The Contribution of Constructed Response Items to Large Scale Assessment: Measuring and Understanding their Impact

Robert W. Lissitz[1] and Xiaodong Hou[2]

University of Maryland

Sharon Cadman Slater

Educational Testing Service

©Journal of Applied Testing Technology, 2012, Volume 13, Issue #3

[1] Send reprint requests to Dr. Robert W. Lissitz, 1229 Benjamin Building, University of Maryland, College Park, MD, 20742.

[2] We would like to thank the Maryland State Department of Education (MSDE) for their support of the Maryland Assessment Research Center for Education Success (MARCES) and the work pursued in this study. The opinions expressed here do not necessarily represent those of MSDE, MARCES, or the Educational Testing Service.

Abstract

This article investigates several questions regarding the impact of different item formats on measurement characteristics. Constructed response (CR) items and multiple choice (MC) items obviously differ in their formats and in the resources needed to score them. As such, they have been the subject of considerable discussion regarding the impact of their use and the potential effect of ceasing to use one or the other item format in an assessment. In particular, this study examines the differences in constructs measured across different domains, changes in test reliability and test characteristic curves, and interactions of item format with race and gender. The data for this study come from the Maryland High School Assessments that are high stakes state examinations whose passage is required in order to obtain a high school diploma.

Our results indicate that there are subtle differences in the impact of CR and MC items. These differences are demonstrated in dimensionality, particularly for English and Government, and in ethnic and gender differential performance with these two item types.

Key words: constructed response items, multiple choice items, large-scale testing

The Contribution of Constructed Response Items to Large Scale Assessment: Measuring and Understanding their Impact

Introduction

Both multiple choice (MC) items and constructed response (CR) items have been widely used in large-scale educational testing. MC items require examinees to select a response from given options, while CR items present an item stem and require examinees to construct a response "from scratch." MC items generally demonstrate greater content validity than do CR items (Newstead & Dennis, 1994). Further, because MC items can be answered relatively quickly by examinees, a broader portion of the domain can be assessed efficiently by administering more such items. Further, it is well known that tests comprised of more items tend to have higher test reliability (Angoff, 1953; Crocker & Algina, 1986). MC items can also be easily and accurately scored making them cost-efficient. The ease of scoring also permits score reporting to be accomplished more quickly, thus providing students and teachers with feedback on performance in a timelier manner. All these elements make MC items very attractive. Many educators have come to rely increasingly upon closed-ended (primarily MC) rather than CR items due to their efficiency (Bleske-Rechek, et al., 2007).

However, such reliance raises the question of whether the exclusive use of MC items is the best decision. In fact, the Race to the Top Program's application for new grants for Comprehensive Assessment Systems, among many other requirements, calls for a system that "elicits complex student demonstrations or applications of knowledge and skills" (U.S. Department of Education, 2010). And both consortia – the Smarter Balanced Assessment Consortium and the Partnership for the Assessment of Readiness for College and Career – have

CR items, as well as more extended performance assessments as part of their assessment designs (Center for K-12 Assessment, 2012). Some practitioners argue that MC items fail to elicit the higher levels of cognitive processing and that MC items engage examinees in a type of guessing game (Campbell, 1999). It is believed by some researchers that MC items are unable to tap higher order thinking and allow for a higher probability of guessing correctly which causes lower reliabilities in the test for lower ability students (Cronbach, 1988). Nevertheless, many suggest that MC items can measure essentially the same cognition as CR items (Kennedy & Walstad, 1997). Hancock (1994) pointed out that proponents of the MC format believe that MC items can be written to tap complex thinking though it is more difficult to write such MC items than CR items.

Although MC items can be designed to measure reasoning skills, many researchers think that they cannot elicit the important constructive cognitive processes as effectively as CR items do. CR items are believed to be best suited to test reasoning abilities such as evaluating, synthesizing, and analyzing (Stiggins, 2005). Because the learning and development process involves the active construction of knowledge, learning theorists such as Piaget and Vygotsky believe active construction of knowledge is needed in education (Bedrova & Leong, 1996; Berk & Winsler, 1995). Since CR items require the active construction of knowledge in the examinees' reasoning process by using their own knowledge to produce a solution (Reiss, 2005) these items are seen as more like the reasoning that learning theorists encourage. CR items allow for a range of answers, all of which are provided by examinees requiring the origination of ideas rather than the recognition of them. CR items reduce the probability of correct guessing to essentially zero because the correct answer is not shown in a CR item. In addition, CR items may directly show what examinees think and expand the possibility of creative thinking since they require that the

examinee construct a response in their own words. Therefore, in many large-scale tests, CR items are included in spite of the relatively expensive and inevitable subjectivity in scoring.

Assessment practices are also believed to influence what and how something is taught. If the ability to construct rather than select a solution is not assessed, it might be neither taught nor learned (Wiggins, 1993; Rodriguez, 2002; Reiss, 2005). In other words, if only MC items are included in the assessment, then only the skill to select from given options might be taught. If MC items are exclusively used in testing, reasoning skills such as evaluating, synthesizing, and analyzing might not be the focus of instruction and learning. If we use only MC items, we may risk the loss of the active construction of knowledge, which is important in the learning process. Also, examinees seem to prepare harder for CR items than MC items (Snow, 1993).

Some educators have expressed concern that exclusive use of one type of item may put some students at a disadvantage. For example, Bridgeman and Morgan (1996) suggest that some students may perform poorly on MC items but do quite well on essay assessments; others may do well on MC items but poorly on essays. Further, CR items allow examinees to receive partial credit on an item, whereas MC items typically do not. Students who perform well on only one type of item may be unintentionally disadvantaged in assessments that rely on only the other item format.

There are many research papers showing that performance on a test might be affected by item format. Performance differences between genders are often found in these studies (Garner & Engelhard, 1999; Mislevy, 1993; Reiss, 2005; Traub, 1993; Willingham & Cole, 1997). Females have been found to have an advantage on CR items, which might be explained by their generally better performance on tests of language ability (Halpern, 2004; Reiss, 2005).

Investigating the relationship between MC and CR items

In order to answer the question of what CR items may uniquely contribute to large-scale assessments, we need to know whether typical MC items assess the same thing as CR items (Bennett et al., 1991; Stout, 1990; Thissen, Wainer & Wang, 1994). This is certainly critical to the use of the usual IRT models, classical test and generalizability theory methods that depend upon unidimensionality. A number of researchers have investigated MC and CR items and the contributions of the two types of items to test information about student achievement, but the evidence is inconclusive (Martinez, 1999, Rodriguez, 2003). In addition, whether CR and MC items measure different constructs may also be dependent on the content domain. Based on the review of nine studies, Traub (1993) concluded that item type appears to make little difference in the reading comprehension and quantitative domains, but for the writing domain, different item types measure different constructs.

Bennett and his colleagues (1990) used the College Board's Advanced Placement Computer Science (APCS) examination as construct validity criteria for an intermediary item type and indirectly examined the relationship of the two item formats. They found little support for the existence of construct differences between the formats. Later, Bennett and his colleagues (1991) employed confirmatory factor analysis (CFA) again in the APCS examination to assess the equivalence of MC and CR items by comparing the fit of a single factor and a two-factor model, where each item type represents its own factor. Both MC and CR items were written to measure the same content, but CR items were intended to assess selected topics more deeply. It was found that a single factor provided a more parsimonious fit. CFA was also used in Manhart's (1996) study to investigate whether the MC and CR science tests measured the same construct. Each test was divided into several parcels of items where each parcel contained either all MC or all CR

items. The fit of the one-factor and two-factor models was compared by inspecting the chi-square tests and standardized residuals. They concluded that the two-factor model was generally more appropriate for explaining the covariance between the parcels than the one-factor model.

Bridgeman and Rock (1993) used exploratory principal components analysis to examine relationships among existing item types and new computer-administered item types for the analytical scale of the Graduate Record Examination General Test. By analyzing the correlation matrix of item parcels with a principal components model, the number of factors to extract was determined. The MC and CR items of the same task were found to load on the same factor and the new CR version, which was a counterpart of the MC version, did not tap any different dimension, significantly.

Thissen et al. (1994) proposed a model assuming a general factor for all items plus orthogonal factors specific to CR items. They found that both MC and CR items largely loaded on the general factor, and CR items loaded significantly on CR specific factors.

Whether the two formats are able to measure the same construct is an important issue to investigate. Many commonly used measurement models assume unidimensionality to calibrate and score the test and to construct appropriate score reporting strategies. If there is no effect of item format, the dimensionality of the mixed-format assessment will depend on the nature of the items themselves—that would be whether the two formats are designed to be counterparts of one another or tap different skills (Thissen, Wainer & Wang, 1994).

Using Maryland High School Assessments (MDHSAs), this article investigates several questions regarding the impact of the two item formats on measurement characteristics. In particular, the study examines the differences in constructs measured across different domains,

changes in test reliability and test characteristic curves, and interactions of item format with race and gender.

Participants

The analysis is conducted on the results from the 2007 MDHSA Form E. Table 1 shows the number of participants in each test by race and gender. The mean number of participants across the four content areas is 10,555. For the four content areas, on average, about 49% of the examinees were white and 39% were African American. The remaining examinees were Hispanic (about 7% on average) and Asian/Pacific (about 6% on average). The American Indian group (about 0.3% on average) had sample sizes too small to be included in our analysis. Gender was distributed evenly with a few more male than female students in the Algebra, Biology and Government tests, and a few more female than male students in the English test. Overall the percentages of male and female examinees are 50.7% and 49.3%, respectively.

