Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 29

Multivariate Generalizability Analyses of Mixed-Format Advanced Placement Exams

Sonya Powers and Robert L. Brennan[†]

November 2009

[†]Robert L. Brennan is the E. F. Lindquist Chair in Measurement and Testing and Director of the Center for Advanced Studies in Measurement and Assessment (CASMA), 210 Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: robert-brennan@uiowa.edu).

Sonya Powers is a research assistant in the Educational Measurement and Statistics Department, College of Education, University of Iowa (email: sonya-powers@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 combining item types within assessments, it is important to consider the reliability of scores for each item type and the reliability of composite scores. This study investigated the effect of scoring procedures, section weights, and numbers of items per section on reliability, error variance, and conditional standard errors of measurement, using multivariate generalizability theory techniques. Results indicate that the multiple-choice scoring method (number correct vs. formula scoring) may impact multiple-choice section and composite score reliability, and result in different optimal section weights. Although the multiple-choice section contributes between 50 and 60 percent of the composite score points operationally, optimal weighting from a reliability or error variance perspective would increase the multiple-choice contribution to the composite to 80 or 90 percent. Increasing the number of free-response items often improves the reliability of the composite, but administration time limits the number of free-response items that can be administered. The results of this study are necessarily specific to the Advanced Placement (AP) Biology and AP World History Exams considered, but these data serve to illustrate the usefulness of generalizability theory for answering test development questions that arise with mixed-formats exams.

1 Introduction

When combining item types within an assessment, it is important to consider the reliability of scores for each item type and the reliability of composite scores. Often, items of the same type are grouped into sections, for instance, a section comprised of multiple-choice (MC) items and a section comprised of free-response (FR) items. These sections contribute to the composite score in proportions that can be specified a priori by test developers or by psychometricians after the assessment is administered and scored. Changing the numbers of items within sections or choosing different section weights may affect the reliability of composite scores. The type of scoring used may also affect the reliability of test sections and composite scores. Human scoring of FR items impacts the reliability of the FR section and composite scores. Also, different methods of scoring MC items may impact MC and composite score reliability. Two types of MC scoring methods are number correct (NC) scoring and correction-for-guessing scoring, also known as formula scoring (FS). The amount of measurement error at specific score points is also important to monitor when decisions about examinees are based on cut scores.

This study uses a series of multivariate generalizability (G) studies and decision (D) studies (see Brennan, 2001a) to address the following research questions for an AP Biology and an AP World History exam:

1. What effect does the MC scoring method (FS or NC) have on reliability, error variance, and conditional standard errors of measurement (CSEMs)?
2. How do different section weights impact reliability, error variance, and CSEMs?
3. What effect does changing the number of items per section have on reliability, error variance, and CSEMs?
4. Given administration time constraints, what are optimal numbers of items and weights for MC and FR sections?

Answers to these research questions are necessarily test specific but this study serves to illustrate the usefulness of generalizability theory for answering test development questions that arise with mixed-formats exams.

2 Methods

2.1 Description of Instruments

2.1.1 Scores and Scales

An AP Biology exam and a AP World History exam are considered in this study. Both exams have a MC section and a FR section. The MC and FR sections are weighted and summed to form a composite score. The AP Biology exam has 100 MC items with five alternatives, four 10-point FR items, and a

composite score scale ranging from 0 to 150. The AP World History exam has 70 MC items with five alternatives, three 9-point FR items, and a composite score scale ranging from 0 to 120.

MC items were scored operationally using FS. With FS, wrong responses on these two exams received a score of -0.25, correct responses received a score of 1, and omitted responses received a score of 0. Clearly, using FS may result in MC total scores less than 0. In practice, few examinees receive negative MC total scores and operationally, negative MC total scores are set to 0. For the generalizability analyses discussed later in this paper, negative MC total scores were not set to zero for two reasons. First, the resulting rescaling of item scores for those examinees with negative MC total scores appears arbitrary. Second, as noted above, there were relatively few negative MC total scores.

For comparison purposes, the data were rescored using NC scoring procedures in which all incorrect and omitted items received a score of 0 and all correct items received a score of 1. Because examinees were informed that their exams would be scored using a correction for guessing, the results for NC scoring reported later are at best an approximation and should be interpreted with caution.

Means and standard deviations of unweighted MC and FR sections are provided in Table 1 for the two exams and the two scoring methods. FR means and standard deviations are the same for FS and NC conditions because the scoring method only affects the MC section. Using NC scoring rather than FS results in higher means but less variability for MC scores.

2.1.2 Operational, Relative, and Nominal Weights

For both exams, test developers wanted a composite that had a particular proportion u_m of the composite score range associated with the MC section, and a proportion $u_f = 1 - u_m$ of the composite score range associated with the FR section. We will call u_m and u_f *relative* weights. (Note that this terminology is not entirely consistent with terminology in other literature.) The relative weights are defined a priori by test developers. Operationally, for AP Biology $u_m = 0.6$ and $u_f = 0.4$; for AP World History $u_m = 0.5$ and $u_f = 0.5$.

For both exams, in order to obtain composite scores, the relative weights, prespecified composite score range, and other characteristics of the exam are used to obtain *nominal* weights such that

$$w_m X_m + w_f X_f = C, \tag{1}$$

where w_m is the nominal weight for the MC section, w_f is the nominal weight for the FR section, X_m is the MC total score, X_f is the FR total score, and C is the composite score. Brennan (2009) provides a detailed discussion of how to obtain these nominal weights. Operationally AP Biology nominal weights are $w_m = 0.9$ and $w_f = 1.5$. For AP World History they are $w_m = 0.8571$ and $w_f = 2.2222$.

It is important to note that X_m and X_f are observed scores in the total-score metric (TSM). In generalizability theory, the usual convention is to do computa-

tions in the mean-score metric (MSM). This convention is used in the computer program mGENOVA (Brennan, 2001b), which was used to calculate the results in this study. Therefore, to obtain composite results using mGENOVA, it is necessary to convert terms on the left side of Equation 1 to their analogues in the MSM. Doing so gives

$$v_m \bar{X}_m + v_f \bar{X}_f = C, \tag{2}$$

where v_m is the nominal weight for the MC section based on the MSM, v_f is the nominal weight for the FR section on the MSM, $\bar{X}_m$ is the MC mean score, and $\bar{X}_f$ is the FR mean score. Using the methods discussed by Brennan (2009), it can be shown that for AP Biology, operational nominal weights on the MSM are $v_m = 90$ and $v_f = 6$. For AP World History they are $v_m = 60$ and $v_f = 6.6667$.

