Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 32

Measurement Error Variability for Advanced Placement (AP) Composite Scores and Grades

Benjamin J. Andrews[†]
Michael J. Kolen
Won-Chan Lee

January 2011

A previous version of this paper was presented at the Annual Meeting of the National Council on Measurement in Education, Denver, May, 2010.

[†]Benjamin Andrews is a research assistant in Educational Measurement and Statistics, College of Education, University of Iowa (email: benjamin-andrews@uiowa.edu). Michael J. Kolen is Professor, 224B1 Lindquist Center, University of Iowa, Iowa City, IA, 52242 (email: michael-kolen@uiowa.edu). Won-Chan Lee is Associate Professor and Associate Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, IowaCity, 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

When a test contains both multiple-choice and free-response items, there is additional complexity in estimating conditional standard errors of measurement. This article compares different methods for estimating conditional standard errors of measurement for mixed-format tests. Three different procedures are used including a unidimensional IRT procedure (Wang, Kolen & Harris, 2000), a multidimensional IRT procedure (Kolen & Wang, 1998, 2007) and a compound multinomial procedure (Lee, 2007). Conditional standard errors of measurement and reliabilities for both composite scores and AP grades are estimated using the three methodologies. The effects of different IRT models and different weighting schemes are also investigated.

1 Introduction

Standard errors of measurement quantify the amount of measurement error in test scores. The magnitude of measurement error typically varies depending on true score for both raw and scale scores. These are called conditional standard errors of measurement (CSEMs). Composite scores for most Advanced Placement (AP) examinations are constructed from both multiple-choice items and free-response items. This use of mixed item types complicates the computation of CSEMs for these composite scores. In addition, because AP scores are reported as grades of 1, 2, 3, 4, and 5, the CSEMs for AP grades are also important to consider. Given that there are so few grade categories, it also seems reasonable to estimate conditional measurement error variability in terms of the conditional distribution of grades given proficiency.

Various methods can be used to estimate CSEMs for composite scores and grades for AP examinations. An IRT method was developed by Wang, Kolen, and Harris (2000) that assumes a single unidimensional IRT model can be used to fit both multiple-choice and free-response items. This model is implemented using the publicly available computer program POLYCSEM (Kolen, 2004). This method is referred to here as the unidimensional IRT method. Another IRT-based method was developed by Kolen and Wang (1998, 2007) in which separate unidimensional IRT models are fit to items for each of the item types. The composite score is calculated as a composite of the scores from the different item types. This method is referred to here as the multidimensional IRT method. Lee (2007) presented a non-IRT approach to estimating conditional standard errors of measurement for situations in which there are multiple-choice and free-response items. Lee's (2007) approach uses a multinomial error model when all items are scored using the same polytomous categories and a compound multinomial error model is used when a test contains both polytomous and dichotomously scored items. This method is referred to here as the compound multinomial method. In addition to being able to estimate CSEMs, each of these methods can be used to estimate reliabilities of AP composite scores and grades.

In the present paper, estimates of CSEMs are compared using different methods. The effects of various conditions such as IRT model and weighting scheme on CSEMs are also investigated.

2 Methodology

2.1 Data

The analyses were conducted using data from the AP Biology, AP World History and AP English Language and Composition tests. Two forms of each test were used. Because formula scoring is used for the operational administration of the AP tests, certain changes needed to be made to use these three methods. Not having examinees respond to all items complicates the use of IRT. To deal with this issue, examinees were included for the current study only if they responded to all the items. Also, number-correct scoring was used for the multiple-choice section. This was done for both forms of all three tests and the same examinees were used for both the IRT procedures and the compound multinomial procedure. The resulting sample sizes for the 2004 and 2005 AP World History test forms were 5,952 and 6,136 respectively. The sample sizes for the 2004 and 2006 AP Biology test forms were 2,739 and 3,517 respectively. The AP English Language and Composition test forms had sample sizes of 6,304 for 2004 and 6,709 for 2007.

Composite scores on the AP tests are weighted summations of scores on the multiple-choice section and the summed scores on the free-response items. The weights used operationally typically are non-integer. The computer programs that were used in this study allow only for integer weights. Two different weighting schemes were compared in the current research. The first involves rounding the weights for each section to the nearest integer. These are referred to here as rounded weights. The resulting weights were 1 for each item in the multiple-choice sections and 2 for each item in the free-response sections for the World History and Biology AP tests. These differ slightly from the operational weights. For the English Language and Composition test, the multiple-choice section was given a weight of 1 and the free-response section was given a weight of 3. These too differ slightly from the weights used in the operational administrations. The second weighting scheme is referred to here as integer weights. Instead of rounding weights to the nearest integer, weights were chosen so that the proportion of total points for each section mirrors the operational conditions as closely as possible. Integer weights were 3 for the multiple-choice items and 8 for the free-response items for the AP World History test, 3 and 5 for the AP Biology test and 2 and 5 for the AP English Language and Composition test.

To convert both types of composite scores to AP grades, cut scores for the AP grades were set so that the percentages of examinees that were given each AP grade in the study were as close as possible to the actual percentages in the operational administrations.

2.2 The Unidimensional IRT Method

In unidimensional IRT, it is assumed that examinee proficiency is described by a single variable, θ. For polytomously scored tests in unidimensional IRT, the probability of a particular response $V_i = v_i$ to item i conditional on θ is symbolized $P(V_i = v_i|\theta)$ and is referred to as the category characteristic function. Based on examinee data, category characteristic functions are estimated for each score category for each item and proficiency is estimated for each examinee. IRT estimation requires a local independence assumption in which, conditional on proficiency, examinee responses are assumed to be independent. Dichotomously scored items can be viewed as polytomously scored items with two score categories (wrong and right). In this paper, the three-parameter logistic model was used for dichotomous items. For the polytomous items, either the generalized partial credit model (Muraki, 1992) or the graded response model (Samejima,

1997) was used.

The general process for calculating CSEMs for the unidimensional IRT method is as follows: First, a distribution of raw scores (composite scores) is found conditional on θ. The standard deviation of this distribution is the CSEM for raw scores. The CSEM for AP grades is estimated using a similar process. The raw scores conditioned on θ are converted to AP grades. The standard deviation of this distribution is considered to be the CSEM for AP grades. The unidimensional IRT method assumes that a single θ underlies responses to both the multiple-choice and free-response items.

