Multiple Choice Items and Constructed Response Items: Does It Matter? [1]

Robert W. Lissitz Xiaodong Hou

University of Maryland

Introduction

Both multiple choice (MC) items and constructed response (CR) items are widely used in large-scale educational testing. MC items are usually considered to have a relatively reliable scoring procedure and therefore are fairer to examinees. Because they can be answered relatively quickly by the examinees, then a wider portion of the domain can be assessed with more item coverage. 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 in the schools with feedback on performance in a timelier manner. All these elements make MC items very attractive.

However, many practitioners argued that MC items fail to elicit the higher levels of cognitive processing and that MC items engaged examinees in a type of guessing game (Campbell, 1999). It is believed by some researchers that MC items have an inability to tap higher order thinking and allows for a higher probability of guessing correctly which causes lower reliabilities in the test for lower ability students (Cronbach, 1988). Nevertheless, more and more measurement literature suggests that MC items can measure essentially the same thing as do CR items (Kennedy & Walstad, 1997). Hancock (1994)

pointed out that proponents of the MC format believe that the MC items can be written to tap complex thinking though it is more difficult to write such MC items than CR items.

In contrast, many practitioners think that CR items can elicit the constructive cognitive processes that they believe are important. CR items allow a range of answers, all of which are provided by the examinee. They are more effective in reducing the probability of correct guessing because the correct answer is not shown in a CR item, although the truth of this depends upon having an accurate scoring system. In addition, CR items may directly show what the examinees think since they require that the examinee must construct a response in their own words. Therefore, in many large-scale tests, the CR items are still retained in spite of relatively expensive score reporting due to the time and cost of scoring CR items, as well as inevitable subjectivity in scoring.

In order to answer the question whether it matters if CR items are completely removed from assessments and only MC items remain, we need to know whether MC items are able to assess the same traits from examinees as the ones that CR items assess. There have been a number of researchers investigating MC and CR items and their trait-equivalence, but the evidence is inconclusive (Martinez, 1999, Rodriguez, 2003).

Among the methods to assess the trait-equivalence of MC and CR item formats, there are two broad approaches in the literature. The first approach uses stem-equivalent items in both formats, which means that MC and CR items employ the same stem to control content differences and isolate the format effect (Ackerman & Smith, 1988; Frisbie & Cantor, 1995). In the second approach, items in both formats have independent stems, tapping similar or different domain or cognitive ability (Rodriguez, 2003). The format correlations or corrected correlations are often reported in the literature. A high

correlation indicates trait equivalence and a low correlation indicates the two formats may examine different traits or constructs. Besides the correlation method, factor analysis is also employed to study the trait equivalence of the two formats. For example, several studies reported that MC items loaded on one factor and CR items loaded on a separate factor in confirmatory factor analysis (Bennett et al., 1991; Bridgeman & Rock, 1993; Rodriguez, 2003).

In our research, we have tried to use each of these methods to investigate the impact of removing CR items from the Maryland High School Assessments (HSA). The correlations between total combined, MC only and CR only test forms are compared for each content test. The reliabilities are also examined for each test. What's more, gender-related and race-related item format differences are investigated in our analysis. Finally, we have looked at the loading of each item format in a principal components analysis.

Research questions

An analysis of the technical implications of removing CR items from Maryland's High School Assessment testing program (HSA) is conducted for the four HSA assessments in algebra/data analysis, English II, biology, and government.

Specifically, in our research we are trying to answer the following research questions for each content area:

1. What are the correlations between total scores with and without CR items?
2. What are the correlations between the total scores without CR items (all MC items) and the scores for only CR items?

3. Are there differences in the means of total scores and standard deviations among different ethnic groups when there are only CR items, when there are no CR items, and when there is a mixture of both?
4. Are there differences in the means of total scores and standard deviations across gender when we have a mixture of CR and MC items and when we remove CR items?
5. How does the reliability change for a test when CR items are removed? Are the Standard Errors of Measurement (SEM) different?
6. Do the CR items and MC items load on the same components in principal components analysis?
7. Is there any difference between the different HSA tests in regards to their sensitivity to inclusion or exclusion of CR items?