2.1.3 Alternative Relative and Nominal Weights

Two additional sets of weights are considered for the AP Biology and AP World History exams. For AP Biology, alternative relative weights of $u_m = 0.5$ and $u_f = 0.5$ and $u_m = 0.4$ and $u_f = 0.6$ are considered. The TSM nominal weights for the first set of alternative weights are 0.75 and 1.875 for the MC and FR sections, respectively. For the MSM, the weights are 75 and 7.5, respectively. For the second set of alternative relative weights, the TSM nominal weights are 0.6 and 2.25 and for the MSM the weights are 60 and 9 for the MC and FR sections, respectively.

For AP World History, alternative relative weights of $u_m = 0.6$ and $u_f = 0.4$ and $u_m = 0.4$ and $u_f = 0.6$ were chosen for comparison. The first set of alternative relative weights have TSM-nominal weights of 1.0286 and 1.7778, respectively, and MSM-nominal weights of 72 and 5.3333, respectively. The second set of alternative relative weights have TSM-nominal weights of 0.6667 and 2.6667, respectively, and MSM-nominal weights of 48 and 8, respectively.

2.1.4 Alternative Numbers of Items per Section

In addition to considering alternative section weights, two alternative numbers of items per section were considered. Operationally, both the AP Biology and the AP World History exams have approximately three-hour administration times. Therefore, in considering how many items to add or remove from a section, the administration time was held as constant as possible. With this constraint, one alternative to the operational AP Biology exam was to remove one FR item, and replace it with 30 MC items. A second alternative for AP Biology was the addition of one FR item and the elimination of 30 MC items. Therefore, three D studies were conducted: one for the operational condition which consisted of 100 MC and 4 FR items, one for 130 MC and 3 FR items, and one for 70 MC and 5 FR items.

For AP World History, the only reasonable alternative was to remove one FR item and replace it with 50 MC items. With only 70 MC items operationally,

adding one FR item would result in a 20-item MC section which would probably not provide adequate content representation. Therefore, for the AP World History exam the operational condition consisted of 70 MC and 3 FR items, and the alternative condition consisted of 120 MC and 2 FR items.

After combining alternative relative section weights with alternative numbers of items per section, there were eight alternative conditions (A1-A8) for the AP Biology exam and five alternative conditions (A1-A5) for the AP World History exam, in addition to the original operational condition (O). Relative weights, nominal TSM and MSM weights, and score ranges for the MC and FR sections are provided in Tables 2 and 3.

2.2 Description of Samples

The G study calculations were based on samples of 20,000 examinees per exam. For both exams, females outnumbered males although to a smaller extent in AP World History. The majority of examinees were white/Caucasian. Complete descriptive statistics are provided in Table 4 including the breakdown in participation rates by ethnicity.

2.3 Procedures

A series of D studies was conducted to answer the four research questions. G study variance and covariance components, D study variance and covariance components, generalizability coefficients, and CSEMs were estimated using mGENOVA software (Brennan, 2001b). The CSEMs reported in this study were fitted absolute standard errors because operationally, decisions about examinees are based on a cut score, not on relative performance. Separate D studies were computed based on operational and alternative weights and items per section, and based on scoring technique (FS vs. NC). The multivariate design used to estimate variance and covariance components and generalizability coefficients was the $p^\bullet \times i^\circ$ design where p represents persons and i represents items. Each person takes each item in both the MC and the FR sections. Therefore persons are fully crossed with the fixed "item-type" facet (indicated by the closed circle). Items are instead nested within item-type-that is, an item is either a MC item or a FR item (indicated by the open circle). Covariance terms can be estimated for the facets crossed with item-type (namely, persons), but covariance terms cannot be estimated for the facets nested within item-type (namely, items). The superscripts follow the notation in Brennan (2001a) for multivariate designs. Persons and items were considered random facets in this study.

3 Results

3.1 What effect does the MC scoring method (FS or NC) have on reliability, error variance, and CSEMs?

3.1.1 G Study Results

MSM variance and covariance component estimates for the FS and NC G studies are provided in Table 5. FS and NC scoring methods apply only to MC items; therefore, although the covariance component between the two sections for persons ($\sigma_{mf}(\mathrm{p})$) is affected by MC scoring method, the FR variance components are unaffected. For the MC section, the variance components for FS were uniformly larger than the variance components for NC scoring. The G study results provided in Table 5 are the basis for all further D study comparisons, reliability estimates, and error variance estimates. It should be noted that G and D study variance components are all reported on the MSM, as is the usual convention in G theory.

3.1.2 D Study Results for MC and FR Sections

D study variance and covariance components and reliability indices for MC and FR sections are provided in Table 6 for the AP Biology exam and in Table 7 for the AP World History exam. Three D studies were conducted for AP Biology operational and alternative conditions, and two D studies were conducted for AP World History operational and alternative conditions. The first two-thirds of Tables 6 and 7 provides a comparison of D study results for the MC section using FS and NC scoring. The final third of the two tables provides the D study results for the FR section. Across all conditions and both exams, the MC variance components were uniformly larger for FS than for NC scoring.

Two types of reliability indices are typically reported for generalizability analyses: a generalizability coefficient, symbolized $E\rho^2$, which involves relative error variance, and an index of dependability, symbolized Φ, which involves absolute error variance. Φ will be smaller than $E\rho^2$ in the $p^\bullet \times i^\circ$ design whenever the D study variance component for items ($\sigma^2(\mathrm{i})$) is greater than zero. When $\sigma^2(\mathrm{i})$ equals zero, the two coefficients are equal. Because $\sigma^2(\mathrm{i})$ was greater than zero for all D studies in this study, Φ was always smaller than $E\rho^2$.

Although universe score variance was smaller for NC scoring than for FS, NC error variance was also smaller, resulting in slightly higher MC reliability indices for NC than for FS. MC reliability indices are around 0.9 or higher for most conditions and both exams. The FR section had substantially lower reliability– as low as 0.481 for AP World History Alternatives 3-5 for Φ. However, FR reliability estimates reported here are more difficult to interpret than might be expected: in the available data raters and tasks are completely confounded, because there is only one rating of the response of each person to each task (or item). The assignment of raters to persons and tasks is such that variability attributable to the two facets (raters and tasks) cannot be disentangled.

3.1.3 D Study Results for Composite Scores

Composite universe score variance and error variance components are weighted averages of the corresponding MC and FR section D study variance components. Table 8 provides the composite score variance components and reliability indices for nine composite score D studies involving the AP Biology exam. Table 9 provides the same information for six D studies involving the AP World History exam. The results in Tables 8 and 9 were obtained using mGENOVA with the MSM weights listed in Tables 2 and 3. $E\rho^2$ in the current design is equal to stratified alpha. As noted previously, Φ is smaller than $E\rho^2$ across all conditions for MC and FR sections; Φ is also smaller for the composite.