The raw scores that are investigated in this paper are calculated by finding a weighted sum of the item scores using integer weights. The weighted summed score is

$$X = \sum_{i=1}^{n} w_i V_i, \tag{1}$$

where w_i is an item weight and n is the number of items on the test. Scale scores (AP grades) are a function of these weighted summed scores, so that $S = S(X)$. Given IRT proficiency and item parameter estimates for all items on a test, the conditional probability of a particular integer-weighted summed score on a test is defined as $P(X = x|\theta)$ and can be calculated using a recursive algorithm (Hanson, 1994; Thissen, Pommerich, Billeaud & Williams, 1995). The true scale score (i.e., expected scale score) is

$$\tau_{S|\theta} = \sum_{j=minX}^{maxX} S(j) \cdot P(X = j|\theta), \tag{2}$$

where $minX$ and $maxX$ are the minimum and maximum integer-weighted summed scores. Conditional error variance of scale scores is

$$\sigma^2_{S|\theta} = \sum_{j=minX}^{maxX} \left[S(j) - \tau_{S|\theta}\right]^2 \cdot P(X = j|\theta). \tag{3}$$

The conditional standard error of measurement for scale scores is the square root of this variance. Reliability of scale scores is defined as

$$\rho_{SS'} = 1 - \frac{\int\limits_{\theta} \sigma^2_{S|\theta} g(\theta)\, d\theta}{\sigma^2_S}, \tag{4}$$

where σ^2_S is the variance of scale scores in the population. Note that if the identity function is substituted for S in Equations 2 through 4, the resulting quantities are for the raw scores, X.

2.3 The Multidimensional IRT Method

Kolen and Wang (1998, 2007) generalized the unidimensional procedures to be applicable to tests that are fit with multidimensional IRT models. For this generalization, the proficiency parameter, θ, is replaced by a vector of proficiency parameters in Equations 2 through 4. Also, the integral in Equation 4 is replaced by multivariate integrals over each of the proficiencies.

The three-parameter logistic model was used for the multiple-choice items and either the generalized partial credit model or the graded response model was used for the free-response items as they were for the unidimensional method. The parameters were estimated using separate calibrations of the computer program PARSCALE (Muraki & Bock, 1993).

The procedure described by Mislevy (1984) was used to estimate the multivariate distribution of the proficiencies. Observed data are used in this procedure to estimate the correlations between the θ for the multiple-choice section and the θ that represents the free-response section. The simulation approach described by Kolen and Wang (1998, 2007) was used in the present research. In this approach, it is assumed that the joint distribution is bivariate normal with means of 0 and standard deviations of 1. A random variable was then drawn from the population and the CSEM was calculated. This process was done a total of 5,000 times. As a result, there were 5,000 simulees with CSEMs.

2.4 The Compound Multinomial Method

The compound multinomial method has assumptions that differ from the IRT methods. First, the standard errors of measurement are conditioned on each individual instead of on θ. Under the compound multinomial assumptions, the dichotomous items make up one set, and the polytomous items the other. For the dichotomous item set, $\boldsymbol{\pi_1} = (\pi_1, \pi_2)$ represents the proportion of items in the universe that an individual would get scores of 0 and 1. For the polytomous item set, there is a $\boldsymbol{\pi_2} = (\pi_1, \pi_2, \ldots, \pi_k)$ that represents the proportion of items in the universe that an individual would get scores of $c_1, c_2, \ldots, c_k$, where k is the number of response categories for the polytomous items. Raw scores for each item set (X_1 for the dichotomous items and X_2 for the polytomous item set) are represented by the sum of item scores over all items in the item set. The weights for the dichotomous items and the polytomous items are represented by w_1 and w_2, respectively.

If it is assumed that errors across item sets are uncorrelated, the conditional summed score error variance for an individual is obtained by

$$\sigma^2_{E|\boldsymbol{\pi_1},\boldsymbol{\pi_2}} = w_1^2\sigma^2_{E_1|\boldsymbol{\pi_1}} + w_2^2\sigma^2_{E_2|\boldsymbol{\pi_2}}, \tag{5}$$

where $\sigma^2_{E_1|\boldsymbol{\pi_1}}$ and $\sigma^2_{E_2|\boldsymbol{\pi_2}}$are the conditional error variances for the two item sets. The conditional error variance for the polytomous item set can be obtained as

$$\sigma^2_{E_2|\boldsymbol{\pi_2}} = n_2\sum_{i=1}^{k} c_i^2\pi_i(1-\pi_i) - 2n_2\sum_{i<j}\sum c_ic_j\pi_i\pi_j, \tag{6}$$

where n_2 is the number of items in the polytomous item set, and π_i is the proportion of items for which the individual would receive a score of c_i in the universe of items. The results for the dichotomous item set can be obtained in the same manner. The square roots of Equations 5 and 6 give CSEMs.

Lee (2007) discussed two different approaches for calculating the CSEMs for scale scores. For this study, the direct approach, which is based on the distribution of scale scores, was used. Let $P(S|\boldsymbol{\pi_1}, \boldsymbol{\pi_2}) \equiv f(S)$ be the conditional scale score distribution for an individual. The conditional scale score distribution is computed using the conditional raw score distribution, which is modeled by the compound multinomial distribution (see Lee, 2007 for more details). The conditional error variance for the individual can be expressed as

$$\sigma^2_{S|\boldsymbol{\pi_1},\boldsymbol{\pi_2}} = \sum S^2 \cdot f(S) - \left[\sum S \cdot f(S)\right]^2, \tag{7}$$

where the summations are taken over all possibile scale score points. The CSEM for scale scores is the square root of this variance.

Equations 5 through 7 can be estimated by replacing π parameters by observed proportion scores. As for the IRT methods, reliability for both raw scores and scale scores can be calculated by subtracting the ratio of average error variance to total score variance from one.

3 Results

Selected results are shown in this section. In many instances, results were similar across test forms. The following results are those that are representative of important findings over the various conditions.

An attempt was made to fit the multidimensional IRT model for all tests. However, the IRT proficiencies for the multiple-choice and the free-response items were too highly correlated to be able to be fit by the multidimensional IRT model for the AP Biology and AP World History tests. Therefore, the multidimensional IRT model was fit only for the two AP English Language and Composition forms.