Table 1. Participant Demographic Information

2007 HSA Form E		Algebra	English	Biology	Government
Total (counts)		13030	9263	9438	10491
Race (%)	White	47.8	48.7	50.7	47.6
	African American	38.9	38.6	36.4	39.3
	Hispanic	7.6	6.6	6.6	6.8
	Asian/Pacific Islander	5.4	5.8	6.0	5.9
	American Indian	0.3	0.3	0.3	0.4
Gender (%)	Male	51.2	49.4	51.0	51.2
	Female	48.8	50.6	49.0	48.8

The analyses used four random samples of 2,000 students from the whole population taking the 2007 MDHSA Form E so that model-fit indexes are more comparable across content areas. The distributions of gender and race of the samples are very similar to their population, which are shown in Table 2. The completely random sampling in this study makes it possible to generalize the conclusions to the population, while making the results a little easier to compare and to understand.

Table 2. Subgroup Participant Demographic Information (Sample=2000)

2007 HSA Form E		Algebra	English	Biology	Government
Race (%)	White	48.2	49.3	50.8	48.7
	African American	38.7	38.1	35.9	38.2
	Hispanic	7.4	6.7	7.4	7.3
	Asian/Pacific Islander	5.9	6.0	6.0	5.9
Gender (%)	Male	51.8	49.6	51.4	51.9
	Female	48.2	50.5	48.7	48.1

Instruments

The MDHSAs are end-of-course tests that cover four subjects: Algebra, Biology, English, and Government. The 2007 tests were composed of MC items and CR items. MC items were machine-scored and CR items were scored by human raters[3]. In addition, the Algebra tests have student-produced response items or "gridded" response (GR) items which require students to grid in correct responses on the answer document. Because they are not clearly MC or CR test items, they were not included in some analyses.

In all four tests, MC and CR items were designed to test the same expectations in the content areas, hence the knowledge and skills required to answer them were originally expected to be very similar. The content coverage of each test is shown in Tables 3 to 6 (Maryland State Department of Education, 2008).

[3] For a detailed description of how CR items were scored for the MDHSAs, please see http://www.marylandpublicschools.org/NR/rdonlyres/099493D7-805B-4E54-B0B1-3C0C325B76ED/2386/432002ScoringContractorsReport.pdf

Table 3. Algebra Blueprint

Reporting category	Number of items (points)		Total points
	MC(1pt)	CR	
Expectation 1.1 The student will analyze a wide variety of patterns and functional relationships using the language of mathematics and appropriate technology.	8	1 (4pt)	13
Expectation 1.2 The student will model and interpret real world situations, using the language of mathematics and appropriate technology.	10	1 (4pt)	14
Expectation 3.1 The student will collect, organize, analyze, and present data.	4	2 (3pt)	10
Expectation 3.2 The student will apply the basic concepts of statistics and probability to predict possible outcomes of real-world situations.	4	2(3&4pt)	10
Total counts	26	6	47 32

Table 4. Biology Blueprint

Reporting category	Number of items (points)		Total points
	MC(1pt)	CR(4pt)	
Goal 1 Skills and Processes of Biology	8	2	16
Expectation 3.1 Structure and Function of Biological Molecules	8	1	12
Expectation 3.2 Structure and Function of Cells and Organisms	9	1	13
Expectation 3.3 Inheritance of Traits	9	1	13
Expectation 3.4 Mechanism of Evolutionary Change	5	1	9
Expectation 3.5 Interdependence of Organisms in the Biosphere	9	1	13
Total counts	48	7	76 55

Table 5. English Blueprint

Reporting category	Number of items (points)		Total points
	MC(1pt)	CR	
1: Reading and Literature: Comprehension and Interpretation	13	1(3pt)	16
2: Reading and Literature: Making Connections and Evaluation	11	1(3pt)	14
3: Writing – Composing	8	2(4pt)	16
4: Language Usage and Conventions	14	0	14
Total counts	46	4	60 50

Table 6. Government Blueprint

Reporting category	Number of items (points)		Total points
	MC(1pt)	CR(4pt)	
Expectation1.1 The student will demonstrate understanding of the structure and functions of government and politics in the United States	13	3	25
Expectation 1.2 The student will evaluate how the United States government has maintained a balance between protecting rights and maintaining order.	11	2	19
Goal 2 The student will demonstrate an understanding of the history, diversity, and commonality of the peoples of the nation and world, the reality of human interdependence, and the need for global cooperation, through a perspective that is both historical and multicultural.	8	1	12
Goal 3 The student will demonstrate an understanding of geographic concepts and processes to examine the role of culture, technology, and the environment in the location and distribution of human activities throughout history.	7	1	11
Goal 4 The student will demonstrate an understanding of the historical development and current status of economic principles, institutions, and processes needed to be effective citizens, consumers, and workers.	11	1	15
Total counts	50	8	82 58

Methods

The models tested in this study are similar to those used by Bennett et al. (1991). The domains investigated in the studies of Bennett et al. (1991) and Thissen et al. (1994) were computer science and chemistry. In this paper, we proposed two-factor CFA models for the four content areas: Algebra, Biology, English and Government. The factors represent the two item formats. Factors were allowed to be correlated and items were constrained to load only on the factor that was assigned in advance.

Since all the indicators were treated as categorical variables in our study, all testing of the CFA models was based on Robust Maximum Likelihood (ML) estimation in EQS, which can be used when a researcher is faced with problems of non-normality in the data (Byrne, 2006). In other words, the robust statistics in EQS are valid despite violation of the normality assumption underlying the estimation method. Robust ML estimation is used in analyzing the correlation matrix, and chi-square and standard errors are corrected (i.e., Satorra-Bentler scaled chi-square and Robust standard errors) through use of an optimal weight matrix appropriate for analysis of categorical data.

To assess the fit of the two-factor models, factor inter-correlations and goodness-of-fit were checked and the model's fit was compared to two alternative models, a one-factor CFA model and a null model in which no factors were specified. The following goodness-of-fit indicators were considered in our study: Satorra-Bentler scaled chi-square/degrees of freedom ratio (S-B χ^2 /df), Comparative Fit Index (CFI), Bentler-Bonett Normed Fit Index (NFI), Bentler-Bonett Non-normed Fit Index (NNFI), Root Mean-Square Error of Approximation (RMSEA) and Akaike information criterion (AIC). Low values of the ratio of chi-square/degrees of freedom indicate a

good fit. However, there is no clear-cut guideline. An S-B χ^2/df value of 5.0 or lower has been recommended as indicating a reasonable fit but this index does not completely correct for the influence of sample size (Kline, 2005). Therefore, other indexes, which are less affected by sample size, were considered. NFI has been the practical criterion of choice for a long time but was revised to take sample size into account, called the CFI. CFI is one of the incremental fit indexes and the most widely used in structural equation modeling. It assesses the relative improvement in fit of the researcher's model compared with the null model. A value greater than .90 indicates a reasonably good fit (Hu & Bentler, 1999). NNFI assesses the fit of a model with reference to the null model, and occasionally falls outside the 0-1 range. The larger the value, the better the model fit. RMSEA is a "badness-of-fit" index with a value of zero indicating the best fit and higher values indicating worse fit. It estimates the amount of error of approximation per model degree of freedom and takes sample size into account. In general, RMSEA $\leq$ .05 indicates close approximate fit and RMSEA $\leq$ .08 suggests reasonable error of approximation (Kline, 2005). AIC is an index of parsimony, which considers both the goodness-of-fit and the number of estimated parameters; the smaller the index, the better the fit (Bentler, 2004).

Item parcels have been suggested for use in factor analysis modeling in the study of test dimensionality. Cook et al. (1988) believed that using parcels instead of individual items could help insure the covariance matrix is not a function of item difficulty if approximately equal numbers of easy and difficult items are placed in each parcel. Bennett et al. (1991) used item parcels in their study investigating the relationship between MC and CR items where the mean difficulty values for the MC parcels were similar. In this paper, we used a similar strategy to Thissen et al. (1994) to build the MC item parcels without respect for item content in hope that

the parcels would be approximately equally correlated. In addition, two factors were considered in the decisions on the size and number of the parcels in each content area. First, each parcel included an equal or similar number of items within a content area. Second, the total number of both MC parcels and CR items (i.e., the total number of loadings in factor analysis) for each content area remained equal or similar across four content areas. Items were ordered in difficulty and then selected with equal interval in ranking order (in other words, the range in ranks would be equal) so that each parcel has approximately equal difficulty with maximum variation in parcel-summed scores. For example, if there are 18 items that are divided into 3 item parcels, the items 1, 4, 7, 10, 13, and 16 might be in the first parcel. Similarly, item 2, 5, 8, 11, 14, 17 could be in the parcel 2, and the third parcel goes from item 3 to item 18, so that the range of the ranks is equal in these three parcels.

Reliability was investigated and compared for the four different content area tests before and after the CR items were removed. Spearman Brown prediction was used when investigating reliability issues to counter the effect of changing the number of items of the test. Test Characteristic Curves were also compared with and without CR items, with various strategies used to replace the CR items with MC items. The interaction of item format with gender and ethnicity was examined by looking at the consistency of the changes in the percentage points obtained when going from MC to CR items.

Results

The 2007 HSA form E of the Algebra, English, Biology and Government tests were analyzed to investigate the implications of removing CR items from the tests. Number-right scoring was used in the analysis. Omitted responses were considered missing values and were deleted from the analysis.

Reliability

Algebra: The reliability of the Algebra test decreased from .91 to .88 when CR items were removed from the test. The reader will note in Table 7 that both the reliability of the Algebra test and the SEM decreased after the CR items were removed. It may be that simply increasing the number of MC items would counter this effect. In order to examine whether increasing the number of MC items would counter the effect, the Spearman Brown Prophecy Formula

$$\rho_{xx'} = \frac{k\rho_{jj'}}{1+(k-1)\rho_{jj'}},$$

was employed to calculate reliability for a new test in which new parallel items are hypothesized to be added to compensate for dropping the CR items, where $\rho_{jj'}$ is the reliability of the test without CR items, and *k* is the ratio of new test length to original test. The Spearman-Brown prophecy formula assumes that any additional items would have similar characteristics to the items on which the initial estimate is based. Therefore in this study it was assumed that the intercorrelations of the newly added items are similar to those of the existing items when the new reliabilities were calculated. The new reliability for the lengthened HSA Algebra test is .93, slightly higher than the original test.

