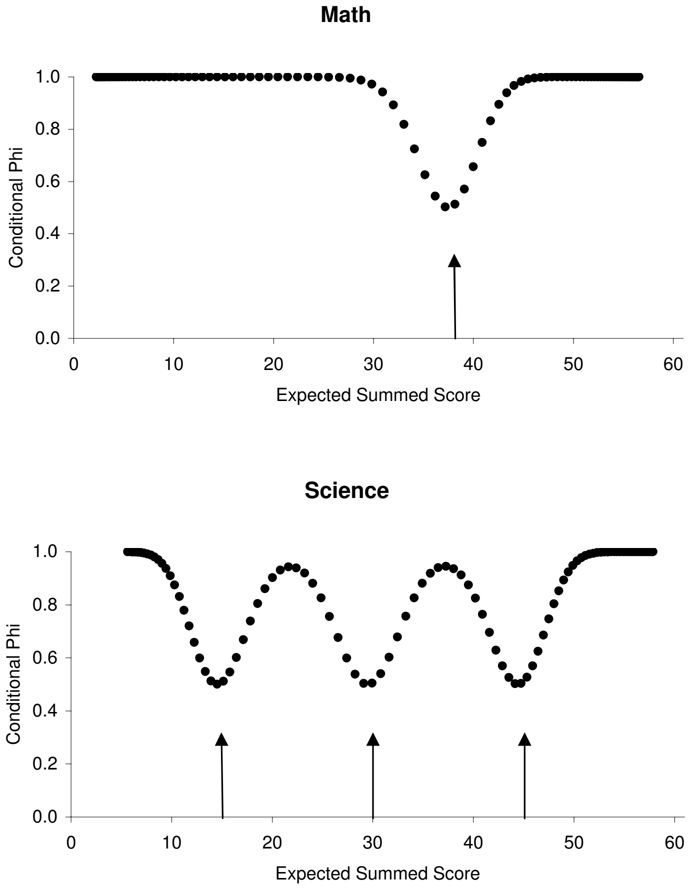

Figure 1: Estimates of ϕ_θ

near the cut scores. This is consistent with findings from previous studies (e.g., Lee, 2005; Lee et al., 2002). It is intuitively obvious that more classification errors will be made for examinees with observed scores near the cut scores. Likewise, classification errors will be less likely to be made for those examinees with observed scores that are farther away from the cut scores.

Estimates of conditional classification accuracy index, $\hat{\gamma}_\theta$, based on the 2PL+GPC model combination were plotted in Figure 2. The pattern of $\hat{\gamma}_\theta$ was similar to that of $\hat{\phi}_\theta$, except that $\hat{\gamma}_\theta$ tended to show a sharper increase as the expected summed scores move away from the cut scores.

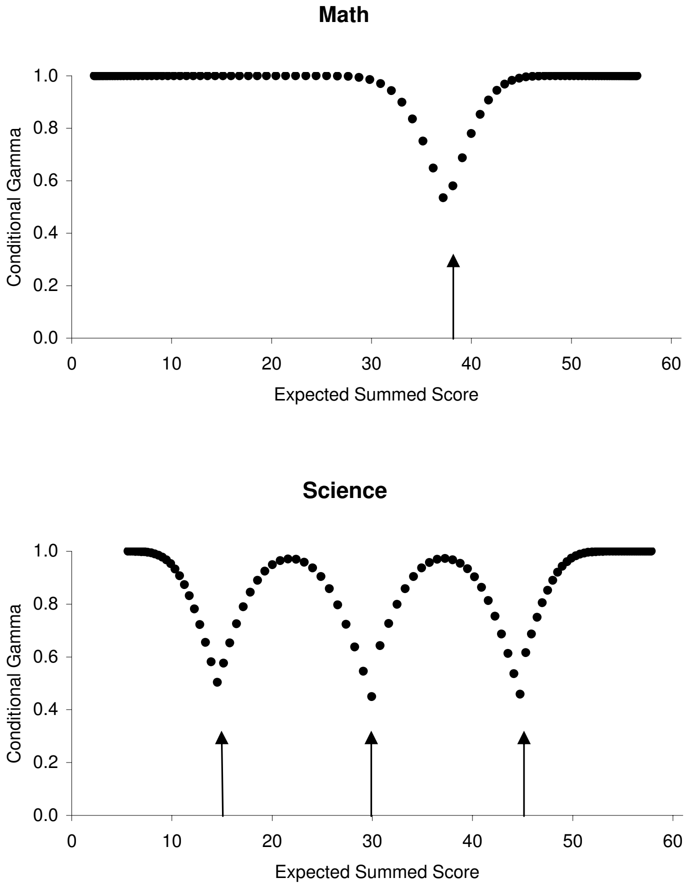

Figure 2: Estimates of γ_θ

6 Discussion

Building upon previous research (Huynh, 1990, Schulz et al., 1999, Wang et al., 2000, Lee et al., 2002), the principal purpose of the present paper is to present a general IRT procedure that can be used for estimating various classification consistency and accuracy indices for tests consisting of a single item type or a mixture of dichotomous and polytomous items. The IRT procedure is illustrated using two real data sets that consist of both dichotomous and polytomous item types. Six different IRT model combinations are considered, as well as two non-IRT approaches.