Instruments

2007 HSAs are end-of-course tests and consist of Algebra, Biology, English, and Government. The tests are composed of MC items and CR items. Algebra tests have student-produced response items or "gridded" response (GR) items which require students to grid in correct responses on the answer document. Since it can be graded by machine, GR items are also considered a form of MC item in this analysis. MC items are machine-scored and CR items are scored by raters.

Participants

The analysis is conducted on the results from the 2007 HSA form E which has one of the largest examinee populations in the Algebra, Biology, English, and Government areas. Table A below shows the information of participants on race and gender. In

algebra, there are 13,030 examinees, of which 51.2% were male and 48.8% female. Algebra examinees were 47.8 % white, 38.9% are African American, 7.6% Hispanic, 5.4% Asian/Pacific islander, and 0.3% is American Indian. In the English test, there are 9263 examinees of which 49.4% were male and 50.6% female. The English examinees were 48.7% white, 38.6% African American, 6.6% Hispanic, 5.8% Asian/Pacific islander, and 0.3% is American Indian. In the Biology test, there are 9438 examinees. Male constituted 51.0% of the examinees, and females were the remaining 49.0%. The percentage of examinees who were white, African American, Hispanic, Asian/Pacific islander and American Indian are 50.7%, 36.4%, 6.6%, 6.0%, and 0.3%, respectively. In the Government test, there are 10491 examinees. 51.2% are male and 48.8% are female. The percentage of examinees of white, African American, Hispanic, Asian/Pacific islander and American Indian are 47.6%, 39.3%, 6.8%, 5.9%, and 0.4% respectively.

Table A. Participants' information on race and gender

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

Results

2007 HSA form E of algebra, English, biology and government tests were analyzed to investigate the implications of removing CR items from the tests. Omit responses and drop responses were considered as missing values and number-right scoring was used in the analysis.

2007 Maryland HSA Algebra

The correlation between the total score of the test containing only MC and GR items with no CR items and the total score of the test containing MC, GR and CR items is .961. The correlation between the total score of the test containing only CR and GR items and the total score of the test containing only CR items is .796. The correlation between the total score of the test containing all MC, GR and CR items and the total score of the test containing only CR items is .932.

Table1. Correlation between MC scores, CR scores and total scores

AlgE07	**Correlation**
Corr (MC+GR, CR)	.796
Corr (MC+GR, Total)	.961
Corr (CR, Total)	.932

The reliability of the test decreases when CR items are removed from the test (i.e. only MC and GR items remained in the test). The reliability of the test containing MC, GR and CR items is .905 with unconditional SEM equal to 3.37, and the reliability of the test containing only MC and GR items is .879 with unconditional SEM of 2.28. 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,

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

is employed to calculate reliability for the new test into which some items are hypothesized to be added so that the points of these items added are equal to the number of points lost from dropping the CR items, where $\rho_{jj'}$ is the reliability of the test without CR items, and k is the factor by which the length of the test without CR items must be increased. The new reliability for the new lengthened algebra test is .934. We can see that the reliability is slightly higher than the original test. Therefore, we believe that increasing the number of MC items (at least making the test without CR items have the same number of points with the test with CR items) may counter the effect of decreasing reliabilities by removing CR items in the algebra test.

Table2. Reliability comparison when with CR and without CR

AlgE07	with CR	without CR
Reliability (Cronbach's Alpha)	.905	.879
Unconditional standard error of measurement	3.37	2.28

The descriptive statistics are shown in table 3. We can see from figures 1, 2 and 3 that the pattern of mean scores for different races remains essentially unchanged for the tests containing only MC and GR, or containing MC, GR and CR items. Asian/pacific islander always ranks highest followed by White. And African American always ranks lowest.