Consistent with the findings from the MC section, the variance components for the composite were uniformly larger when FS was used than when NC scoring was used. However, unlike the D study results found for the MC section, the reliability of the composite is somewhat *smaller* for NC scoring than for FS (recall that the reliability was *larger* when the MC section was scored NC). The composite error variance components were smaller for NC than for FS so the apparent discrepancy between the NC MC section and composite score reliabilities is due to the universe score variance component. Universe score variance for the composite is

$$\sigma_C^2(p) = v_m^2\sigma_m^2(p) + v_f^2\sigma_f^2(p) + 2v_m v_f \sigma_{mf}(p), \tag{3}$$

where the variance components and weights are on the MSM.

The universe score variance component for the FR section, σ_f^2(p) , is unaffected by MC scoring method. The previous finding that the reliability of the MC section was greater for NC scoring indicates that the decrease in σ_m^2(p) from FS to NC scoring was not very large. Therefore, it is the lower covariance between the MC and FR sections for NC scoring that largely explains the decrease in composite reliability from FS to NC scoring. For AP Biology, σ_{mf}(p) decreased from 0.34 to 0.29 when comparing FS to NC results (see Table 6). The multiplication of σ_{mf}(p) by 2 in equation 3 further amplifies the decrease in composite universe score variance for the NC scoring method. The lower NC error variances were not small enough to compensate for the substantial decrease in universe score variance when changing from FS to NC scoring. Therefore the composite reliability was lower for NC than for FS.

The comparative results for NC and FS discussed above should not be interpreted as a definitive indication of which type of scoring is better. The differences between NC and FS are much more complicated, and a discussion of all issues relevant to the choice of scoring method is beyond the scope of this study.

Despite somewhat lower NC composite reliabilities, a comparison of NC and FS CSEMs across the composite score scale indicated that absolute error CSEMs for NC were substantially smaller than CSEMs for FS for a large portion of the score scale (see Figure 1). Smaller CSEMs for NC scoring is not surprising; Tables 8 and 9 already indicated that the composite error variance for NC is smaller than that for FS. However, these findings illustrate that consideration of

score precision from a measurement error perspective could lead to a preference for one scoring method (namely NC), but consideration of score precision from a reliability perspective could lead to a preference for a different scoring method (namely FS). Again, these results do not establish the superiority of either scoring method. Such considerations must be based on more than CSEMs and reliability estimates. For example, true scores using NC scoring are not the same as true scores using FS.

3.2 How do different section weights impact reliability, error variance, and CSEMs?

3.2.1 Relative Weights

Three sets of relative weights were considered for both exams: the operational set (O), and two alternative sets (A1 and A2). These relative weights and the corresponding TSM and MSM nominal weights were provided in Tables 2 and 3. The D study variance components and reliability estimates for the MC and FR sections are provided in the third column (O,A1,A2) of Table 6 for AP Biology and the third column of Table 7 for AP World History. Because section weights are only a composite score consideration, all three conditions (O, A1, and A2) share the same MC and FR section D study results.

Composite D study results differ as a function of section weights. The D study results for O, A1, and A2 are provided in columns 3-5 of Table 8 for AP Biology and columns 3-5 of Table 9 for AP World History. Clearly different section weights result in different composite universe score variance, error variances, and reliability indices. For both AP Biology and AP World History, the effect of increasing the relative weight assigned to the MC section was an increase in universe score variance, a decrease in error variance, and an increase in composite score reliability. Conversely, increasing the relative FR weight (i.e., decreasing the relative MC section weight) resulted in lower universe score variance, higher error variances, and lower reliability indices.

In Figure 2, CSEMs for the O, A1, and A2 conditions are provided for AP Biology and AP World History, using NC and FS results. From Figure 2 it can be seen that the greater the relative MC section weight, the lower the CSEMs. For both AP Biology and AP World History the greatest relative MC weight considered was 0.6. For AP Biology, a 0.6 relative weight was operationally assigned to the MC section. Therefore, for AP Biology, the operational condition had the lowest CSEMs for both NC and FS. For AP World History, the operational MC weight was 0.5. The A1 condition, which had a relative MC weight of 0.6, resulted in lower CSEMs than the operational condition across the score scale for both NC and FS.

In Figure 3, the relationship of the relative weight assigned to the MC section and the composite score reliability (Φ) is provided for AP Biology and AP World History, and separately for NC and FS. The highest reliability was achieved with a MC weight between 0.8 and 0.9. It appears that the optimal MC relative weight under NC scoring would be slightly higher than the optimal

MC relative weight using FS. The relationship between relative MC weight and composite reliability was steeper for AP World History, indicating that the same amount of increase in relative weight for the MC section will result in greater increases in composite score reliability for AP World History than for AP Biology. The greater impact of relative MC weights on score reliability for AP World History is due to the lower reliability of scores on the FR section for AP World History. Clearly, all conditions included in this study had less-than-optimal relative weights.

3.2.2 Effective Weights

Effective weights quantify the proportion of composite variance contributed by a given fixed facet. In this study, item type is the fixed facet and effective weights represent the proportion of composite score variance attributable to the MC and FR sections. Effective weights are similar to relative weights in that they are proportions that sum to one across sections. However, the weights we have been calling relative weights are based on the proportion of composite score points assigned to each section. Effective weights are based instead on the proportion of composite score variance attributable to each section–either universe score variance or error variance. Effective weights provide an alternative conceptualization of the importance placed on each test section. Effective weights were calculated based on the results of the operational and alternative relative weighting conditions. See Brennan (2009) for details about computing effective weights.

Table 10 provides the effective weights that resulted from all conditions for AP Biology, including NC and FS. Table 11 provides the same information for AP World History. For both exams, FS resulted in greater effective weights for the MC section. This is not surprising given that effective weights are based on score variability, and the variance components for the MC section were larger under FS than NC scoring. Also, use of NC or FS does not change the variance components of the FR section, so the relative contribution of the MC section to composite variance must be larger for FS.

Effective weights based on universe score variance ($ew(\mathrm{p})$) are not affected by changes in the number of items within sections, but they are affected by changes in relative section weights. This result was also expected because $ew(p)$ is based on the variability of persons, not items. Effective weights based on error variance ($ew(\delta)$ and $ew(\Delta)$) are affected by both person and item variance.

Although the $ew(\mathrm{p})$ tended to be similar to the relative weights, $ew(\delta)$ and $ew(\Delta)$ were very different. In general, the values of $ew(\delta)$ and $ew(\Delta)$ for the MC section were quite small, ranging from 0.54 to 0.07 for AP Biology (FS), 0.45 to 0.05 for AP Biology (NC), 0.46 to 0.04 for AP World History (FS), and 0.36 to 0.03 for AP World History (NC). Clearly the MC section contributed fairly substantially to universe score variance, but in some cases very little to error variance. This is expected given that the FR section was much less reliable than the MC section.