The CSEMs for raw scores using the unidimensional IRT method and rounded weights are shown in Figure 1 for both forms of the AP Biology and English Language and Composition tests. The patterns of CSEMs were similar for both IRT models. For both the AP World History (not shown) and AP Biology tests, the CSEMs when the graded response model was used were slightly higher than those when the generalized partial credit model was used throughout the score range. The CSEMs were quite similar at the extreme true score values and also in the middle of the true score range as shown in the top two graphs of Figure 1. For the AP English Language and Composition test, the CSEMs when the graded response model was used were somewhat smaller for lower true scores and slightly larger for the higher true score values as shown in the bottem two graphs of Figure 1.

Figure 2 shows CSEMs for AP grades for the 2004 AP World History and 2004 AP English Language and Composition tests, again using rounded weights. There appeared to be very little difference in CSEM estimates for AP grades

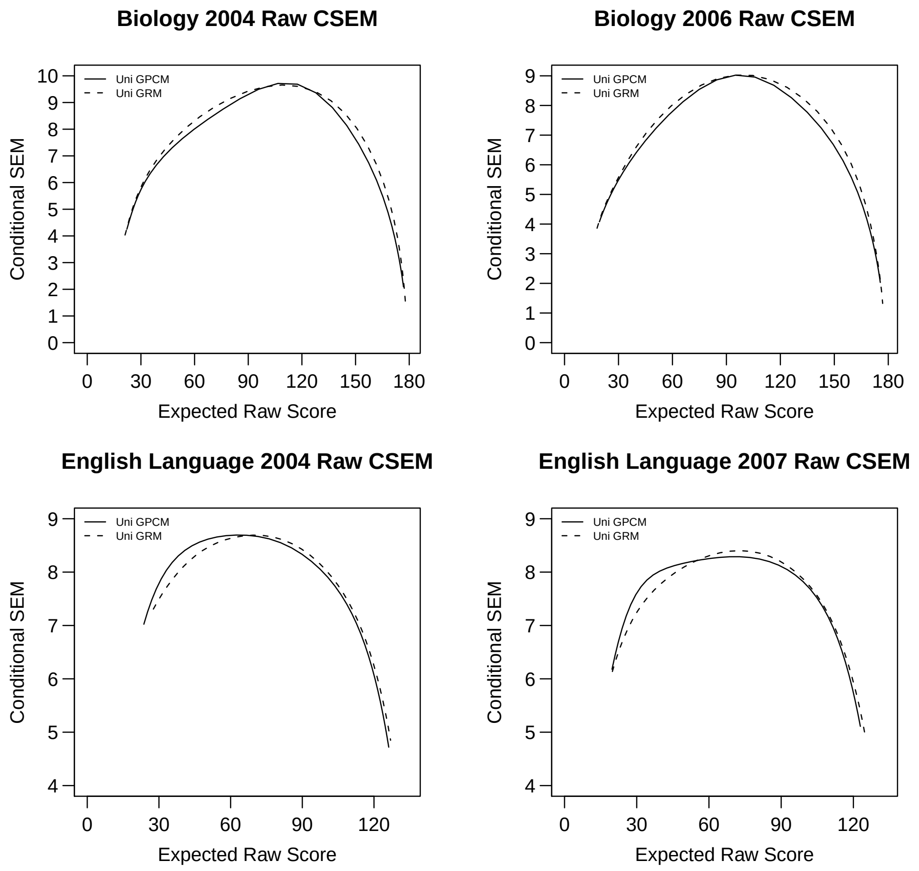


Figure 1: Raw score CSEMs for the unidimensional IRT method using rounded weights.

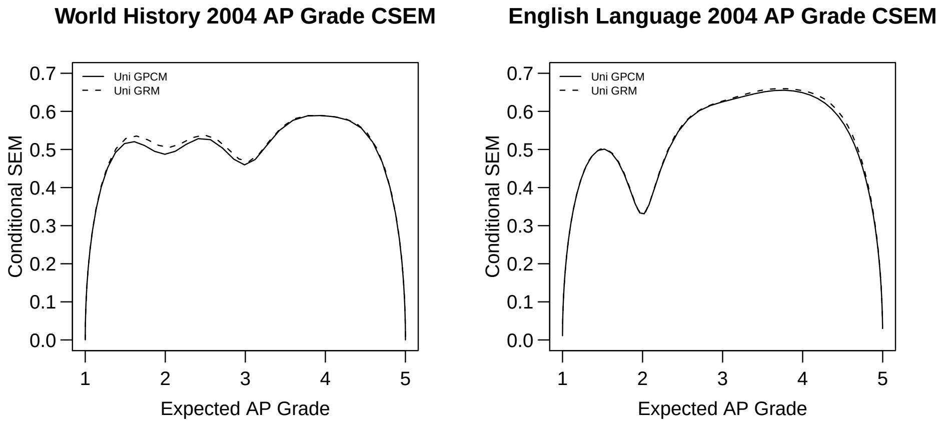


Figure 2: Unidimensional IRT AP grade CSEMs using rounded weights.

for the two polytomous IRT models. There were only slight differences between the two models at the lower AP Grades for the World History test forms. For both the AP Biology and AP English Language and Composition test forms, the patterns were very similar for both models. Note that the graphs of the CSEMs are a bit wavy. Relatively lower CSEMs tend to occur at integer true AP grades, whereas relatively higher CSEMs occur at true scores that are in the middle of two integer true AP grades. This pattern is likely due to the rounding of AP grades to integers.

Figure 3 shows the fitted and observed score distributions for one form of each of the three tests using rounded weights. Comparing the observed distributions with the raw score distribution based on the IRT model serves as a partial check of model fit. Results were similar across both forms of each test. The fitted distributions differed slightly based on use of the generalized partial credit model or the graded response model. Though there are differences, the model appears to fit the observed data reasonably well for the AP World History and AP Biology tests.

For the English Language and Composition test, the unidimensional fit was virtually identical regardless of polytomous IRT model. The fitted distribution for the multidimensional IRT method differed from that for the unidimensional IRT model as shown in the bottom of Figure 3. The multidimensional IRT method appears to fit somewhat higher relative frequencies than the observed frequencies at the high end of the score scale and lower relative frequencies in the middle of the score distribution. This result was similar to that for the other form of the English Language and Composition test (not shown).