The results from this empirical study reveal that:

- Some (but not substantial) differences are found in the estimated marginal classification consistency and accuracy indices across different model combinations;

- both D and P methods produce very similar results;
- when compared to non-IRT procedures, the IRT procedures tend to produce larger estimated marginal classification consistency and accuracy indices; and
- estimated conditional classification consistency and accuracy indices show wavy patterns with the lowest values near the cut scores.

It should be emphasized that the focus of the paper is *not* on evaluating the degree of goodness-of-fit for various IRT model combinations. As exercised in this paper, comparison of actual and model-predicted category proportions could be used as an additional piece of information for selecting a model to analyze a particular data set. However, it should not be used as a main tool for evaluating IRT model fit or as a sole criterion for selecting a best fitting model. It is found in this study that the item-level evaluation of goodness-of-fit based on item fit statistics and the comparison of actual and estimated category proportions do not necessarily lead to a choice of the same model combination. For example, analysis of the science data shows that the 3PL+GPC combination produces the least number of misfitting items, while the 2PC+GPC combination results in the best fitting category proportions. Although it is not completely unreasonable to choose a model combination with estimated proportions that are closest to the actual ones in the context of estimating classification indices, other model selection criteria should also be considered (e.g., Kang, Cohen, & Sung, 2005).

It is suggested that the D method is tied to a distributional approach and the P method relates to an individual approach. Although the estimated classification consistency and accuracy indices are found to be similar for the two methods, the individual approach (i.e., P method) might be preferred when the focus of an examination is on each individual examinee. One such example would be computer adaptive testing. By contrast, the distributional approach (i.e., D method) might be preferred when the main interest is on computing group-level statistics. A further investigation, however, is necessary.

In this study, the assumption of IRT unidimensionality is made for analyzing data from mixed format tests. Previous research (e.g., Ercikan et al., 1998) indicates, however, that the underlying factor structure of a mixed format test can be multidimensional. If a mixed format test is believed to be truly multidimensional due to the format effects, a multidimensional IRT model could provide a basis for estimating the indices discussed in this paper. Notwithstanding some theoretical and empirical evidence of multidimensionality for data from a mixed format test, the use of mixed IRT models under the unidimensionality assumption could be justifiable from a practical point of view, if a mixed format test is designed to measure a “dominant” underlying dimension with some possible minor dimensions—this is sometimes called essential unidimensionality (Stout, 1990).

7 References

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). [Computer software and manual]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Brennan, R. L., & Lee, W. (2006a). *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 on http://www.education.uiowa.edu/casma).

Brennan, R. L., & Lee, W. (2006b). *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 on 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 on 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.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35*, 137–154.

Hanson, B. A. (1994). *An extension of the Lord-Wingersky algorithm to polytomous items.* Unpublished research note.

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.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1996a). Iowa Tests of Basic Skills: Form M: Levels 13–14. Itasca, IL: Riverside Publishing.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (1996b). Constructed-Response Supplement to The Iowa Tests Form 1: Levels 13–14. Itasca, IL: Riverside Publishing.

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

Huynh, H. (1990). Computation and statistical inference for decision consistency indexes based on the Rasch model. *Journal of Educational Statistics, 15*, 353–368.

Kang, T., Cohen, A. S., & Sung, H.-J. (2005, April). *IRT model selection methods for polytomous items.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

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

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

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

Lee, W., & Kolen, M. J. (2008). *IRT-CLASS: A computer program for item response theory classification consistency and accuracy (Version 2.0).* Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

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

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

Li, H.-H., & Stout, W. F. (1995, April). *Assessment of unidimensionality for mixed polytomous and dichotomous item data: Refinements of Poly-DIMTEST.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Measurement in Education, 8*, 452–461.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–185). New York: Springer-Verlag.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–185). New York: Springer-Verlag.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4: IRT item analysis and test scoring for rating-scale data.* [Computer program]. Chicago, IL: Scientific Software International, Inc.

Peng, C. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17*, 359–368.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment Research & Evaluation*, 7(14). Available online: http://pareonline.net/getvn.asp?v=7&n=14.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment Research & Evaluation*, 10(13). Available online: http://pareonline.net/getvn.asp?v=10&n=13.

Samejima, F. (1997). Graded response model. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer-Verlag.

Schulz, E. M., Kolen, M. J., & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23*, 347–362.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293–325.

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

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. L. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39–49.

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.