Table3. Mean and SD for Algebra form E07

	MC + GR + CR		MC+GR		CR	
Ethnicity	Mean	SD	Mean	SD	Mean	SD
Asian/Pacific islander	38.19	9.32	24.37	5.41	13.46	4.77
White	35.62	9.96	23.00	5.92	12.27	4.79
American Indian	33.85	9.97	21.82	6.05	11.08	5.25
Hispanic	30.40	9.89	20.00	6.16	9.97	4.54
African American	26.91	10.42	18.05	6.44	8.19	4.71
Total	32.31	10.94	21.11	6.56	10.57	5.15

Figure1. Mean score plot for different ethnicity

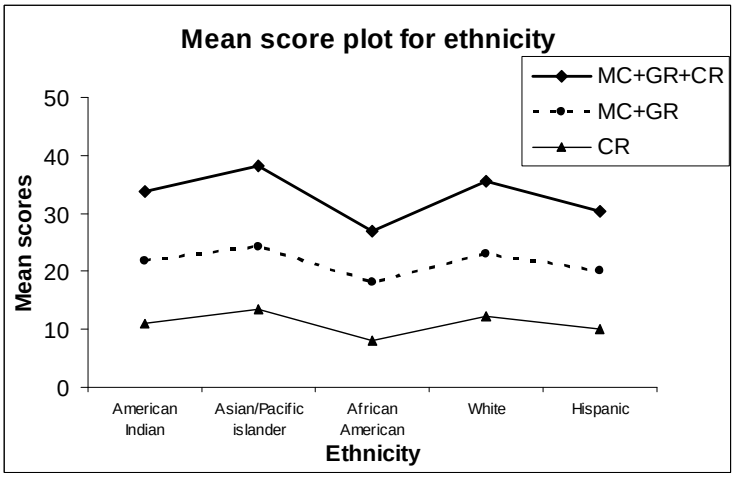


Figure 2. Boxplots for MC+GR+CR test scores by Ethnicity

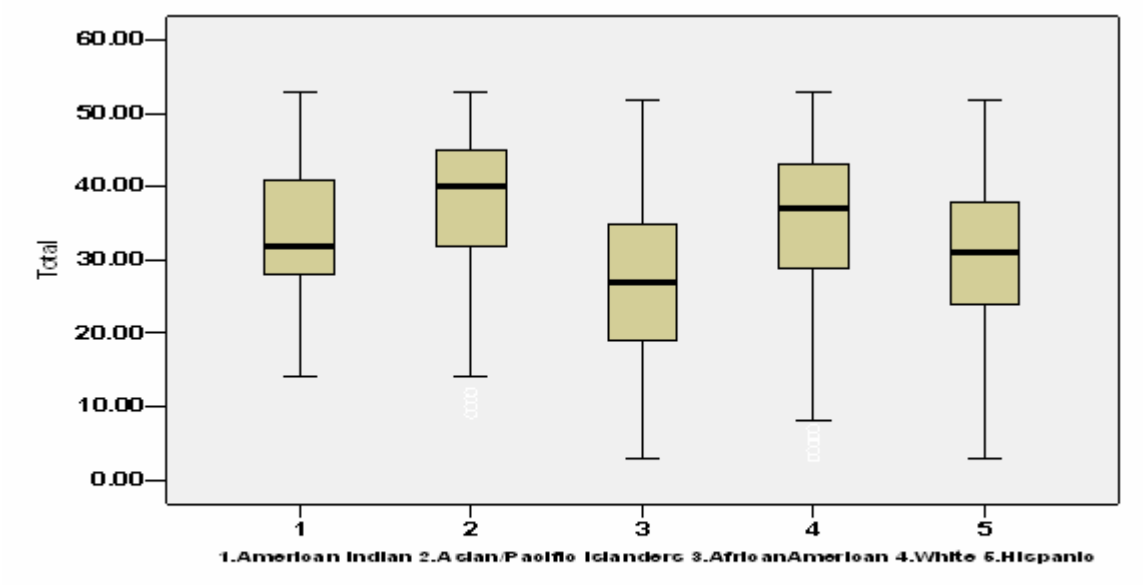
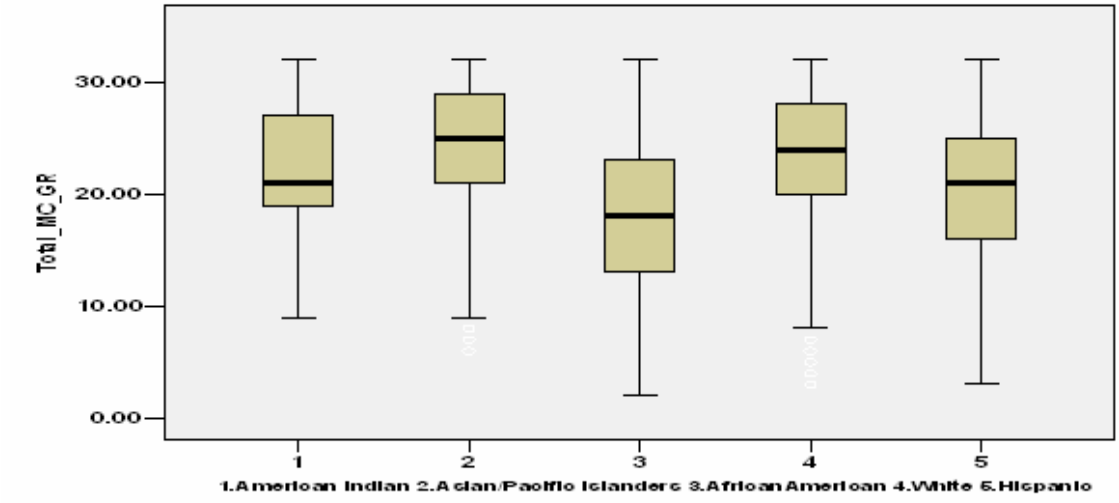