Raw score and AP grade CSEMs using the compound multinomial method are shown in Figures 4 and 5, respectively, using rounded weights. Results from only two tests are shown. The top two graphs in Figure 4 show the individual CSEMs for raw scores. There is a great deal of scatter because many different

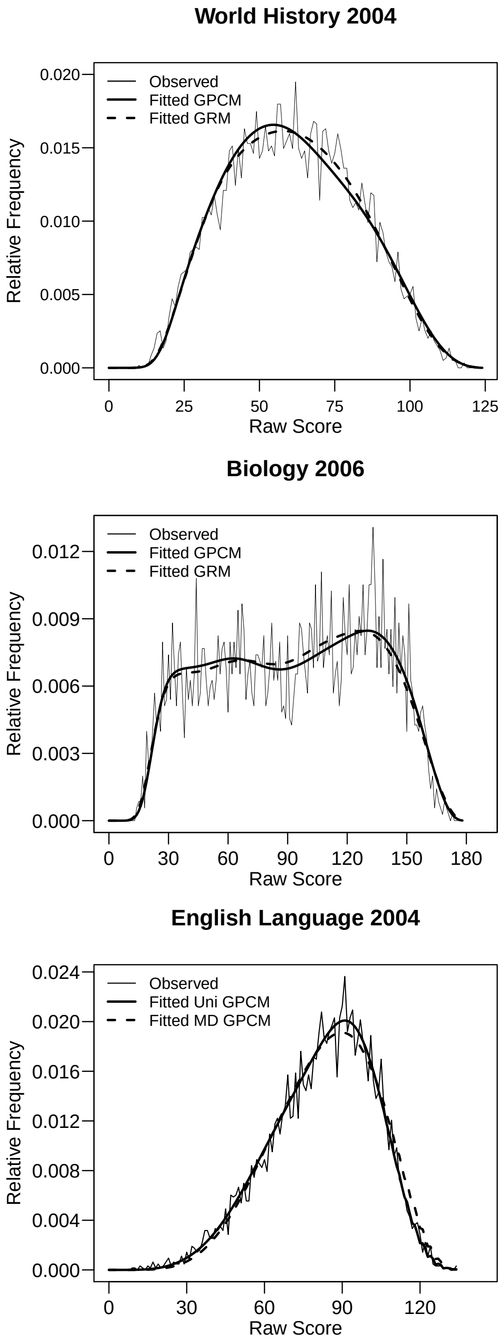


Figure 3: Fitted and observed score distributions using rounded weights.

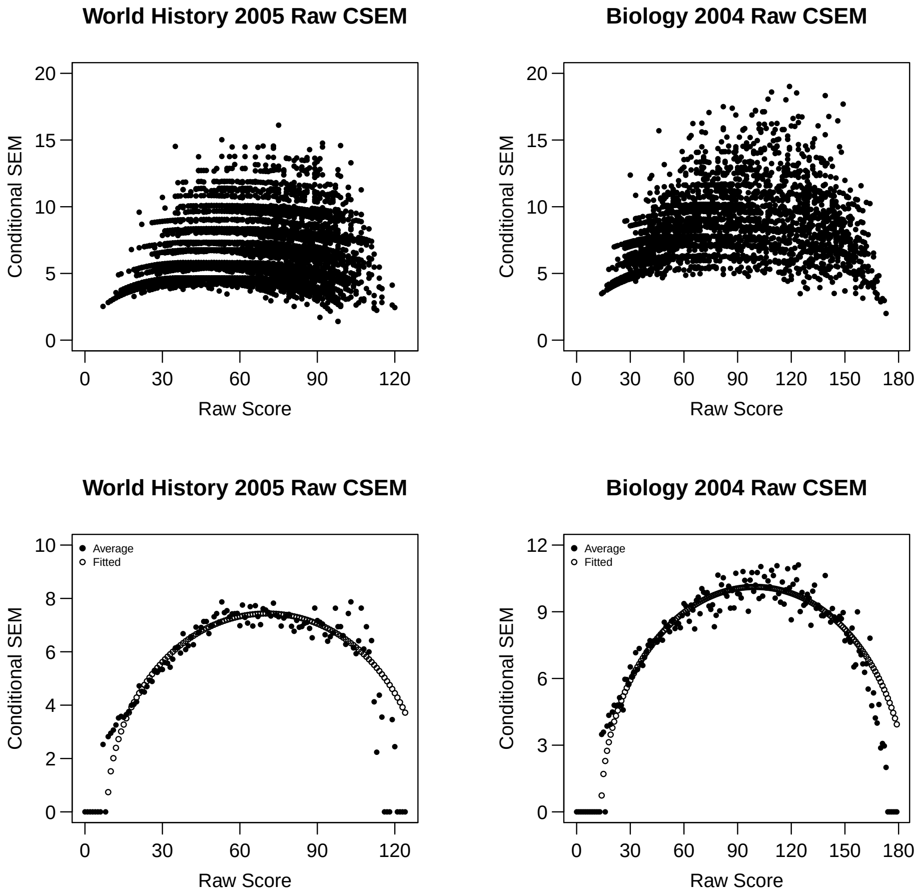


Figure 4: Raw score CSEMs for the compound multinomial method using rounded weights.