In Figure 4, the relationship of MC effective weights $ew(\mathrm{p})$ and $ew(\Delta)$, and

composite score reliability (Φ) is provided for AP Biology and AP World History, and separately for NC and FS, based on operational numbers of items. The relationship of ew(p) and reliability was very similar to the relationship of MC relative weights and reliability (see Figure 3). Therefore, if weights were chosen based on consideration of ew(p) instead of relative weights, similar nominal weights might be obtained. If optimal ew(p) weights are desired, nominal weights would be computed such that the MC section contributed approximately 80% to composite universe score variance, and the FR section contributed the remaining 20%.

The relationship of $ew(\Delta)$ and reliability was much different from the relationships for relative weights or ew(p). In terms of composite score reliability, little is gained by increasing the MC $ew(\Delta)$ beyond 0.2. It is difficult to discern the maximum value of the curvilinear relationship between the $ew(\Delta)$ for the MC section and composite score reliability, but it might also be around 0.8 or higher.

The relationship between relative MC weights and effective MC weights is provided in Figure 5 for AP Biology and AP World History exams, and separately for NC and FS. The NC relationship is shifted slightly to the left of the FS relationship. Clearly for these two exams, the relationship between relative weights and ew(p) for the MC section was almost linear. However, the relationship between relative weights and $ew(\Delta)$ was monotonic increasing and nonlinear.

3.3 What effect does changing the number of items per section have on reliability, error variance, and CSEMs?

3.3.1 D Study Results for Sections

Holding section weights constant, two alternative numbers of items per section were considered for the AP Biology exam and one alternative was considered for the AP World History exam. In terms of the conditions listed in Tables 2 and 3, the operational (O) condition was compared to A3 and A6 for AP Biology, and the operational condition was compared to A3 for AP World History. D study results are provided in Table 6 for AP Biology and Table 7 for AP World History. Each condition compared in this section has different D study results for the MC and FR sections.

For AP Biology, decreasing the FR section by one item and increasing the MC section by 30 items caused an increase in the MC section score reliability indices and a decrease in the FR section score reliability indices (see O and A3 in Table 6). For AP World History, decreasing the FR section by one item and increasing the MC section by 50 items also resulted in increased MC score reliability indices and decreased FR score reliability indices (see O and A3 in Table 7). The A6 condition for AP Biology involved increasing the FR section by one item and decreasing the MC section by 30 items. As expected, MC section score reliability decreased and FR section score reliability increased for A6 compared to the operational section reliabilities.

3.3.2 D Study Results for Composite

Composite variance components and reliability coefficients for conditions O, A3, and A6 are provided in Table 8 for AP Biology, separately for NC and FS. The same information is provided for AP World History in Table 9 for conditions O and A3. For AP Biology, the A3 condition resulted in lower composite score reliability compared to the O condition. Therefore, removing a FR item and adding 30 MC items did not improve operational composite score reliability. For AP Biology FS, condition A6, which added a FR item and removed 30 MC items, also resulted in lower composite score reliability than the operational condition. For NC, the reliabilities of O and A6 were almost equivalent.

For AP World History, condition A3, which removed a FR item and added 50 MC items, resulted in lower reliability than the operational condition. Therefore, in terms of items per section, it appears that the operational condition was superior to other possible conditions for both AP Biology and AP World History, given administration time constraints.

Inspection of CSEMs for the O, A3, and A6 conditions, provided in Figure 6, also indicated that the operational condition was superior to the other conditions across both exams and both MC scoring methods. The differences between the CSEMs for O, A3, and A6 were fairly small for AP Biology. The difference in the CSEMs for O and A3 was much larger for AP World History.

3.4 Given administration time constraints, what are optimal numbers of items and weights for MC and FR sections?

Reliability and error variances across the nine AP Biology conditions are provided in Table 8. Figure 7 contains plots of CSEMs for all AP Biology study conditions. The first column of plots contains CSEMs for the operational condition and conditions A1-A4 for both NC and FS. The second column of plots contains CSEMs for the operational condition and conditions A5-A8 for both NC and FS. Based on Table 8, Figure 7, and results presented previously, it appears that greater relative weights for the MC section and operational numbers of items within the MC and FR sections resulted in the greatest composite score reliability indices, the smallest error variances, and the smallest CSEMs. Therefore, for the AP Biology exam, the operational condition could only be improved upon by increasing the relative MC section weight to about 0.8, given administration time constraints. Among the nine AP Biology conditions in this study, the operational condition was optimal.

For the AP World History exam, operational numbers of MC and FR items were optimal but A1, with a MC relative weight of 0.6 instead of 0.5, resulted in more reliable composite scores. Reliability and error variances across the six AP World History conditions are provided in Table 9. AP World History CSEMs for the operational condition and conditions A1-A5 are plotted for NC and FS in Figure 8. The impact of relative section weights on CSEMs and reliability was greater than the impact of changing the numbers of MC and FR items.

Therefore, the CSEMs for A4 were also lower than the operational CSEMs, and the reliability for A4 was greater than the operational reliability. Among the six AP World History conditions in this study, A1 was optimal. Increasing the relative MC weight from 0.6 to approximately 0.8 would have provided even higher composite score reliability.

4 Conclusions

MC section score reliability was around 0.9 for both exams, but FR section score reliability was much lower, especially for AP World History, which had a Φ of 0.58 operationally. Composite score reliabilities were smaller than the score reliabilities of the MC sections because of the relatively large weight given to the FR section.

One way to increase composite score reliability and decrease composite CSEMs would be to give the MC section higher relative weight in forming composite scores. For both exams, relative weights for the MC section must be approximately 0.8 to maximize reliability for FS. When using NC scoring the MC relative weights must be even higher to maximize reliability. Increasing the relative weights for the AP World History exam would be especially important to consider for increasing composite score reliability because of the low reliability of FR section scores.

Consideration of the relationship between composite score reliability and effective weights also leads to the conclusion that high MC section weights are necessary to maximize reliability. For $ew(\mathrm{p})$, a MC section effective weight of approximately 0.8 is associated with maximum reliability. For $ew(\Delta)$ higher MC section weights also correspond to higher composite score reliability estimates for most of the score range.

Changing the number of items within sections can lead to increased composite score reliability in some cases. In the current study, changes to the number of items within the FR and MC sections resulted in changes to the section score reliabilities, but the operational numbers of items resulted in the highest composite score reliability for both AP Biology and AP World History. Given administration time constraints, the results reported here support continued use of the operational numbers of items.

Finally, the relationship between reliability and error variance for NC and FS is somewhat complicated. For the MC section, NC scoring results in smaller error variances and higher reliabilities. For the composite, NC scoring results in smaller error variances and smaller CSEMs for the majority of the AP Biology and AP World History score ranges. However, the covariance component for persons between the fixed MC and FR facets is much smaller when using NC scoring than when using FS. The smaller NC covariance and variance components resulted in a substantially smaller composite universe score variance for the NC scoring method compared to the composite universe score variance for the FS method. Because the impact of NC scoring is greater for the composite universe score variance than the composite error variances, composite reliability

for the NC scoring method is lower than for the FS method. In short, if one used reliability as a criterion, FS might be preferred. If one used CSEMs as a criterion, NC scoring might be preferred. But these statements should not be interpreted as indicators of which type of scoring is better, per se. That is a much more complicated issue that is outside the scope of this paper. That being said, if a statistic is used as a partial basis for choosing between NC and FS, then it matters whether the statistic chosen is a reliability coefficient or an error variance.