Table 7. Internal Consistency Reliability of Tests Scores With and Without Constructed Response Items

Content Area	Reliability& SEM	Test with CR	Test without CR	New Lengthened Test without CR
Algebra	Coefficient Alpha	.91	.88	.93
	SEM	3.37	2.28	----
English	Coefficient Alpha	.90	.88	.91
	SEM	3.03	2.71	----
Biology	Coefficient Alpha	.93	.89	.93
	SEM	3.45	2.93	----
Government	Coefficient Alpha	.94	.91	.95
	SEM	3.61	2.94	----

Biology: The reliability of the Biology test decreased from .93 to .89 when the CR items were removed from the test. The reliability for the new lengthened Biology test increased by .003, using the Spearman Brown Prophecy Formula assuming MC items replaced the CR items.

English: The reliability of the English test dropped by .017 when CR items were removed from the test. When the Spearman Brown Prophecy Formula was employed to calculate reliability for the new lengthened test, reliability was .91, which was higher than the original test by 0.008.

Government: The reliability of the Government test reduced from .94 to .91 when CR items were removed from the test .The new reliability using the Spearman Brown Prophecy Formula for the new lengthened government test was .95, which is larger than the reliability of the original test by 0.006.

Confirmatory Factor Analysis

Algebra: The MC section was divided into five "item parcels," resulting in four 5-MC-item parcels and one 6-MC-item parcel. The MC item parcels and CR items were used in the analysis. Focusing on the ROBUST fit indexes, the two-factor model produced good results with fit indices of CFI value of .97, NFI value of .97, NNFI value of .97, and a RMSEA value of .074, with a 90% C.I. ranging from .068 to .079. Those indices were slightly improved compared with

the values found in the one-factor model. The chi-square difference between the one-factor and two-factor models was 85.06 with 1 degree of freedom, $p<.01$. However, the inter-correlation of the two factors was .94, $p<.01$. In addition, the S-B χ^2/df ratio was relatively large, which indicates poor fit. This lack of fit may be explained when we examine the standardized residuals.

Standardized residuals were between zero to .07 in magnitude, with the exception of one residual value of .25 of CR4 and CR6 that were designed to test the same expectations. This may explain the lack of fit that the chi-square test indicated above. The average off-diagonal absolute standardized residual (AODASR) is .03, which reflects that overall little covariation remained. Similar results were found in the one-factor model. Except for the residual value .29 of CR4 and CR6, all standardized residuals ranged from zero to .07. The AODASR was .03.

Table 8. Confirmatory Factor Analysis Results: Algebra Data

	Fit index					
Model (N=2000)	Chi-square/df	NFI	NNFI	CFI	RMSEA(90% CI)	AIC
Two-factor	510.42/43	.97	.97	.97	.07 (.07-.08)	424.42
One-factor	595.48/44	.97	.96	.97	.08 (.07-.08)	507.48
Null	17571.68/55	--	--	--	--	17461.68

Table 9. Loadings of MC Item Parcels and CR Items for Algebra

	Two-factor model		One-factor model
	MC factor	CR factor	General factor
MC parcel 1	.80	--	.78
MC parcel 2	.77	--	.76
MC parcel 3	.74	--	.73
MC parcel 4	.79	--	.77
MC parcel 5	.77	--	.76
CR 1	--	.84	.83
CR 2	--	.63	.62
CR 3	--	.80	.78
CR 4	--	.66	.62
CR 5	--	.63	.63
CR 6	--	.76	.73

All loadings were significant at the .05 level. Loadings of CR2, CR4 and CR5 on the CR factor were relatively lower than others in both CFA models. This was probably due to the common variance shared by these CR items and MC items which were designed to be parallel or due to the lower reliability of the CR items.

Biology: The MC section was divided into six 8-item parcels. Based on the ROBUST fit indexes of the S-B χ^2/df ratio, CFI, NFI, NNFI and RMSEA, the two-factor model fit the data well. Those indices were slightly improved compared with the values found in the one-factor model. The chi-square difference between one-factor and two-factor models is 231.66 with 1 degree of freedom, $p<.01$. Hence, the two-factor model has statistically better fit than the one-factor model. The intercorrelation between the two factors is .88, $p<.001$.

Standardized residuals were between zero to .09 for the two-factor model. The average off-diagonal absolute standardized residuals (AODASR) is .02, which reflects little overall covariation. Similar results were found in the one-factor model where all standardized residuals ranged from zero to .08, except for the residual value of CR1 and CR2, which was.11. The AODASR is .04. All loadings are significant at the .05 level in both models, which show very similar patterns.

Table 10. Confirmatory Factor Analysis Results: Biology Data

Model (N=2000)	Fit index					
	Chi-square/df	NFI	NNFI	CFI	RMSEA(90% CI)	AIC
Two-factor	425.05/64	.98	.98	.98	.05 (.05-.06)	297.05
One-factor	656.71/65	.98	.98	.98	.07 (.06-.07)	526.74
Null	28577.81/78	--	--	--	--	28421.81

Table 11. Loadings of MC Item Parcels and CR Items for Biology

	Two-factor model		One-factor model
	MC factor	CR factor	General factor
MC parcel 1	.83	--	.78
MC parcel 2	.74	--	.71
MC parcel 3	.80	--	.76
MC parcel 4	.71	--	.68
MC parcel 5	.82	--	.78
MC parcel 6	.72	--	.69
CR 1	--	.72	.70
CR 2	--	.81	.80
CR 3	--	.82	.81
CR 4	--	.84	.84
CR 5	--	.83	.80
CR 6	--	.75	.73
CR 7	--	.83	.80

English: The MC section was divided into nine "item parcels," resulting in eight 5-MC-item parcels and one 6-MC-item parcel. Focusing on the ROBUST fit indexes, the two-factor model had better results than the one-factor model, which is shown in Table 12. The chi-square difference between the one-factor and two-factor models is 587.4 with 1 degree of freedom, $p<.01$, meaning that the two-factor model is statistically better than the one-factor model. In the two-factor model, the inter-correlation of the two factors was .74, $p<.05$, which indicates that there is a certain degree of difference in what the two types of item formats measure. The S-B χ^2 /df ratio is marginally acceptable, and other fit indices were good.

Absolute standardized residuals were between zero to .13, and .03 on average. The average off-diagonal absolute standardized residual (AODASR) was .03, which reflects that little

covariation remained. However, in the one-factor model, the largest standardized residual was .36 between CR1 and CR3, which may indicate some association within or across item formats.

Table 12. Confirmatory Factor Analysis Results: English Data

Model (N=2000)	Fit index					
	Chi-square/df	NFI	NNFI	CFI	RMSEA(90% CI)	AIC
Two-factor	406.37/64	.97	.97	.97	.05 (.05-.06)	278.37
One-factor	993.77/65	.94	.91	.93	.09 (.08-.09)	863.77
Null	13034.69/78	--	--	--	--	12878.69

Table 13. Loadings of MC Item Parcels and CR Items for English

	Two-factor model		One-factor model
	MC factor	CR factor	General factor
MC parcel 1	.61	--	.60
MC parcel 2	.68	--	.67
MC parcel 3	.66	--	.65
MC parcel 4	.76	--	.75
MC parcel 5	.71	--	.70
MC parcel 6	.71		.70
MC parcel 7	.72		.70
MC parcel 8	.72		.71
MC parcel 9	.74		.71
CR 1	--	.70	.57
CR 2	--	.82	.68
CR 3	--	.78	.64
CR 4	--	.79	.65

All loadings were significant at the .05 level. Loadings of MC parcels on the MC-factor in the two-factor model were slightly higher than those on the general factor in the one-factor model, whereas loadings of CR items on the CR-factor in the two-factor model were much higher than those on the general factor in the one-factor models.

Government: The MC section was divided into five 10-item parcels. ROBUST fit indexes showed that the two-factor model fit the data well. See the S-B χ^2/df ratio, CFI, NFI, NNFI and RMSEA, in Table 14. The chi-square difference between the one-factor and two-factor models

was 823.16 with 1 degree of freedom, $p<.01$. Hence, the two-factor model had statistically better fit than the one-factor model. The inter-correlation between the two factors is .83, $p<.001$.

Standardized residuals were between zero to .07. The average off-diagonal absolute standardized residual (AODASR) was .02, which reflects that little covariation remained. These residuals were smaller than those in the one-factor model, where there were eight residuals beyond the value of .10, and the AODASR was .05. All loadings were significant at the .05 level in both models. However, loadings in the two-factor model were higher than those in the one-factor model.

Table 14. Confirmatory Factor Analysis Results: Government Data

	Fit index					
Model (N=2000)	Chi-square/df	NFI	NNFI	CFI	RMSEA(90% CI)	AIC
Two-factor	318.10/64	.99	.99	.99	.05 (.04-.05)	190.10
One-factor	1141.26/65	.97	.97	.97	.09 (.09-.10)	1011.26
Null	39471.73/78	--	--	--	--	39315.73

Table 15. Loadings of MC Item Parcels and CR Items for Government

	Two-factor model		One-factor model
	MC factor	CR factor	General factor
MC parcel 1	.86	--	.79
MC parcel 2	.86	--	.77
MC parcel 3	.83	--	.75
MC parcel 4	.82	--	.75
MC parcel 5	.80	--	.74
CR 1	--	.81	.80
CR 2	--	.83	.82
CR 3	--	.85	.83
CR 4	--	.83	.82
CR 5	--	.82	.83
CR 6	--	.80	.79
CR 7	--	.85	.81
CR 8	--	.85	.82

Changes in Test Characteristic Curves

In order to investigate whether the tests are more or less difficult and discriminating after removing CR items, plots of Test Characteristic Curves (TCCs) were examined to show the differences between the curves with and without CR items. Multiple-choice items were calibrated using the 3-parameter item response model; and constructed-response items were calibrated using the graded partial credit model. An underlying assumption of IRT models is unidimensionality. Previous CFA studies that were not focused on item type have supported that assumption (MSDE, 2009).