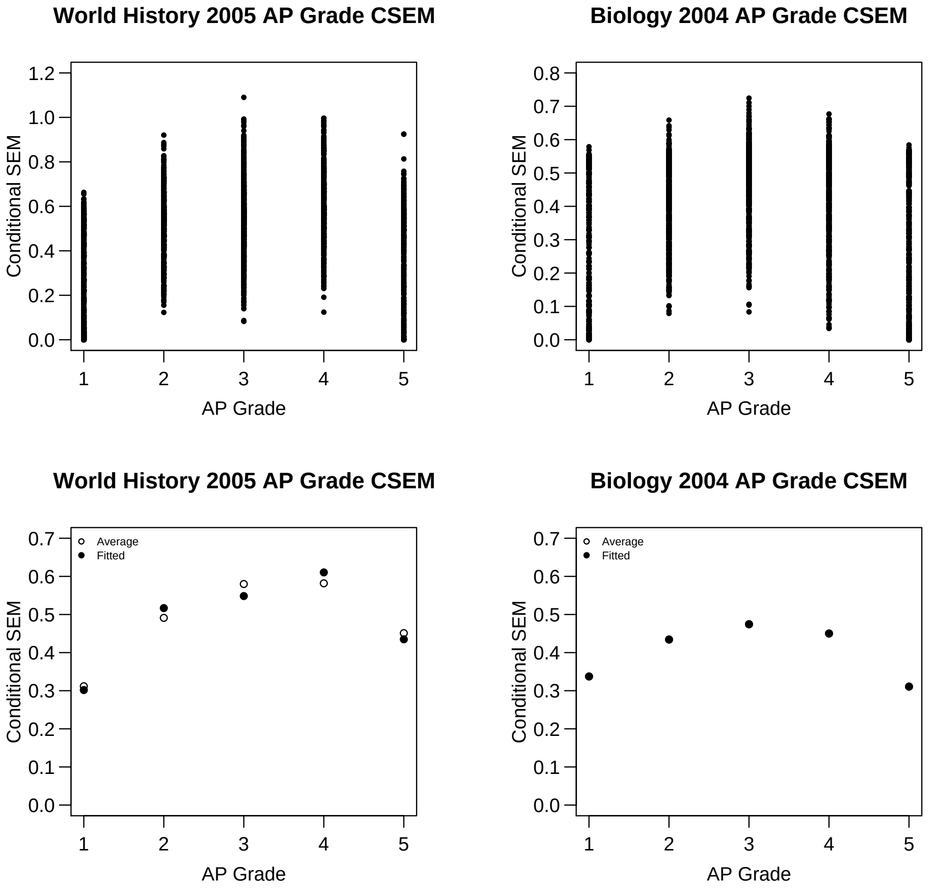


Figure 5: AP grade CSEMs for the compound multinomial method using rounded weights.

combinations of summed scores on the multiple-choice and free-response sections can result in the same total raw score. These results were typical of the other tests as well. The AP English Language and Composition test had patterns very similar to the AP World History example shown. Because some situations warrant that one estimate of a CSEM be reported for each score point, two different ways of doing this are shown in the bottom two graphs of Figure 4. The average CSEM values are the square root of the average error variances. To obtain the fitted CSEM values, conditional error variances were fit with a second degree polynomial. The square root of the fitted conditional error variances are the fitted CSEM values. Both the fitted and average estimates were similar with the exception of the high end of the score scale. This is probably attributable to the fact that few examinees scored at the very high end of the score scale.

Figure 5 shows the compound multinomial results for AP grades using rounded weights. The individual AP grade CSEMs for two tests are plotted in the top two graphs. These were similar across all forms of all tests. At each of the five AP Grade levels, there are many overlapping CSEM values that cause the appearance of a vertical line. Note that for this method, the conditioning variable is the observed AP Grade so CSEM values only exist at each of the five score categories. The average and fitted AP Grade CSEMs are shown in the bottom two graphs. The fitted CSEMs are the square roots of the fitted conditional error variances. A third degree polynomial was chosen because there was very little change in R-square values when higher degrees were used. The fitted and average values appeared to be quite similar. There were small differences for the AP World History and AP English Language and Composition tests. Estimates of CSEMs for the AP Biology tests appeared to be nearly identical for both the fitted and average values.

The weights for the multiple-choice and free-response sections had some effect on the pattern of the CSEMs. For the AP World History test, more weight was given to the free-response section under the integer weighting scheme compared to the rounded weighting scheme. For the AP Biology and AP English Language and Composition tests, more weight was given to the multiple choice section for the integer weighting scheme than under the rounded weighting scheme. Changing the weighting scheme altered the pattern of CSEMs differently for each test. Examples using the unidimensional IRT method are shown in Figure 6. For instance, the 2004 English Language and Composition form had similar patterns of CSEMs for both weighting schemes except for lower values for the integer weights at true scores of 2 and between 2.5 and 3. For the 2007 English Language and Composition test form, the CSEMs were nearly identical, except that under the integer weighting scheme the CSEMs were higher at true scores around 4 and lower at true scores close to 5. Though the integer weighting scheme gave more weight to the multiple-choice section for this test form, there was still a portion of the score scale that had higher CSEMs. Model misfit is one potential reason for this finding. CSEM patterns when the multidimensional IRT method was used were similar to those for the unidimensional IRT method for both forms of the AP English Language and Composition test. The AP Biology test showed different patterns. The two weighting schemes

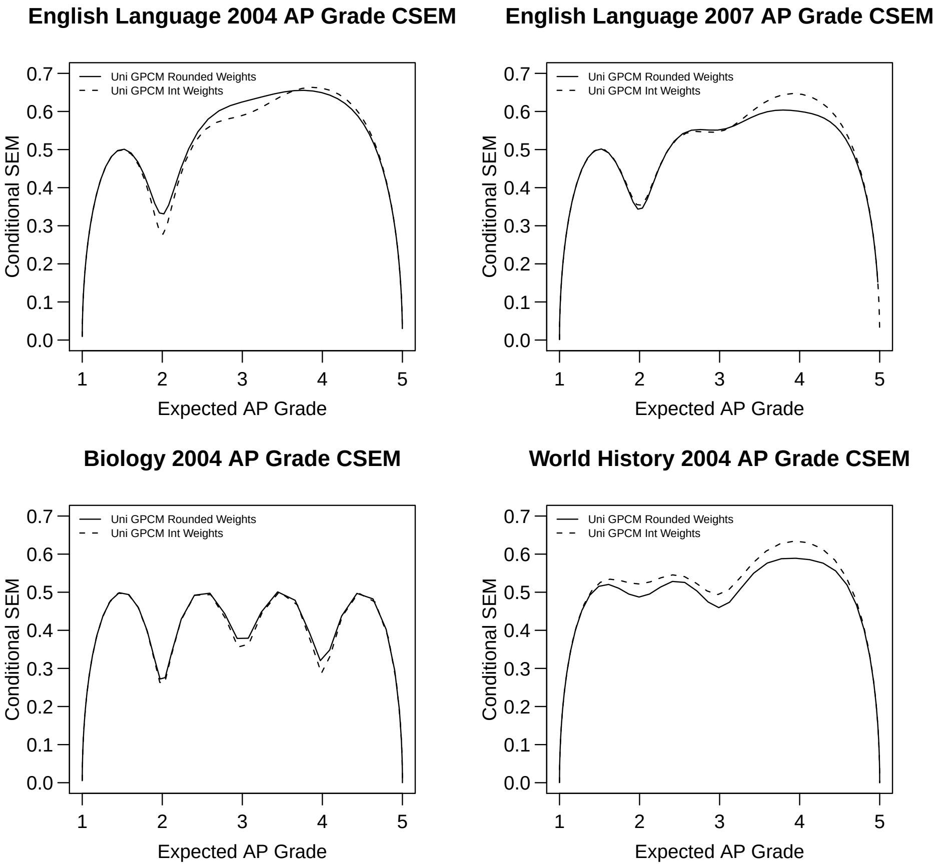


Figure 6: AP grade CSEMs under different weighting schemes for unidimensional IRT method.