One limitation in this study is the lack of necessary data to estimate variance components for raters. Because each item was rated only once operationally, a special G study would have to be conducted in order to estimate the variance components associated with raters. Because this study did not include a rater facet, FR score reliability estimates may not be accurate. If FR score reliability indices change when a rater facet is included, the optimal relative or effective weights for the MC section will also change.

Another possible limitation is the adequacy of the NC scoring comparison. Because examinees were encouraged to skip items they did not know the answers to (because of the correction-for-guessing), it is questionable whether rescoring responses NC (0 for all omitted and incorrect responses) is valid. It is likely that response patterns and section scores would have been different if students were informed that there was no penalty for guessing. However, the choice of scoring method is an important testing concern, and it seems unlikely that the rescoring process used in this study will lead to substantially inaccurate results.

A special G study could be designed to estimate the impact of a rater facet on estimates of FR and composite score reliability. Additionally, it would be possible to give two randomly equivalent examinee groups different scoring instructions so that one group would respond for an exam using FS and the other group would respond for a NC scored exam. If the results are comparable to those found in this study, the differences in reliability, error variance, and CSEMs between scoring methods would be more fully established.

Also, analyses used in this study could also be applied to other exams for test development or redesign purposes. Additional conditions could be considered using more extreme relative weights, such as 0.8 for the MC section and 0.2 for the FR section. However, it might be unlikely that test developers would adopt relative FR weights less than 0.4 despite increases in composite score reliability. Conditions could also be based on considerations about effective weights rather than relative weights.

Generalizability theory can be used to estimate the impact of multiple sources of error on score reliability. The sample of research questions considered in this study shows the potential usefulness of generalizability theory in studying important test development and scoring issues commonly encountered by measurement specialists. Furthermore, mixed-format exams are used widely throughout the United States for making important decisions that affect millions of examinees annually. The validity of these decisions rests in part on the reliability (or more broadly, the generalizability) of the examination scores, which was the focus of this research.

5 References

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

Brennan, R. L. (2001b). *mGENOVA* [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. (2009). Notes about nominal weights in multivariate generalizability theory. *CASMA Technical Note*, No. 4 Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, The University of Iowa. (Available on http://www.education.uiowa.edu/casma).

Table 1: Descriptive Statistics for Exams

		AP Biology		AP World History	
		MC	FR	MC	FR
FS	Mean	45.33	17.13	34.47	8.81
	SD	22.81	8.13	13.36	5.00
NC	Mean	53.63	17.13	40.26	8.81
	SD	17.33	8.13	11.40	5.00

Table 2: AP Biology Scales and Weights Across Study Conditions

	Number of Items	Relative Weights	TSM Weights	MSM Weights	Score Range
			MC		
O	100	0.6	0.9	90	0-90
A1	100	0.5	0.75	75	0-75
A2	100	0.4	0.6	60	0-60
A3	130	0.6	0.6923	90	0-90
A4	130	0.5	0.5769	75	0-75
A5	130	0.4	0.4615	60	0-60
A6	70	0.6	1.2857	90	0-90
A7	70	0.5	1.0714	75	0-75
A8	70	0.4	0.8571	60	0-60
			FR		
O	4	0.4	1.5	6	0-60
A1	4	0.5	1.875	7.5	0-75
A2	4	0.6	2.25	9	0-90
A3	3	0.4	2	6	0-60
A4	3	0.5	2.5	7.5	0-75
A5	3	0.6	3	9	0-90
A6	5	0.4	1.2	6	0-60
A7	5	0.5	1.5	7.5	0-75
A8	5	0.6	1.8	9	0-90

Note. There is 1 score point per MC item and 10 score points per FR item. The composite scale ranges from 0-150.

Table 3: AP World History Scales and Weights Across Study Conditions

	Number of Items	Relative Weights	TSM Weights	MSM Weights	Score Range
			MC		
O	70	0.5	0.8571	60	0-60
A1	70	0.6	1.0286	72	0-72
A2	70	0.4	0.6667	48	0-48
A3	120	0.5	0.5	60	0-60
A4	120	0.6	0.6	72	0-72
A5	120	0.4	0.4	48	0-48
			FR		
O	3	0.5	2.2222	6.6667	0-60
A1	3	0.4	1.7778	5.3333	0-48
A2	3	0.6	2.6667	8	0-72
A3	2	0.5	3.3333	6.6667	0-60
A4	2	0.4	2.6667	5.3333	0-48
A5	2	0.6	4	8	0-72

Note. There is 1 score point per MC item and 9 score points per AP World History FR item. The composite scale ranges from 0-120.

Table 4: Sample Demographics

		AP Biology (%)	AP World History (%)
Gender	Females	58.5	54.9
	Males	41.5	45.1
Ethnicity*	Asian	17.2	14.4
	Black	4.9	6.3
	Hispanic	7.1	11.6
	White	63.9	60.6
	Other	4.3	4.5

*Percentages for ethnicity will not sum to 100% because some examinees did not provide their ethnicity.

Table 5: G Study Variance and Covariance Component Estimates

	FS		NC	
	AP Biology			
σ^2(p)	0.03855	3.17587	0.02812	3.17587
σ_{mf}(p)	0.34158		0.28644	
σ^2(i)	0.04331	0.50529	0.03160	0.50529
σ^2(pi)	0.27030	3.81574	0.18928	3.81574
	AP World History			
σ^2(p)	0.03245	1.90439	0.02377	1.90439
σ_{mf}(p)	0.23144		0.19357	
σ^2(i)	0.04164	1.48838	0.02896	1.48838
σ^2(pi)	0.27984	2.61707	0.19205	2.61707

Table 6: AP Biology D Study Variance and Covariance Components and Reliability Coefficients for Sections