In this analysis, to compensate for the removal of CR items, CR items were replaced on a point-by-point basis with multiple-choice items selected in a number of ways. The number of total points was kept the same by replacing each CR item with as many MC items as matched the point value of the CR item. So, a 3-point CR item was replaced by three MC items, a 4-point CR item was replaced by four MC items, and so on. MC replacement items were selected in three different ways, resulting in three different versions of TCCs for each content area. In Version 1, CR items were replaced with MC items from the bank that matched each of the score point *b*-values on the CR. For example, for a 3-point CR item, one MC replacement item was chosen to best match the CR *b*-value for a score of zero, the second MC replacement item was chosen to best match the CR *b*-value for a score of one, and the third MC replacement item was chosen to best match the CR *b*-value for a score of two. In Version 2, all CR items were replaced with enough MC items as the point value of the CR item, where all MC replacement item *b*-values matched as close as possible to the *b*-value of the highest CR point. And in Version 3, all MC replacement items were selected randomly.

From looking at Figures 1-4 below, a number of patterns can be seen. First, for all four content areas, the lower-asymptote of the TCCs without CR items is consistently higher than the

May 2007 version of each test that contained CR items. This is due to the fact that CR items, for which there is little to no guessing, have been replaced with MC items where examinees have a one in four chance of randomly guessing a correct answer. In fact, in the high stakes environment of the MDHSA, with items with plausible distractors, guessing on MC items can take on a range of non-zero values. The point here is that the CR items do not permit guessing therefore TCCs for tests containing CR items tend to have decreased lower-asymptotes, as can be seen in the solid lines in Figures 1-4. Another observation that can be made from the TCCs is that for Algebra and English, all versions of the MC-only tests are about as discriminating as the May 2007 mixed-format versions. For Biology and Government, Versions 1 and 3 are about as discriminating as the original version, and Version 2 is less discriminating. However, in all content areas but English, the TCCs of Versions 1 and 3 are shifted to the left, meaning that those versions of the tests were less difficult overall than the version that contained CR items. For English, the overall difficulty of the test was closest to the original difficulty for Version 1, where MC replacement items for each CR were selected to match *b*-values of each of the CR score points. For Algebra, Version 2 was a better match of overall test difficulty. Version 2 was created by choosing all of the MC replacement items for each CR to match the difficulty of the highest score point of the CR item. Biology and Government did not appear to have a best MC-only match for overall difficulty. For these two content areas, Versions 1 and 3 were easier, and Version 2 was more difficult at the upper and of the ability scale and somewhat easier at the lower end.

The three versions used for selecting replacement items were chosen because of their likelihood to be employed by assessment developers, who are typically very familiar with selecting items based on item difficulty. However, the critical components in matching TCCs

may be item difficulty and item discrimination, or the *a*-value, rather than difficulty alone. If MC replacement items had been selected based on the best match of both the difficulty and discrimination parameters for each CR score point, even closer results between the mixed item type forms and the multiple-choice only forms may have been obtained.