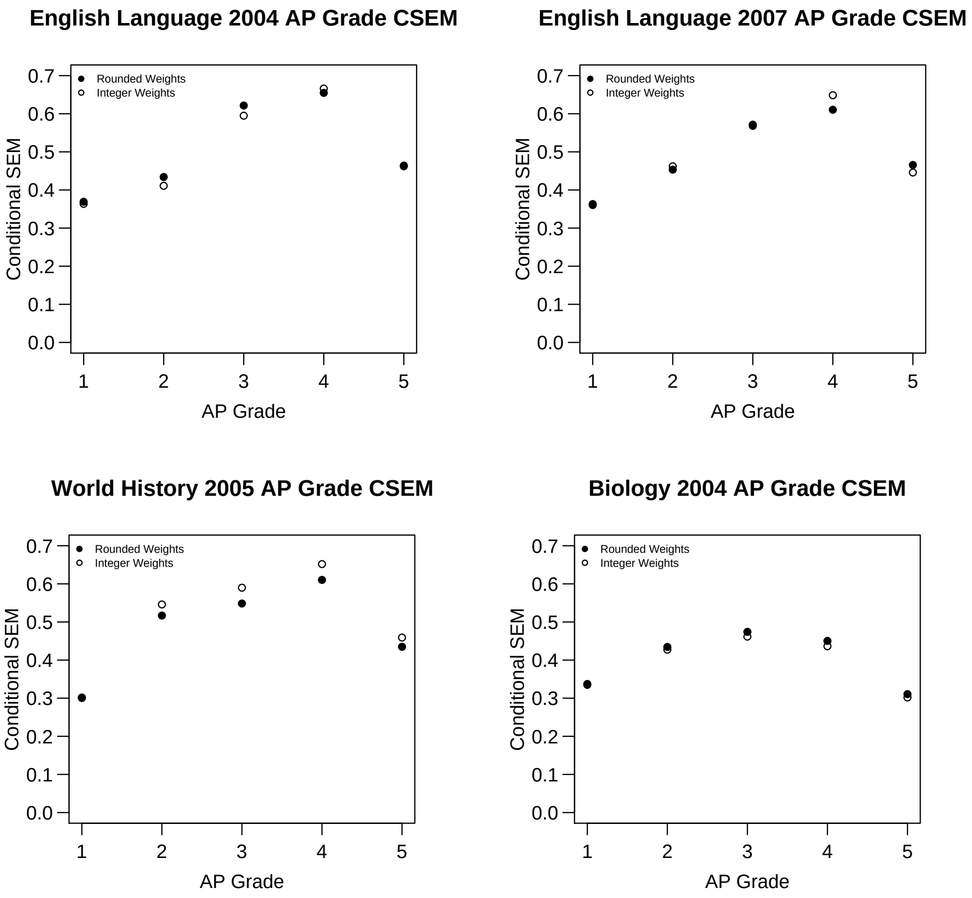
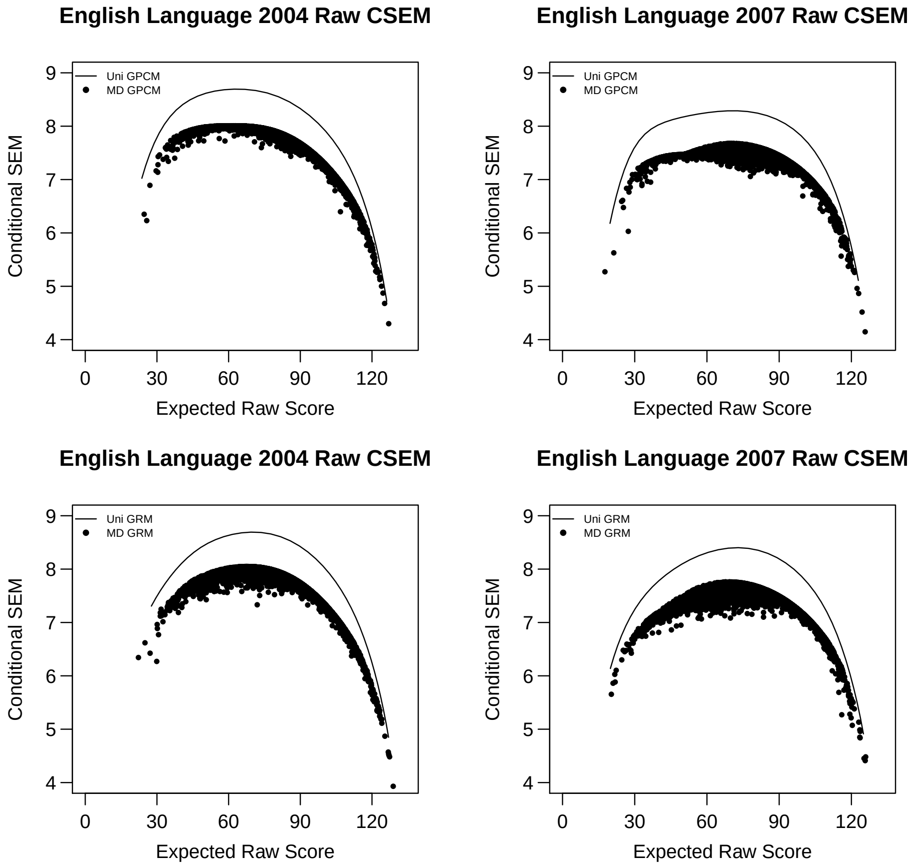


Figure 7: AP grade CSEMs under different weighting schemes for compound multinomial method.

showed nearly identical CSEM patterns except at true scores of 2, 3 and 4. At these values, the integer weight CSEMs were smaller than the rounded weight CSEMs. The AP World History test also had a different pattern. The CSEM patterns were very similar for both weighting schemes except the CSEMs for the integer weights were shifted slightly higher at the middle of the true score range compared to those for the rounded weights.

Figure 7 shows two examples of the effects of the different weighting schemes for CSEMs using the compound multinomial method. The two plots shown are similar to the CSEMs for the other forms of the AP World History and AP Biology tests. For the AP World History test, the middle AP Grade CSEMs were shifted up under the integer weights compared to the rounded weights. For the AP Biology test, the estimates were similar but a little lower for each AP


English Language 2004 Raw CSEM

English Language 2007 Raw CSEM

English Language 2004 Raw CSEM

English Language 2007 Raw CSEM

Figure 8: Unidimensional and multidimensional IRT method raw score CSEMs for rounded weights.