		O, A1, A2	A3-A5	A6-A8
MC (FS)	σ_m^2(p)	0.03855	0.03855	0.03855
	σ_{mf}(p)	0.34158	0.34158	0.34158
	σ_m^2(I)	0.00043	0.00033	0.00062
	σ_m^2(pI)= $\sigma_m^2(\delta)$	0.00270	0.00208	0.00386
	$\sigma_m^2(\Delta)$	0.00314	0.00241	0.00448
	$E\rho_m^2$	0.934	0.949	0.909
	Φ_m	0.925	0.941	0.896
MC (NC)	σ_m^2(p)	0.02812	0.02812	0.02812
	σ_{mf}(p)	0.28644	0.28644	0.28644
	σ_m^2(I)	0.00032	0.00024	0.00045
	σ_m^2(pI)= $\sigma_m^2(\delta)$	0.00189	0.00146	0.00270
	$\sigma_m^2(\Delta)$	0.00221	0.00170	0.00316
	$E\rho_m^2$	0.937	0.951	0.912
	Φ_m	0.927	0.943	0.899
FR	σ_f^2(p)	3.17587	3.17587	3.17587
	σ_f^2(I)	0.12632	0.16843	0.10106
	σ_f^2(pI)= $\sigma_f^2(\delta)$	0.95394	1.27191	0.76315
	$\sigma_f^2(\Delta)$	1.08026	1.44034	0.86420
	$E\rho_f^2$	0.769	0.714	0.806
	Φ_f	0.746	0.688	0.786

Table 7: AP World History D Study Variance and Covariance Components and Reliability Coefficients for Sections

		O, A1, A2	A3-A5
MC (FS)	σ^2_m(p)	0.03245	0.03245
	σ_{mf}(p)	0.23144	0.23144
	σ^2_m(I)	0.00059	0.00035
	σ^2_m(pI)= $\sigma^2_m(\delta)$	0.00400	0.00233
	$\sigma^2_m(\Delta)$	0.00459	0.00268
	$E\rho^2_m$	0.890	0.933
	Φ_m	0.876	0.924
MC (NC)	σ^2_m(p)	0.02377	0.02377
	σ_{mf}(p)	0.19357	0.19357
	σ^2_m(I)	0.00041	0.00024
	σ^2_m(pI)= $\sigma^2_m(\delta)$	0.00274	0.00160
	$\sigma^2_m(\Delta)$	0.00316	0.00184
	$E\rho^2_m$	0.897	0.937
	Φ_m	0.883	0.928
FR	σ^2_f(p)	1.90439	1.90439
	σ^2_f(I)	0.49613	0.74419
	σ^2_f(pI)= $\sigma^2_f(\delta)$	0.87236	1.30854
	$\sigma^2_f(\Delta)$	1.36848	2.05273
	$E\rho^2_f$	0.686	0.593
	Φ_f	0.582	0.481

Table 8: AP Biology Reliability Composite Error Variance Estimates and Coefficients

	O	A1	A2	A3	A4	A5	A6	A7	A8
					FS				
σ^2_C(p)	795.49	779.76	764.93	795.49	779.76	764.93	795.49	779.76	764.93
$\sigma^2_C(\delta)$	56.24	68.86	87.00	62.63	83.24	110.51	58.75	64.65	75.72
$\sigma^2_C(\Delta)$	64.29	78.41	98.79	71.39	94.59	125.35	67.40	73.81	86.13
$E\rho^2$	0.934	0.919	0.898	0.927	0.904	0.874	0.931	0.923	0.910
Φ	0.925	0.909	0.886	0.918	0.892	0.859	0.922	0.914	0.899
					NC				
σ^2_C(p)	651.46	659.06	667.83	651.46	659.06	667.83	651.46	659.06	667.83
$\sigma^2_C(\delta)$	49.67	64.31	84.08	57.58	79.74	108.27	49.38	58.14	71.55
$\sigma^2_C(\Delta)$	56.78	73.19	95.45	65.61	90.58	122.78	56.67	66.36	81.36
$E\rho^2$	0.929	0.911	0.888	0.919	0.892	0.860	0.930	0.919	0.903
Φ	0.920	0.900	0.875	0.908	0.879	0.845	0.920	0.909	0.891

Table 9: AP World History Reliability Composite Error Variance Estimates and Coefficients

		O	A1	A2	A3	A4	A5
FS	σ^2_C(p)	386.61	400.13	374.39	386.61	400.13	374.39
	$\sigma^2_C(\delta)$	53.16	45.54	65.04	66.55	49.31	89.12
	$\sigma^2_C(\Delta)$	77.36	62.73	98.16	100.88	72.28	137.55
	$E\rho^2$	0.879	0.898	0.852	0.853	0.890	0.808
	Φ	0.833	0.864	0.792	0.793	0.847	0.731
NC	σ^2_C(p)	325.07	326.05	325.31	325.07	326.05	325.31
	$\sigma^2_C(\delta)$	48.65	39.04	62.15	63.92	45.52	87.43
	$\sigma^2_C(\Delta)$	72.19	55.29	94.86	97.86	67.94	135.62
	$E\rho^2$	0.870	0.893	0.840	0.836	0.878	0.788
	Φ	0.818	0.855	0.774	0.769	0.828	0.706

Table 10: AP Biology Effective Weights

		FS		NC	
		MC	FR	MC	FR
ew(p)	O	0.62	0.38	0.59	0.41
	A1	0.52	0.48	0.48	0.52
	A2	0.42	0.58	0.38	0.62
	A3	0.62	0.38	0.59	0.41
	A4	0.52	0.48	0.48	0.52
	A5	0.42	0.58	0.38	0.62
	A6	0.62	0.38	0.59	0.41
	A7	0.52	0.48	0.48	0.52
	A8	0.42	0.58	0.38	0.62
$ew(\delta)$	O	0.39	0.61	0.31	0.69
	A1	0.22	0.78	0.17	0.83
	A2	0.11	0.89	0.08	0.92
	A3	0.27	0.73	0.20	0.80
	A4	0.14	0.86	0.10	0.90
	A5	0.07	0.93	0.05	0.95
	A6	0.53	0.47	0.44	0.56
	A7	0.34	0.66	0.26	0.74
	A8	0.18	0.82	0.14	0.86
$ew(\Delta)$	O	0.40	0.60	0.32	0.68
	A1	0.22	0.78	0.17	0.83
	A2	0.11	0.89	0.08	0.92
	A3	0.27	0.73	0.21	0.79
	A4	0.14	0.86	0.11	0.89
	A5	0.07	0.93	0.05	0.95
	A6	0.54	0.46	0.45	0.55
	A7	0.34	0.66	0.27	0.73
	A8	0.19	0.81	0.14	0.86

Table 11: AP World History Effective Weights

		FS		NC	
		MC	FR	MC	FR
$ew(\text{p})$	O	0.54	0.46	0.50	0.50
	A1	0.64	0.36	0.61	0.39
	A2	0.44	0.56	0.40	0.60
	A3	0.54	0.46	0.50	0.50
	A4	0.64	0.36	0.61	0.39
	A5	0.44	0.56	0.40	0.60
$ew(\delta)$	O	0.27	0.73	0.20	0.80
	A1	0.46	0.54	0.36	0.64
	A2	0.14	0.86	0.10	0.90
	A3	0.13	0.87	0.09	0.91
	A4	0.25	0.75	0.18	0.82
	A5	0.06	0.94	0.04	0.96
$ew(\Delta)$	O	0.21	0.79	0.16	0.84
	A1	0.38	0.62	0.30	0.70
	A2	0.11	0.89	0.08	0.92
	A3	0.10	0.90	0.07	0.93
	A4	0.19	0.81	0.14	0.86
	A5	0.04	0.96	0.03	0.97