Figure 1.Test Characteristic Curves for Algebra

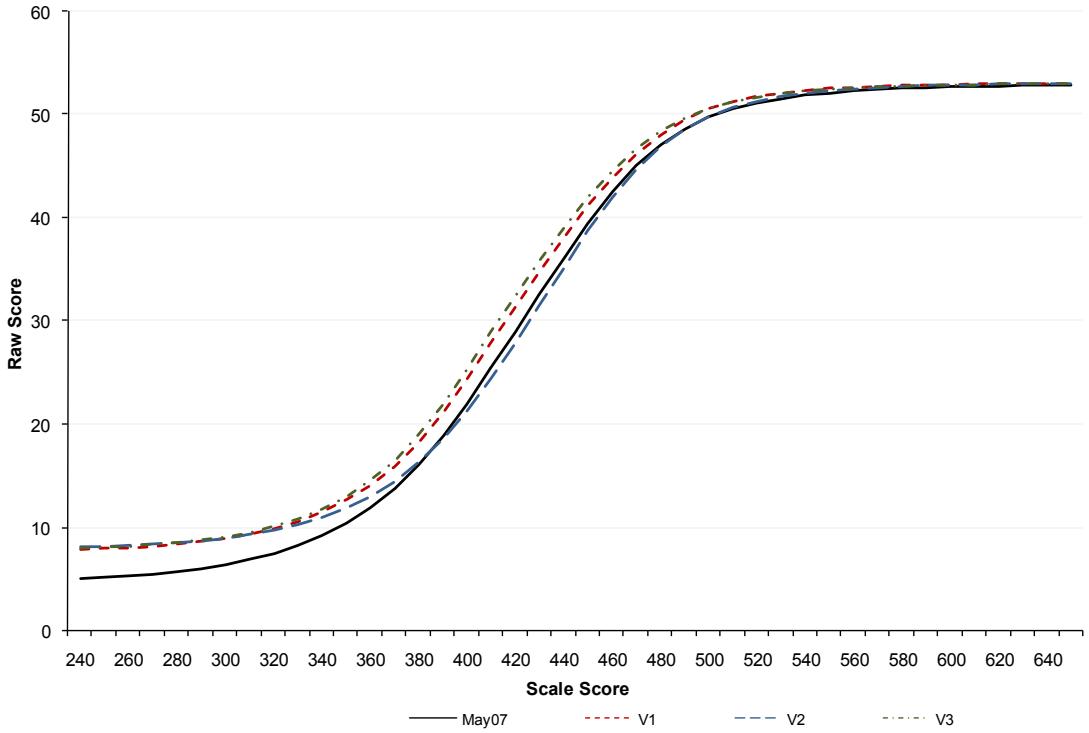


Figure 2.Test Characteristic Curves for Biology

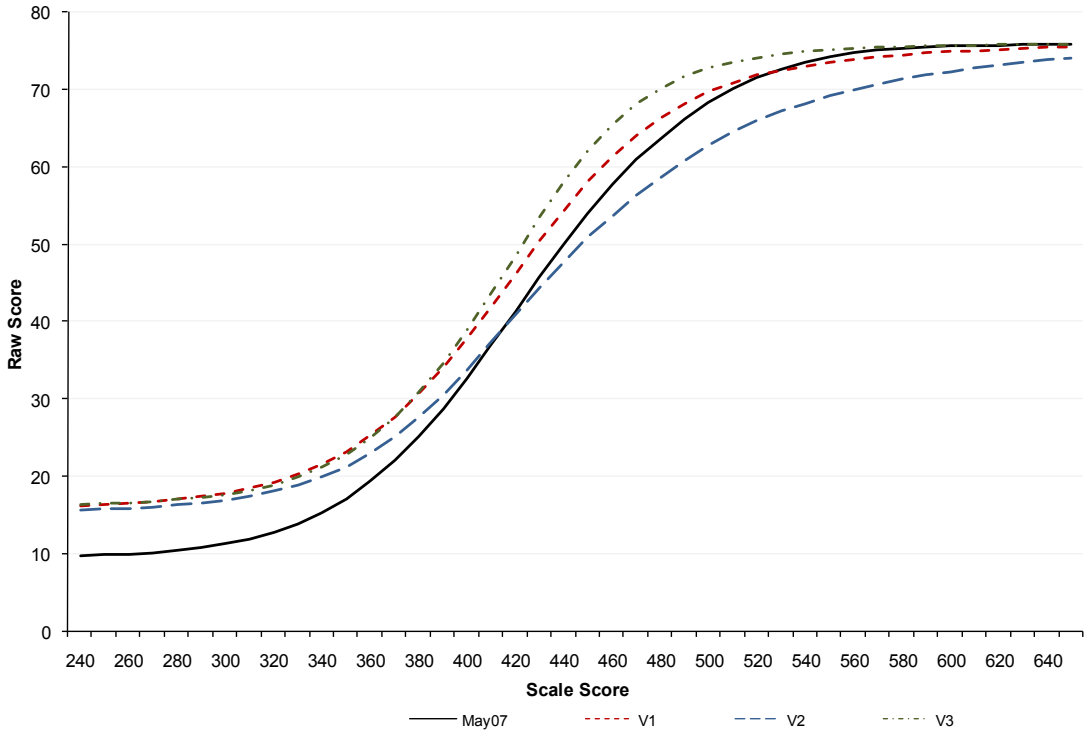


Figure 3.Test Characteristic Curves for English

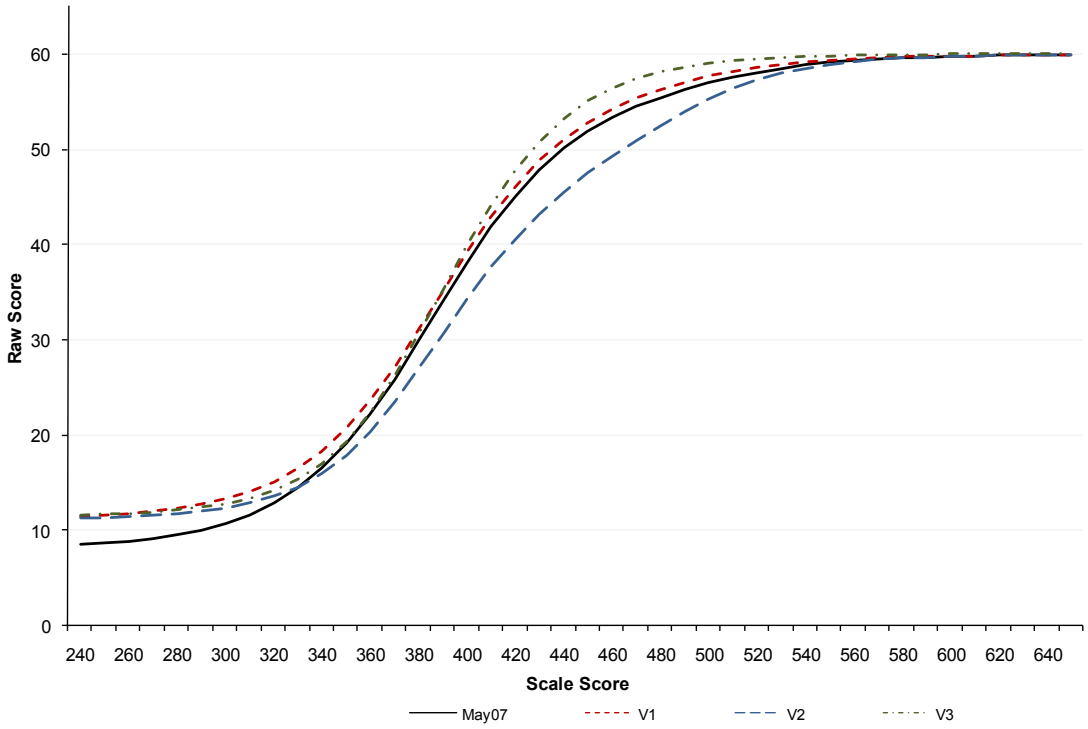
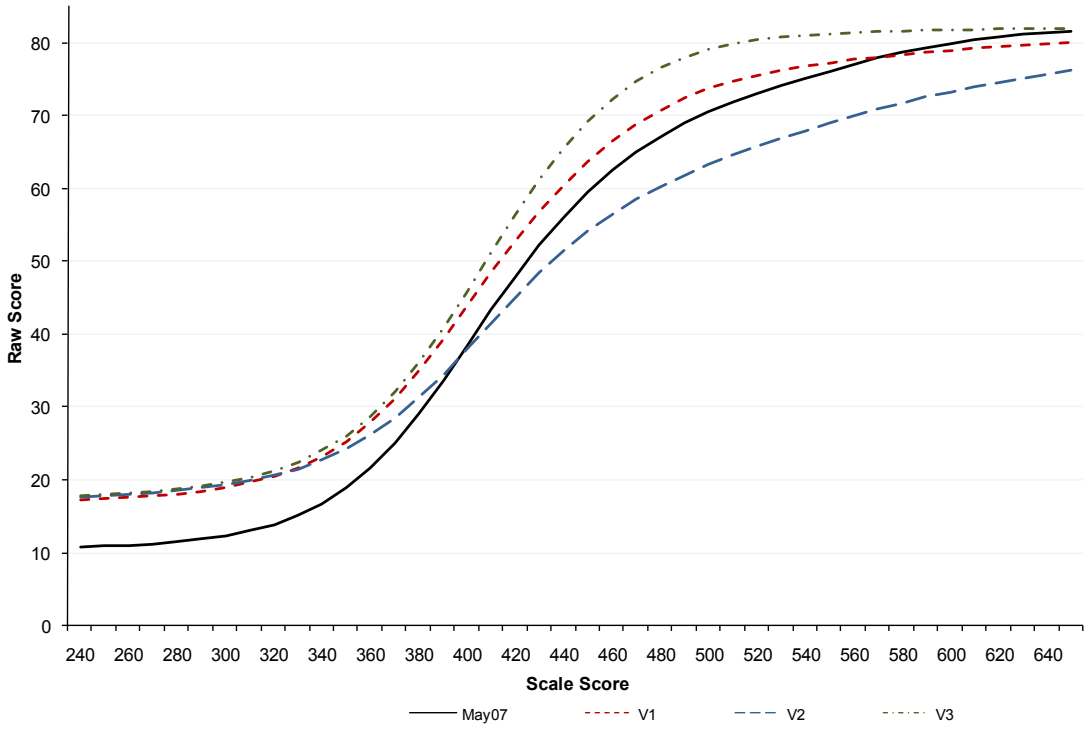


Figure 4.Test Characteristic Curves for Government



Gender and Ethnicity

We compared the performance of the different gender and ethnic groups using descriptive statistics and relative performance patterns on MC and CR items, which are shown in Tables 16-19. The pattern of mean scores for different races was essentially the same for the tests containing only MC and containing MC and CR items. Asian/Pacific Islander students always rank highest followed by White students. African-American students ranked lowest for all four MDHSA content areas.

Table 16. Summary Statistics by Ethnicity and Item Type: Algebra

	MC + GR + CR		MC		CR	
Ethnicity	Mean	SD	Mean	SD	Mean	SD
Asian/Pacific islander	38.19	9.32	19.65	4.66	13.46	4.77
White	35.62	9.96	18.54	4.94	12.27	4.79
American Indian	33.85	9.97	17.41	4.81	11.08	5.25
Hispanic	30.40	9.89	16.15	5.01	9.97	4.54
African American	26.91	10.42	14.38	5.16	8.19	4.71
Total	32.31	10.94	16.89	5.41	10.57	5.15

Table 17. Summary Statistics by Ethnicity and Item Type: English

	MC + CR		MC		CR	
Ethnicity	Mean	SD	Mean	SD	Mean	SD
Asian/Pacific islander	45.61	9.13	36.11	7.33	9.42	2.44
White	44.67	8.58	35.78	6.98	8.86	2.35
American Indian	42.59	8.14	34.14	7.05	8.48	1.90
Hispanic	38.75	9.93	30.75	8.344	7.92	2.38
African American	37.81	9.45	30.21	7.99	7.49	2.29
Total	41.80	9.65	33.41	7.97	8.30	2.44

Table 18. Summary Statistics by Ethnicity and Item Type: Biology

	MC + CR		MC		CR	
Ethnicity	Mean	SD	Mean	SD	Mean	SD
Asian/Pacific islander	47.56	11.78	35.35	8.09	12.06	4.40
White	44.06	11.62	33.34	8.16	10.67	4.29
American Indian	39.08	9.21	29.84	7.38	9.26	2.93
Hispanic	35.58	11.87	27.46	8.55	8.05	4.08
African American	33.52	11.04	26.17	8.04	7.20	3.68
Total	39.98	12.60	30.53	8.89	9.31	4.43

Table 19. Summary Statistics by Ethnicity and Item Type: Government

	MC + CR		MC		CR	
Ethnicity	Mean	SD	Mean	SD	Mean	SD
Asian/Pacific islander	53.83	13.50	37.23	8.89	16.52	5.53
White	50.57	13.79	35.60	9.43	14.83	5.48
Hispanic	42.85	13.51	30.18	9.32	12.55	5.12
American Indian	42.55	12.57	30.24	8.51	12.05	5.67
African American	39.62	13.28	28.17	9.10	11.20	5.22
Total	46.07	14.62	32.50	9.97	13.33	5.67

To investigate the interaction effect of ethnicity and item type, Table 20 was constructed to clearly present the results. Since white and African-American students were the two largest groups in the population, only these two groups are shown in the table. From Table 20, one can see that in every case, whites performed higher than blacks, but there is an interaction effect between item type and race across the four areas. Specifically, in the Algebra test the advantage that white students gained over black students by going from MC to CR items is 3.43. In the other three tests (English, Biology and Government), the advantage for white students decreased by 2.32, 2.55 and 3.52 percentage points respectively when going from MC items to CR items. An interaction effect was also found between gender and item type. The results are remarkably consistent across the four areas. The advantage that females gained over males by going from MC to CR items ranged from 3.20 to 6.26 percentage points and was a significant interaction effect.

Table 20. Interaction between Ethnicity and Item Type

		MC			CR			Difference b/t Ethnicity
Test	Ethnicity	# points	Mean	% of points	# points	Mean	% of points	
Algebra	White	26	18.54	71.31	21	12.27	58.43	-3.43
	African American		14.38	55.31		8.19	39.00	
	Difference			16.00			19.43	
English	White	46	35.78	77.78	14	8.86	63.29	2.32
	African American		30.21	65.67		7.49	53.50	
	Difference			12.11			9.79	
Biology	White	48	33.34	69.46	28	10.67	38.11	2.55
	African American		26.17	54.52		7.20	25.71	
	Difference			14.94			12.39	
Government	White	50	35.60	71.20	32	14.83	46.34	3.52
	African American		28.17	56.34		11.20	35.00	
	Difference			14.86			11.34	

Table 21. Interaction between Gender and Item Type

Test	Gender	MC			CR			Difference b/t Gender
		# points	Mean	% of points	# points	Mean	% of points	
Algebra	Male	26	17.06	65.61	21	10.86	51.71	4.63
	Female		16.71	64.27		11.55	55.00	
	Difference			1.34			-3.29	
English	Male	46	32.53	70.72	14	7.86	56.14	3.65
	Female		34.27	74.50		8.90	63.57	
	Difference			-3.78			-7.43	
Biology	Male	48	30.64	63.83	28	9.07	32.39	3.20
	Female		30.42	63.38		9.84	35.14	
	Difference			.45			-2.75	
Government	Male	50	32.88	65.76	32	12.82	40.06	6.26
	Female		32.11	64.22		14.33	44.78	
	Difference			1.54			-4.72	

Discussion and Summary

Not surprisingly, for all four tests, reliability decreased when the CR items were removed and there were fewer test items as a consequence. There were not large drops in reliability, but they were consistent. By employing the Spearman Brown Prophecy Formula, we observed that increasing the number of MC items (i.e., creating the test without CR items, but having the same number of points on the new test) countered the initial effect of decreasing reliabilities.

Although there was statistically significant improvement from the one-factor to the two-factor CFA model by looking at the chi-square difference, other fit indices were only slightly improved for Algebra and Biology tests. Adding the CR factor did not help reduce the standardized residuals, despite the significant test results. And the correlations between two factors for Algebra and Biology were .94 and .88, respectively, indicating the MC factor and CR factor represented almost the same construct. This could be due to the nature of Algebra and Biology subject matter, in that most of the items require quantitative abilities of the examinees, which are skills for which MC and CR items seem to be relatively equivalent.