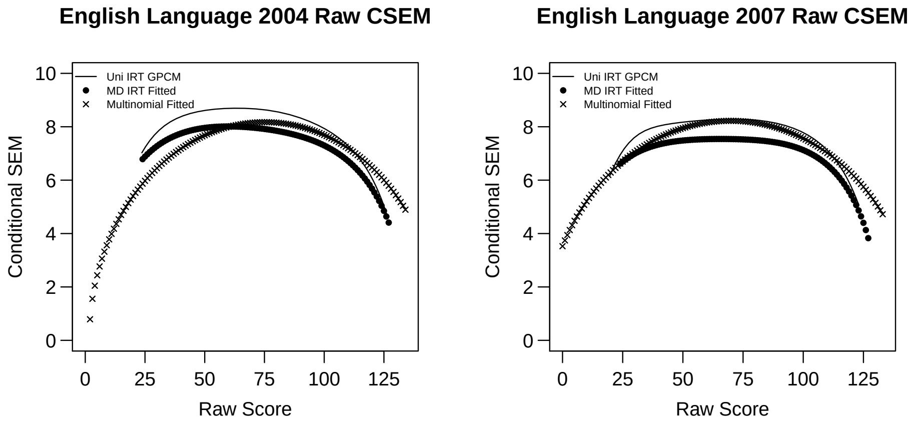


Figure 9: Raw score CSEMs for the English Language and Composition test for rounded weights.

Grade. For the AP English Language and Composition forms, the patterns were similar to the unidimensional IRT method. For instance, for the 2007 English Language and Composition form, the CSEM at 4 is higher for the integer weights and a little lower at an AP Grade of 5.

The plots of raw CSEMS for both IRT methods are shown in Figure 8 for rounded weights. Shapes were similar for both polytomous IRT models. The multidimensional IRT method CSEM estimates were lower than the unidimensional IRT CSEM estimates throughout the entire score range. The biggest differences in magnitude occurred in the middle of the score range. In Figure 9, the raw score CSEMs are compared for the three methodologies using data from the two forms of the AP English Language and Composition test. The multidimensional IRT CSEMs were fit using a fourth degree polynomial. The compound multinomial CSEMs were fit with a second degree polynomial. The AP English Language and Composition test was the only test that had raw score CSEM estimates that were smaller for the compound multinomial method than for the unidimensional IRT method.

For both IRT models, the AP grade CSEM estimates were quite similar. The unidimensional IRT method estimates are slightly larger at true AP Grades of 2.5 to 4.5 but were otherwise quite similar. Results are plotted in Figure 10.

Reliabilities for raw scores and AP grades are shown in Tables 1 and 2 respectively. For the IRT methods, both the graded response and generalized partial credit models produced very similar reliabilities. The reliabilities using the compound multinomial method were lower than the unidimensional IRT reliabilities for the AP World History and AP Biology tests. For the English Language and Composition test, the multidimensional IRT method yielded the highest reliability estimates. The compound multinomial method had reliability estimates that were slightly lower than the multidimensional IRT method. The

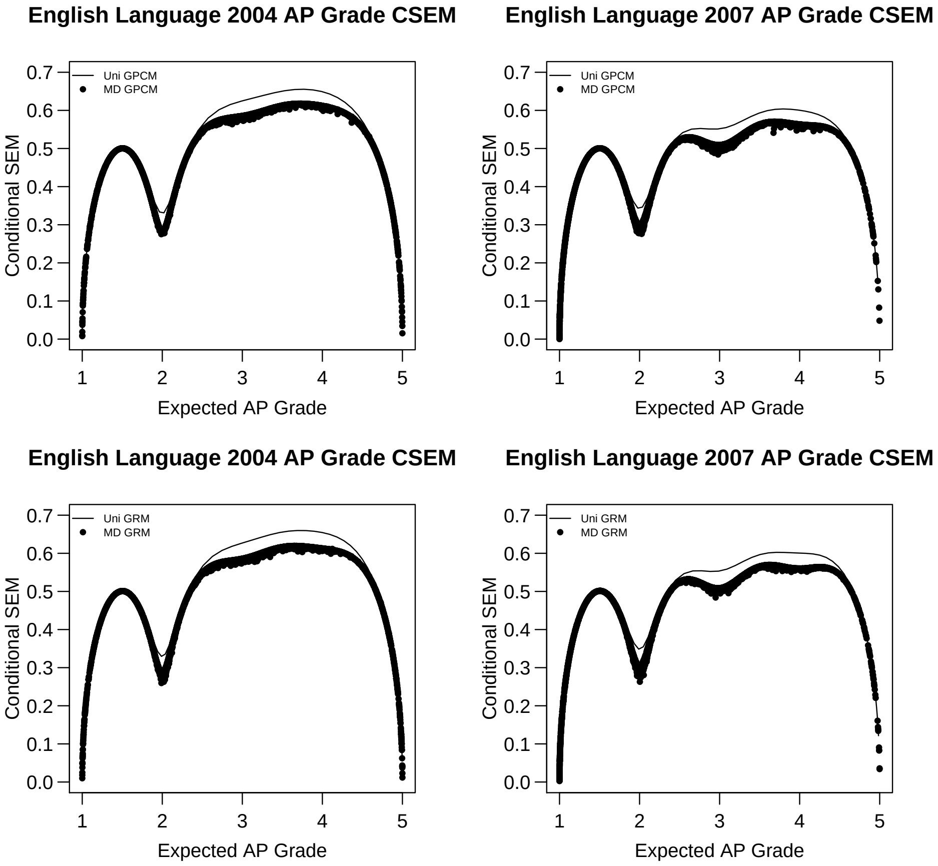


Figure 10: AP grade CSEMs for both IRT methods.