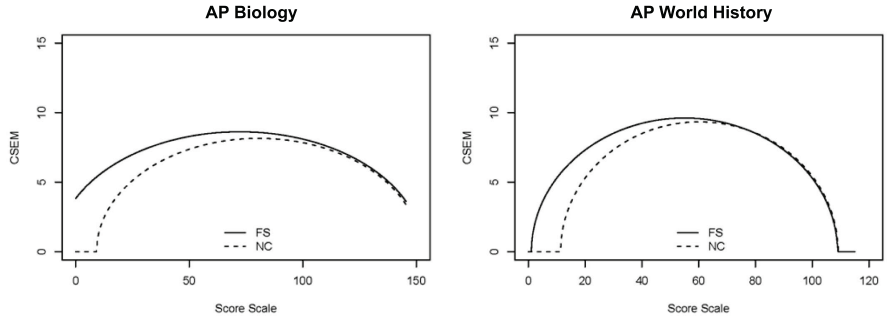

Figure 1: Comparison of Conditional Standard Errors of Measurement for NC and FS

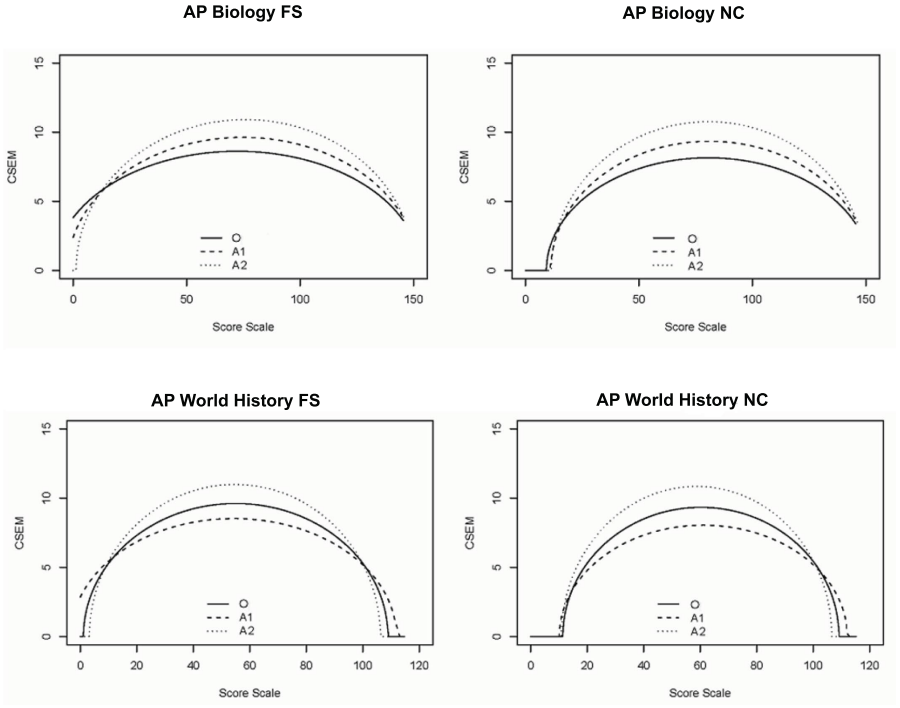

Figure 2: Comparison of CSEM for Operational and Alternative Relative Section Weights

Note. AP Biology MC relative weights: O=0.6, A1=0.5, A2=0.4. AP World History MC relative weights: O=0.5, A1=0.6, A2=0.4.

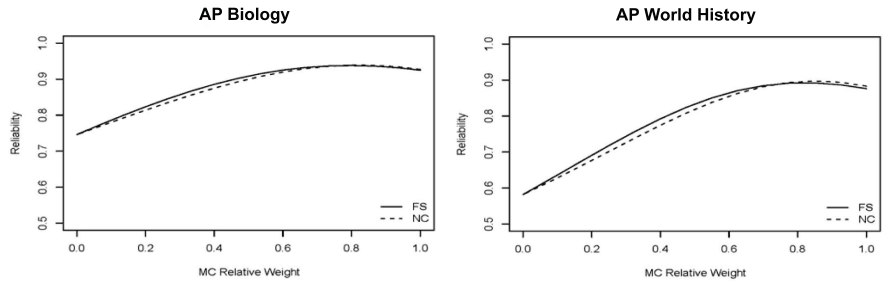

Figure 3: Relationship of MC Relative Weights and Reliability for the Operational Condition (FS and NC)

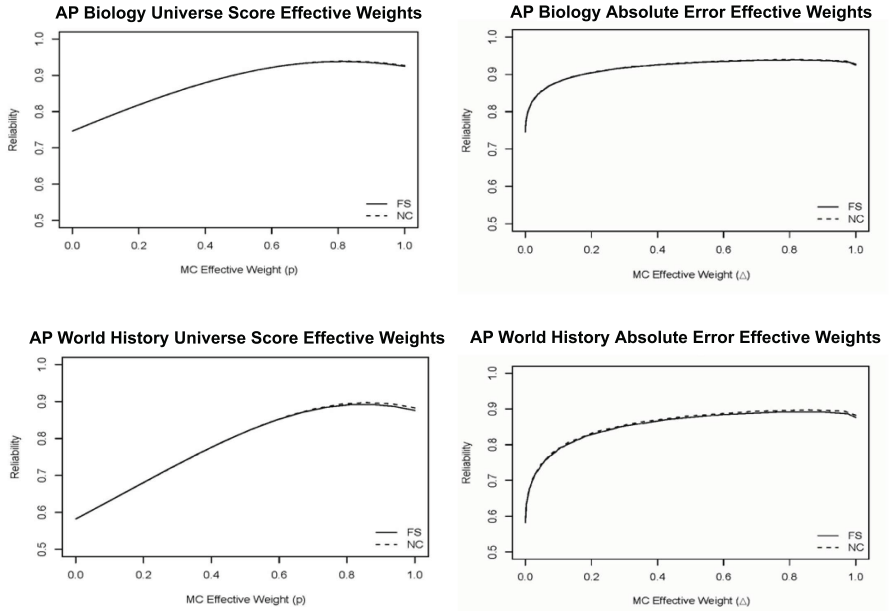

Figure 4: Relationship of MC Effective Weights and Reliability for the Operational Condition (FS and NC)

Note. MC Effective Weight (Δ) represents Absolute Error Effective Weight. MC Effective Weight (p) represents Universe Score Effective Weight.