For English and Government tests, we concluded that the MC and CR items measured different constructs to a greater degree. In both cases, the two-factor CFA models had much better fit indices than the one-factor models. Adding the CR factor effectively reduced the standardized residuals in magnitude. The inter-correlations between MC and CR factors were .74 and .83. This evidence supports the theory that there is multidimensionality of the mixed-format tests. It might be due to the characteristics of the tests, in which the items require the verbal knowledge and skills which CR items are designed to tap.

Many practitioners recognize the existence of multidimensionality but also realize the need to meet the unidimensionality assumption required by many psychometric models. In addition to the factor analytic methods used in this study, some researchers differentiate between traditional and essential dimensionality (Nandakumar, 1991) on a theoretical basis. Essential dimensionality counts only the dominant dimensions in the psychometric assessment of dimensionality. Statistical procedures to assess essential dimensionality have been developed and validated (Stout, 1987, 1990; Nandakumar, 1991, 1993). When test characteristic curves were compared for tests with and without CR items, differences were seen in the various TCCs depending on how replacement MC items were chosen. Overall, selecting MC items randomly from the bank to replace CR items created tests that were easier than those with the CR items. When all MC items were selected to match the *b*-value of the highest CR score point, for examinees of higher ability, the tests were more difficult than those with the CR items; for examinees of lower ability, the tests were easier. This is primarily due to the higher guessing parameters of the MC items at the lower end of the ability scale.

By observing the mean total score for different racial/ethnic groups when having only CR items, when having only MC items, and when having a mixture of both formats, we found that

the rank order in performance of the ethnic mean scores was essentially the same among the race/ethnic groups. Asian/Pacific Islander students always rank highest followed by white students. African-American students rank lowest in all four HSA content areas. However, when examining the interaction of the item format with race/ethnicity, which included only white and African-American ethnic students, we found that the performance gap between the groups is smaller in CR than MC items in English, Biology and Government, and larger in Algebra. It may be that Algebra CR items combine an apparently greater verbal emphasis along with the quantitative emphasis that appears to result in larger race/ethnicity effects. We also found an interaction between item format and gender. The higher level of performance for females relative to males was found for all four tests when moving from MC items only to tests including CR items, providing support for a theory that females' presumed greater verbal abilities result in higher scores on CR items.

Any claim for equivalence of the two item formats needs to be carefully examined. Even where the factor inter-correlation was .94, there are still important differences between MC and CR items. The two item types do not necessarily provide the same information or elicit the same expression of skills that may be especially important when the test is being used for diagnostic purposes. CR tests may provide diagnostic information that the MC tests do not (Birenbaum & Tatsuoka, 1987; Manhart, 1996). As Bennett (1991) pointed out, CR items serve to make visible to teachers and students cognitive behaviors that might be considered important to course mastery. Without CR items, instruction might only emphasize the tasks posed by MC items, and these tasks seem to be subtly different as evidenced by differences between black/white and male/female performance levels. In fact, the issue of "teaching to the test" when the test is primarily comprised of MC items, is being criticized in the call for new assessments as part of

the Race to the Top federal initative. Constructed response items may assess skills not equally assessed with the MC items. Particularly for Algebra, the CR items may require far more English language skills than the Algebra MC items. Such effects would be expected to increase as the number of CR items was increased.

Future Research

Overall, we found that content area may matter when investigating the differences in test results that are associated with the elimination of the CR items. The findings in this research are limited to tasks presented in these four domains, of course. Future research may be needed to see whether this finding is a special case for these tests or if it generalizes to other large-scale assessments characterized by their cognitive domain, as we expect. Research that depends upon items created specifically to test the mix of verbal and quantitative skills might permit teasing apart these differences and their effects. From very different work, Schafer, et al. (2001) found that educating teachers about CR item scoring rubrics led to increased student performance on CR items in Algebra and Biology tests, but not Government and English, suggesting again that differences in these item formats may be important for students and teachers. If the educational system were interested in these cognitive differences both from the standpoint of their instruction and their assessment, increased numbers of each item type would be necessary to obtain more reliable results. The work by Schafer and the work presented here may have implications for value added measurement since our research suggests that performance effects are sensitive to fine item distinctions such as those epitomized by CR and MC items.

References

Angoff, W. (1953). Test reliability and effective test length. *Psychometrika*, 18, 1-14.

Bedrova, E., & Leong, D. (1996). *Tools of the mind: The Vygotskian approach to early childhood education*. Englewood Cliffs, NJ: Prentice-Hall.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement, 14,* 151-162.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple choice items. *Journal of Education of Measurement, 28* (l), 77-92.

Bentler, P. M. (2004). *EQS structural equations program manual*. Los Angeles: BMDP Statistical Software.

Berk, L, & Winsler, A. (1995). *Scaffolding children's learning: Vygotsky and early childhood education*. Washington, DC: National Association for the Education of Young Children.

Birebaum, M. & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats – It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*, 385-395.

Bleske-Rechek, Zeug, N., & Webb, R. M. (2007). Discrepant performance on multiple-choice and short answer assessments and the relation of performance to general scholastic aptitude. *Assessment and Evaluation in Higher Education, 32* (2), 89-105.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple choice formats. *Journal of Educational Measurement,* 29,253-27 1.

Bridgeman, B. & Morgan, R. (1996) Success in college for students with discrepancies between performance on multiple-choice and essay tests, *Journal of Educational Psychology, 88*, 333–340.

Bridgeman, B. & Rock, D. A. (1993). Relationships among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement, 30* (4), 313-329.

Byrne, M. B. (2006). *Structural equation modeling with EQS: basic concepts, applications, and programming*. Hillsdale, NJ: Lawrence Erlbaum Associate.

Campbell, J. R. (1999). Cognitive processes elicited by multiple-choice and constructed-response questions on an assessment of reading comprehension. Doctoral Dissertation, Temple University. (UMI No. 9938651)

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory*. Fort Worth, TX: Holt, Rinehart and Winston, Inc.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H.I. Braun (Eds.), *Test Validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates.

Garner, M., & Engelhard, G. J. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12* (1), 29-51.

Center for K-12 Assessment. (2012). *Coming together to raise achievement: New assessments for the common core state standards*. Retrieved October 10, 2012 from http://www.k12center.org/rsc/pdf/20847_consortiaguide_sept2012.pdf

Cook, L. L., Dorans, N. J., & Eignor, D. R. (1988). An assessment of the dimensionality of three SAT-Verbal test edition *Journal of Educational Statistics, 13*, 19-43.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Belmont, CA: Wadsworth.

Halpern, D. F. (2004). A cognitive-process taxonomy for sex differences in cognitive abilities. *Current Directions in Psychological Science, 13*(4), 135-139.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Lawrence Erlbaum associates, NJ: Mahwah.

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *Journal of Experimental Education, 62*(2), 143-157.

Hu, L.- T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6, 1-55*.

Kennedy, P. & Walstad, W. B. (1997). Combining multiple-choice and constructed-response test scores: An economist's view. *Applied Measurement in Education, 10*(4), 359-375.

Kline, R. B. (2005). *Principle and practice of structural equation modeling*. New York: The Guilford Press.

Manhart, J. J. (1996). *Factor analytic methods for determining whether multiple-choice and constructed-response tests measure the same construct.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

Maryland State Department of Education (2008). *Maryland High School Assessments 2007 Technical Report: Algebra/Data Analysis, Biology, English, and Government*. Retrieved

January 20, 2009 from http://www.marylandpublicschools.org/NR/rdonlyres/B31FA6C8-B834-48BF-8B03-EEE757386ED5/17916/MDHSA_2007_TechnicalReport_FINAL_33108.pdf

Maryland State Department of Education (2009). *Maryland High School Assessments 2008 Technical Report: Algebra/Data Analysis, Biology, English, and Government*. Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (1993). A framework for studying differences between multiple-choice and free-response test items. In R. E. Bennett & W.C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 75-106)*. Hillsdale, NJ: Erlbaum.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement, 1*, 29-38.

Nandakumar, R. (1991). Traditional dimensionality vs. essential dimensionality. *Journal of Educational Measurement, 28,* 99-117.

Newstead, S. & Dennis, I. (1994) The reliability of exam marking in psychology: examiners examined, *Psychologist,* 7, 216–219.

Reiss, P. P. (2005). Causal models of item format and gender-related differences in performance on a large-scale mathematics assessment for grade three to grade ten. Dissertation. University of Hawaii.

Rodriguez, M.C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213-231). Mahwah, NJ: Erlbaum.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-

response items: a random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163-184.

Schafer, W. D., Swanson, G., Bené, N., & Newberry, G. (2001). Effects of teacher knowledge of rubrics on student achievement in four content areas. *Applied Measurement in Education*, 14, 151-170.

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

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52*, 589-617.

Snow, R. E. (1993). Construct validity and construed response tests. In R. E. Bennett, & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 45-60). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stiggins, R. J. (2005). *Student-involved assessment for learning* (4th ed.). Upper Saddle River: Pearson.

Thissen, D., Wainer, H., & Wang, X-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*, 113-123.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W.C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 75-106)*. Hillsdale, NJ: Erlbaum.

Wiggins, G. (1993). *Assessing student performance*. San Francisco: Jossey-Bass.

Willingham, W.W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Elbaum.

Note: Figures and tables below have been integrated into text.