Table 1: Raw Score Reliabilities

World History

	2004		2005	
	Rounded	Integer	Rounded	Integer
Unidimensional IRT				
GPCM	.9045	.8894	.9135	.8971
GRM	.8989	.8820	.9103	.8931
Compound Multinomial	.8671	.8430	.9085	.8932

Biology

	2004		2006	
	Rounded	Integer	Rounded	Integer
Unidimensional IRT				
GPCM	.9532	.9573	.9617	.9645
GRM	.9502	.9548	.9591	.9624
Compound Multinomial	.9468	.9513	.9528	.9565

English Language and Composition

	2004		2007	
	Rounded	Integer	Rounded	Integer
Unidimensional IRT				
GPCM	.8230	.8434	.8348	.8519
GRM	.8191	.8397	.8325	.8499
Compound Multinomial	.8506	.8649	.8516	.8645
Multidimensional IRT				
GPCM	.8529	.8670	.8697	.8811
GRM	.8566	.8703	.8677	.8792

Table 2: AP Grade Reliabilities

World History

	2004		2005	
	Rounded	Integer	Rounded	Integer
Unidimensional IRT				
GPCM	.8618	.8463	.8698	.8519
GRM	.8549	.8374	.8644	.8449
Compound Multinomial	.8122	.7849	.8553	.8357

Biology

	2004		2006	
	Rounded	Integer	Rounded	Integer
Unidimensional IRT				
GPCM	.9133	.9178	.9244	.9273
GRM	.9097	.9147	.9203	.9237
Compound Multinomial	.9024	.9078	.9099	.9136

English Language and Composition

	2004		2007	
	Rounded	Integer	Rounded	Integer
Unidimensional IRT				
GPCM	.7405	.7558	.7466	.7784
GRM	.7382	.7535	.7449	.7763
Compound Multinomial	.7557	.7653	.7462	.7803
Multidimensional IRT				
GPCM	.7845	.7948	.7951	.8193
GRM	.7885	.7985	.7910	.8141

Table 3: Correlations Between Scores on the Multiple-Choice and Free-Response Sections

World History	2004	2005
Pearson Correlation	.7718	.7618
Disattenuated Correlation	.9449	.9084

Biology	2004	2006
Pearson Correlation	.8871	.9096
Disattenuated Correlation	.9818	.9613

English Language	2004	2007
Pearson Correlation	.6319	.6570
Disattenuated Correlation	.8029	.8147
Estimated True θ Correlation		
GPCM	.8007	.8073
GRM	.7983	.8049

unidimensional IRT method had the smallest estimates of reliability.

The correlations between raw scores on the multiple-choice and free-response sections are reported in Table 3. Both the Pearson correlations and disattenuated correlations are reported for both forms of all three tests. Disattenuated correlations were calculated using Cronbach's (1951) coefficient α. The disattenuated correlations shown in Table 3 give some indication that the AP World History and AP Biology tests are fairly close to unidimensional. For the English Language and Composition test forms, the estimated true correlation between the θ for the multiple-choice section and the θ for the free-response section when both polytomous IRT models were used are also reported. The estimated true correlations were very similar to the disattentuated correlations for the AP English Language and Composition test forms.

4 Discussion

For the AP World History and AP Biology tests, the CSEM estimates for the unidimensional IRT method were lower and reliability estimates were higher than those for the compound multinomial method for raw scores. This is consistent with the assumptions of the two methodologies. That is, the unidimensional IRT procedure assumes strictly parallel forms whereas the compound multinomial method assumes test forms are randomly parallel. Forms that are randomly parallel have slightly more error due to variation in content sampling. For AP grades, reliability estimates were also higher for the unidimensional IRT method for these tests. The fact that the AP grades had so few score categories made it difficult to compare the CSEMs for the IRT method with the CSEMs for the compound multinomial method however.

Multidimensionality appears to play a role in CSEM and reliability estimates for the AP English Language and Composition test. The unidimensional IRT method produced larger CSEMs, on average, than the compound multinomial method. Also, reliability estimates were lower for the unidimensional IRT method. One potential reason for this finding is that the multidimensionality is causing more overall error than would be the case if the data were nearly unidimensional. The multidimensional IRT method produced reliability estimates that were higher than both the compound multinomial and unidimensional IRT method. Further research should be done to see if this is a consistent finding and to see how the compound multinomial method may be affected by differing correlations between abilities for each section. Also, based on the observed and fitted distributions for the AP English Language and Composition test, it was difficult to tell whether the unidimensional or multidimensional model fit better, even though evidence such as the estimated true correlations between the θ for each section indicated that the test is not unidimensional. Additional work is needed to explain which method actually fit better and how that fit affects estimates of CSEMs.

The weighting scheme also appeared to be a factor in determining the patterns of AP grade CSEMs. Making alterations to the weighting scheme caused slight changes in the CSEM patterns. These differences were not consistent across tests. For the AP English Language and Composition test using both IRT methods, AP grade CSEMs changed only over small intervals but remained close elsewhere. The AP Biology test estimates only changed at true AP grades that were integers when the unidimensional IRT method was used. For the unidimensional IRT method, the AP World History AP grade CSEMs in the middle of the score range differed in magnitude by a fairly constant amount. It is not clear what caused these differences. These changes may or may not be similar to what would happen for other tests.

The three methodologies used in the present paper have several theoretical and practical differences. First of all, the model assumptions are quite different. For the IRT methods, the data must fit the model and IRT assumptions must not be violated. The compound multinomial method assumes the error distribution is compound multinomial. The conditioning variables are also different between the two IRT methods and the compound multinomial method. The CSEMs are conditioned on IRT proficiency parameters for both IRT methods. The CSEMs for the compound multinomial method are conditioned on an individual's $\boldsymbol{\pi_1}$ and $\boldsymbol{\pi_2}$ vectors as described in Section 2.4. In practice, the observed proportion of items an examinee gets each score is substituted for the expected proportion of the universe of items. This could be considered a limitation of the compound multinomial method.

Many practical reasons may make the use of one of the methods preferable to others in a particular context. If IRT is used for other purposes in a testing program, the unidimensional IRT method requires little additional work if the data are truly unidimensional. If there is reason to believe the data are not unidimensional, the multidimensional IRT method may provide better estimates. The multidimensional IRT procedure is more computationally complex, however. If IRT assumptions are likely violated, the multinomial method may serve as a reasonable alternative.

5 References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (1999). *Standards for educational and psychological testing.* Washington, DC: AERA, APA, NCME.

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

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

Kolen, M. J. (2004). *POLYCSEM* (computer program). Iowa City, Iowa: CASMA. (http://www.education.uiowa.edu/casma/computer_programs.htm#other)

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement, 33*(2), 129-140.

Kolen, M. J., & Wang, T. (1998, 2007). *Conditional standard errors of measurement for composite scores using IRT.* Unpublished Manuscript.

Lee, W. (2007). Multinomial and compound multinomial error models for tests with complex item scoring. *Applied Psychological Measurement, 31*(4), 255-274.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*(3), 359-381.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Muraki, E., & Bock, R. D. (1993). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks.* Chicago: Scientific Software International.

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.

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*(1), 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*(2), 141-162.