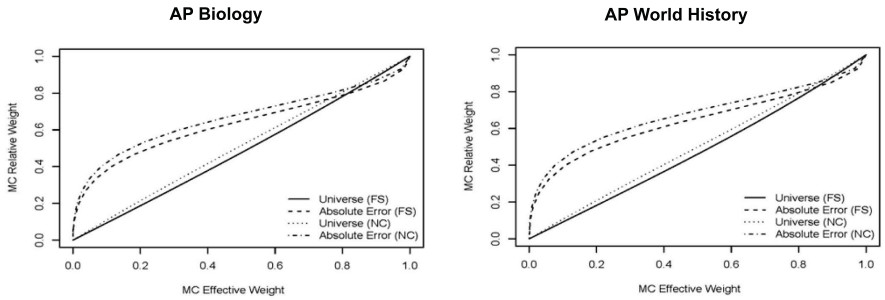

Figure 5: Relationship between Relative and Effective Weights for FS and NC Scoring

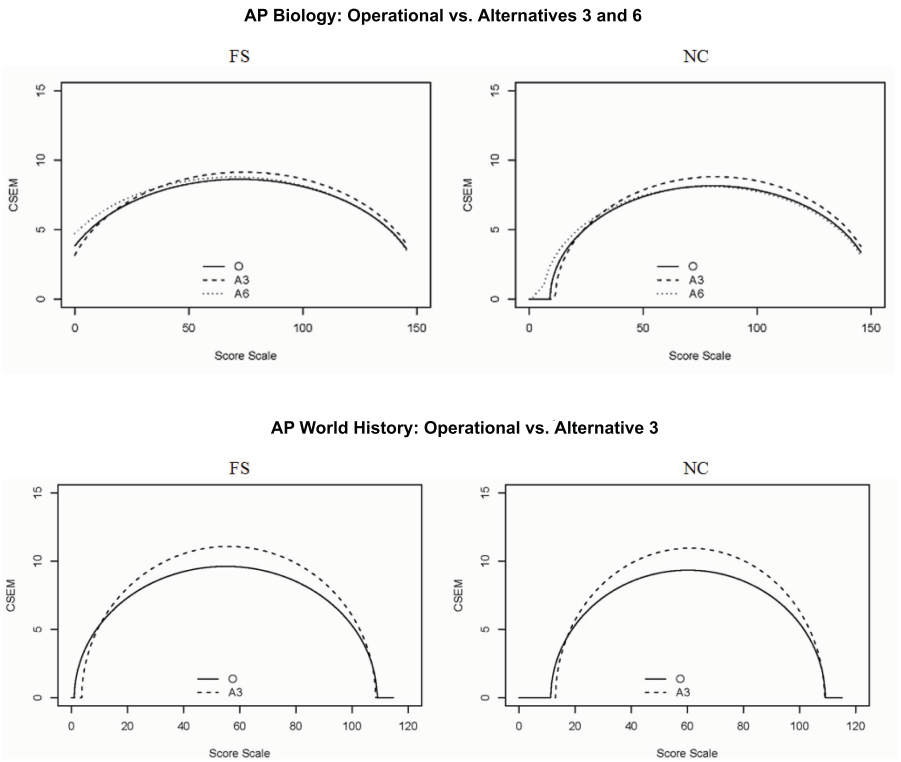

Figure 6: Comparison of CSEM for Operational and Alternative Numbers of MC and FR Items

Note. AP Biology O=100 MC items, 4 FR items; A3=130 MC items, 3 FR items; A6=70 MC items, 5 FR items.
AP World History O=70 MC items, 3 FR items; A3=120 MC items, 2 FR items.

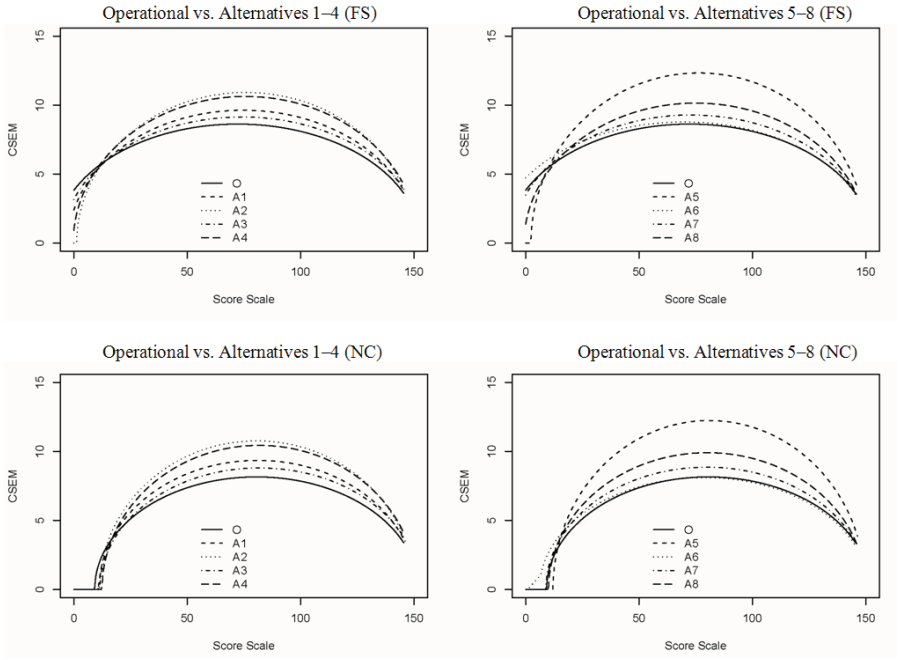

Figure 7: Comparison of CSEM across all Conditions for AP Biology

Note. See Table 2 for a description of weights and numbers of items for both MC and FR sections across Operational and eight Alternative conditions.

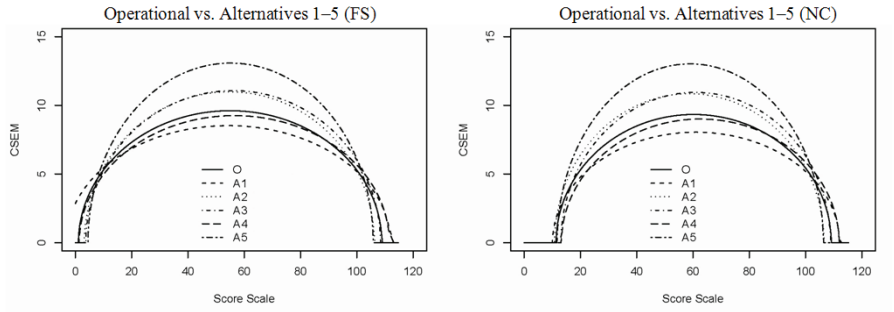

Figure 8: Comparison of CSEM across all Conditions for AP World History

Note. See Table 3 for a description of weights and numbers of items for both MC and FR sections across Operational and five Alternative conditions.