Table 1. Participant Demographic Information

2007 HSA Form E		Algebra	English	Biology	Government
Total (counts)		13030	9263	9438	10491
Race (%)	White	47.8	48.7	50.7	47.6
	African American	38.9	38.6	36.4	39.3
	Hispanic	7.6	6.6	6.6	6.8
	Asian/Pacific Islander	5.4	5.8	6.0	5.9
	American Indian	0.3	0.3	0.3	0.4
Gender (%)	Male	51.2	49.4	51.0	51.2
	Female	48.8	50.6	49.0	48.8

Table 2. Subgroup Participant Demographic Information (Sample=2000)

2007 HSA Form E		Algebra	English	Biology	Government
Race (%)	White	48.2	49.3	50.8	48.7
	African American	38.7	38.1	35.9	38.2
	Hispanic	7.4	6.7	7.4	7.3
	Asian/Pacific Islander	5.9	6.0	6.0	5.9
Gender (%)	Male	51.8	49.6	51.4	51.9
	Female	48.2	50.5	48.7	48.1

Table 3. Algebra Blueprint

Reporting category	Number of items (points)		Total points
	MC(1pt)	CR	
Expectation 1.1 The student will analyze a wide variety of patterns and functional relationships using the language of mathematics and appropriate technology.	8	1 (4pt)	13
Expectation 1.2 The student will model and interpret real world situations, using the language of mathematics and appropriate technology.	10	1 (4pt)	14
Expectation 3.1 The student will collect, organize, analyze, and present data.	4	2 (3pt)	10
Expectation 3.2 The student will apply the basic concepts of statistics and probability to predict possible outcomes of real-world situations.	4	2(3&4pt)	10
Total counts	26	6	47 32

Table 4. Biology Blueprint

Reporting category	Number of items (points)		Total points
	MC(1pt)	CR(4pt)	
Goal 1 Skills and Processes of Biology	8	2	16
Expectation 3.1 Structure and Function of Biological Molecules	8	1	12
Expectation 3.2 Structure and Function of Cells and Organisms	9	1	13
Expectation 3.3 Inheritance of Traits	9	1	13
Expectation 3.4 Mechanism of Evolutionary Change	5	1	9
Expectation 3.5 Interdependence of Organisms in the Biosphere	9	1	13
Total counts	48	7	76 55

Table 5. English Blueprint

Reporting category	Number of items (points)		Total points
	MC(1pt)	CR	
1: Reading and Literature: Comprehension and Interpretation	13	1(3pt)	16
2: Reading and Literature: Making Connections and Evaluation	11	1(3pt)	14
3: Writing - Composing	8	2(4pt)	16
4: Language Usage and Conventions	14	0	14
Total counts	46	4	60 50

Table 6. Government Blueprint

Reporting category	Number of items (points)		Total points
	MC(1pt)	CR(4pt)	
Expectation1.1 The student will demonstrate understanding of the structure and functions of government and politics in the United States	13	3	25
Expectation 1.2 The student will evaluate how the United States government has maintained a balance between protecting rights and maintaining order.	11	2	19
Goal 2 The student will demonstrate an understanding of the history, diversity, and commonality of the peoples of the nation and world, the reality of human interdependence, and the need for global cooperation, through a perspective that is both historical and multicultural.	8	1	12
Goal 3 The student will demonstrate an understanding of geographic concepts and processes to examine the role of culture, technology, and the environment in the location and distribution of human activities throughout history.	7	1	11
Goal 4 The student will demonstrate an understanding of the historical development and current status of economic principles, institutions, and processes needed to be effective citizens, consumers, and workers.	11	1	15
Total counts	50	8	58 / 82

Table 7. Internal Consistency Reliability of Tests Scores With and Without Constructed Response Items

Content Area	Reliability& SEM	Test with CR	Test without CR	New Lengthened Test without CR
Algebra	Coefficient Alpha	.91	.88	.93
	SEM	3.37	2.28	----
English	Coefficient Alpha	.90	.88	.91
	SEM	3.03	2.71	----
Biology	Coefficient Alpha	.93	.89	.93
	SEM	3.45	2.93	----
Government	Coefficient Alpha	.94	.91	.95
	SEM	3.61	2.94	----

Table 8. Confirmatory Factor Analysis Results: Algebra Data

	Fit index					
Model (N=2000)	Chi-square/df	NFI	NNFI	CFI	RMSEA(90% CI)	AIC
Two-factor	510.42/43	.97	.97	.97	.07 (.07-.08)	424.42
One-factor	595.48/44	.97	.96	.97	.08 (.07-.08)	507.48
Null	17571.68/55	--	--	--	--	17461.68

Table 9. Loadings of MC Item Parcels and CR Items for Algebra

	Two-factor model		One-factor model
	MC factor	CR factor	General factor
MC parcel 1	.80	--	.78
MC parcel 2	.77	--	.76
MC parcel 3	.74	--	.73
MC parcel 4	.79	--	.77
MC parcel 5	.77	--	.76
CR 1	--	.84	.83
CR 2	--	.63	.62
CR 3	--	.80	.78
CR 4	--	.66	.62
CR 5	--	.63	.63
CR 6	--	.76	.73

Table 10. Confirmatory Factor Analysis Results: Biology Data

	Fit index					
Model (N=2000)	Chi-square/df	NFI	NNFI	CFI	RMSEA(90% CI)	AIC
Two-factor	425.05/64	.98	.98	.98	.05 (.05-.06)	297.05
One-factor	656.71/65	.98	.98	.98	.07 (.06-.07)	526.74
Null	28577.81/78	--	--	--	--	28421.81

Table 11. Loadings of MC Item Parcels and CR Items for Biology

	Two-factor model		One-factor model
	MC factor	CR factor	General factor
MC parcel 1	.83	--	.78
MC parcel 2	.74	--	.71
MC parcel 3	.80	--	.76
MC parcel 4	.71	--	.68
MC parcel 5	.82	--	.78
MC parcel 6	.72	--	.69
CR 1	--	.72	.70
CR 2	--	.81	.80
CR 3	--	.82	.81
CR 4	--	.84	.84
CR 5	--	.83	.80
CR 6	--	.75	.73
CR 7	--	.83	.80

Table 12. Confirmatory Factor Analysis Results: English Data

	Fit index					
Model (N=2000)	Chi-square/df	NFI	NNFI	CFI	RMSEA(90% CI)	AIC
Two-factor	406.37/64	.97	.97	.97	.05 (.05-.06)	278.37
One-factor	993.77/65	.94	.91	.93	.09 (.08-.09)	863.77
Null	13034.69/78	--	--	--	--	12878.69

Table 13. Loadings of MC Item Parcels and CR Items for English

	Two-factor model		One-factor model
	MC factor	CR factor	General factor
MC parcel 1	.61	--	.60
MC parcel 2	.68	--	.67
MC parcel 3	.66	--	.65
MC parcel 4	.76	--	.75
MC parcel 5	.71	--	.70
MC parcel 6	.71		.70
MC parcel 7	.72		.70
MC parcel 8	.72		.71
MC parcel 9	.74		.71
CR 1	--	.70	.57
CR 2	--	.82	.68
CR 3	--	.78	.64
CR 4	--	.79	.65

Table 14. Confirmatory Factor Analysis Results: Government Data

	Fit index					
Model (N=2000)	Chi-square/df	NFI	NNFI	CFI	RMSEA(90% CI)	AIC
Two-factor	318.10/64	.99	.99	.99	.05 (.04-.05)	190.10
One-factor	1141.26/65	.97	.97	.97	.09 (.09-.10)	1011.26
Null	39471.73/78	--	--	--	--	39315.73

Table 15. Loadings of MC Item Parcels and CR Items for Government

	Two-factor model		One-factor model
	MC factor	CR factor	General factor
MC parcel 1	.86	--	.79
MC parcel 2	.86	--	.77
MC parcel 3	.83	--	.75
MC parcel 4	.82	--	.75
MC parcel 5	.80	--	.74
CR 1	--	.81	.80
CR 2	--	.83	.82
CR 3	--	.85	.83
CR 4	--	.83	.82
CR 5	--	.82	.83
CR 6	--	.80	.79
CR 7	--	.85	.81
CR 8	--	.85	.82

Table 16. Summary Statistics by Ethnicity and Item Type: Algebra

	MC + GR + CR		MC		CR	
Ethnicity	Mean	SD	Mean	SD	Mean	SD
Asian/Pacific islander	38.19	9.32	19.65	4.66	13.46	4.77
White	35.62	9.96	18.54	4.94	12.27	4.79
American Indian	33.85	9.97	17.41	4.81	11.08	5.25
Hispanic	30.40	9.89	16.15	5.01	9.97	4.54
African American	26.91	10.42	14.38	5.16	8.19	4.71
Total	32.31	10.94	16.89	5.41	10.57	5.15

Table 17. Summary Statistics by Ethnicity and Item Type: English

	MC + CR		MC		CR	
Ethnicity	Mean	SD	Mean	SD	Mean	SD
Asian/Pacific islander	45.61	9.13	36.11	7.33	9.42	2.44
White	44.67	8.58	35.78	6.98	8.86	2.35
American Indian	42.59	8.14	34.14	7.05	8.48	1.90
Hispanic	38.75	9.93	30.75	8.344	7.92	2.38
African American	37.81	9.45	30.21	7.99	7.49	2.29
Total	41.80	9.65	33.41	7.97	8.30	2.44

Table 18. Summary Statistics by Ethnicity and Item Type: Biology

	MC + CR		MC		CR	
Ethnicity	Mean	SD	Mean	SD	Mean	SD
Asian/Pacific islander	47.56	11.78	35.35	8.09	12.06	4.40
White	44.06	11.62	33.34	8.16	10.67	4.29
American Indian	39.08	9.21	29.84	7.38	9.26	2.93
Hispanic	35.58	11.87	27.46	8.55	8.05	4.08
African American	33.52	11.04	26.17	8.04	7.20	3.68
Total	39.98	12.60	30.53	8.89	9.31	4.43

Table 19. Summary Statistics by Ethnicity and Item Type: Government

	MC + CR		MC		CR	
Ethnicity	Mean	SD	Mean	SD	Mean	SD
Asian/Pacific islander	53.83	13.50	37.23	8.89	16.52	5.53
White	50.57	13.79	35.60	9.43	14.83	5.48
Hispanic	42.85	13.51	30.18	9.32	12.55	5.12
American Indian	42.55	12.57	30.24	8.51	12.05	5.67
African American	39.62	13.28	28.17	9.10	11.20	5.22
Total	46.07	14.62	32.50	9.97	13.33	5.67

Table 20. Interaction between Ethnicity and Item Type

		MC			CR			Difference b/t Ethnicity
Test	Ethnicity	# points	Mean	% of points	# points	Mean	% of points	
Algebra	White	26	18.54	71.31	21	12.27	58.43	-3.43
	African American		14.38	55.31		8.19	39.00	
	Difference			16.00			19.43	
English	White	46	35.78	77.78	14	8.86	63.29	2.32
	African American		30.21	65.67		7.49	53.50	
	Difference			12.11			9.79	
Biology	White	48	33.34	69.46	28	10.67	38.11	2.55
	African American		26.17	54.52		7.20	25.71	
	Difference			14.94			12.39	
Government	White	50	35.60	71.20	32	14.83	46.34	3.52
	African American		28.17	56.34		11.20	35.00	
	Difference			14.86			11.34	

Table 21. Interaction between Gender and Item Type

Test	Gender	MC			CR			Difference b/t Gender
		# points	Mean	% of points	# points	Mean	% of points	
Algebra	Male	26	17.06	65.61	21	10.86	51.71	4.63
	Female		16.71	64.27		11.55	55.00	
	Difference			1.34			-3.29	
English	Male	46	32.53	70.72	14	7.86	56.14	3.65
	Female		34.27	74.50		8.90	63.57	
	Difference			-3.78			-7.43	
Biology	Male	48	30.64	63.83	28	9.07	32.39	3.20
	Female		30.42	63.38		9.84	35.14	
	Difference			.45			-2.75	
Government	Male	50	32.88	65.76	32	12.82	40.06	6.26
	Female		32.11	64.22		14.33	44.78	
	Difference			1.54			-4.72	

Figure 1.Test Characteristic Curves for Algebra

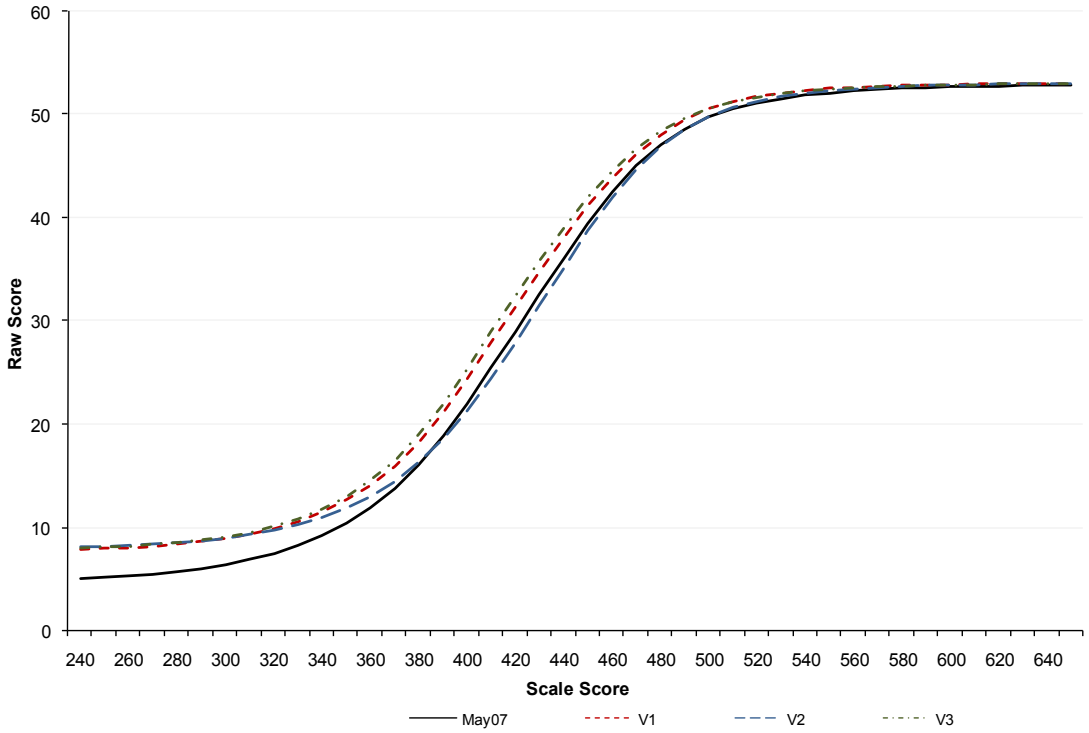


Figure 2.Test Characteristic Curves for Biology

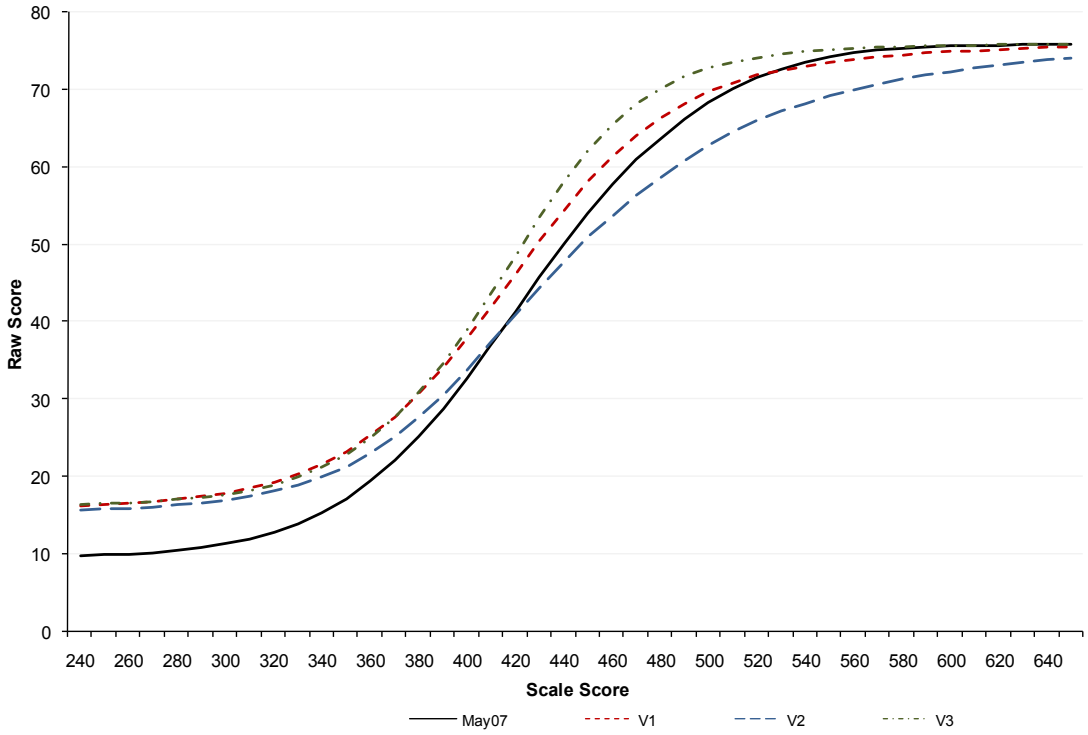


Figure 3.Test Characteristic Curves for English

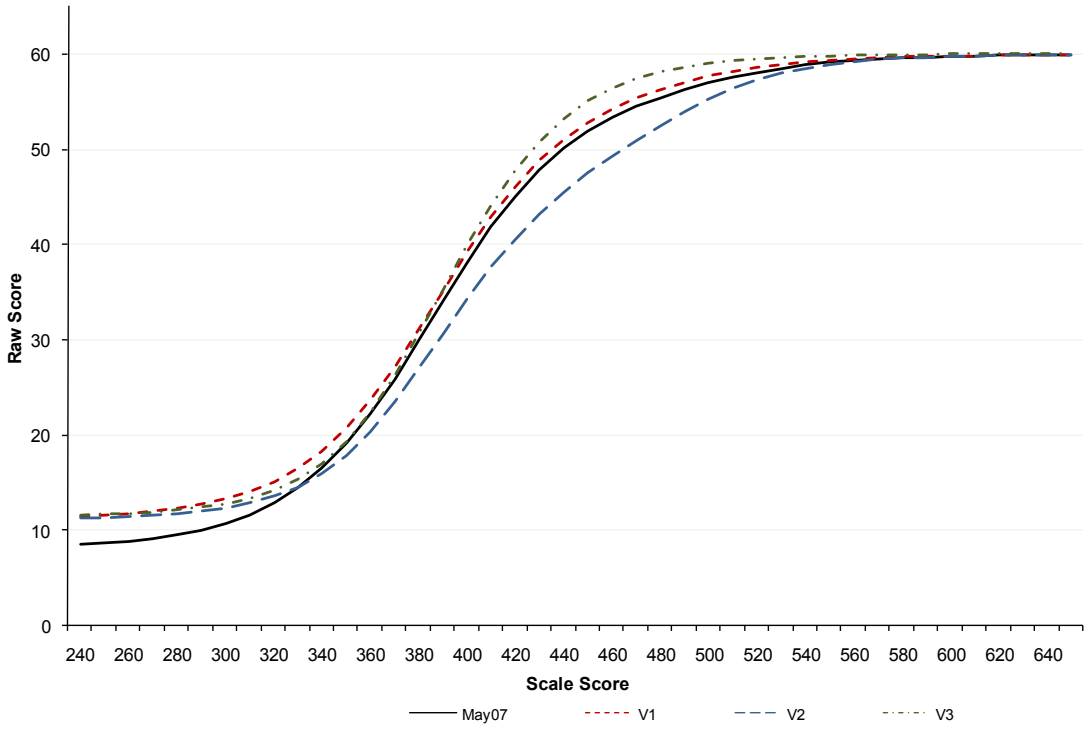


Figure 4.Test Characteristic Curves for Government