Figure 3. Boxplots for only MC and GR test scores by Ethnicity

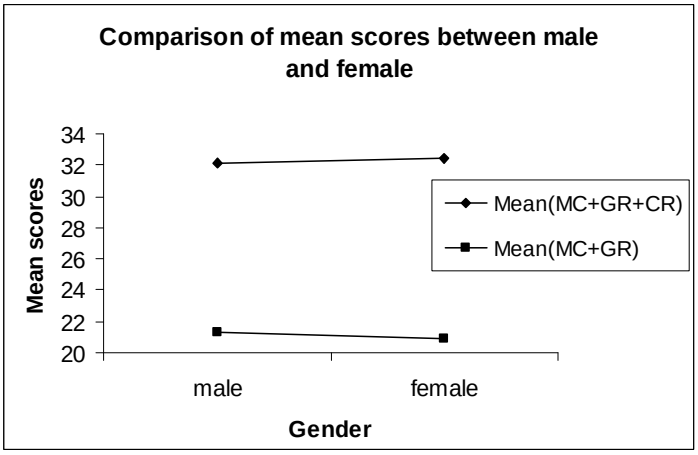


Females and males perform equally when the test contains CR, GR and CR items. When CR items are removed, males tend to perform slightly better than females in the Algebra test.

Table 4. Gender difference in mean scores with and without CR items

	Mean_MC+GR+CR(SD)	T test	Mean_MC+GR(SD)	t test
Male	32.17 (11.15)	t = -1.33, P>.05	21.31(6.64)	t= 3.17, P<.05
Female	32.46 (10.73)		20.91(6.46)	

Figure 4. Comparison of mean scores in gender with and without CR



In algebra test, principal component analysis was conducted. Four factors were extracted by principal component analysis based on a scree plot, and varimax rotation was used. No clear evidence emerged that CR items are loading on the same factor (see Appendix A for results).

2007 Maryland HSA English

The correlation between the total score of the test containing only MC items with removing all CR items and the total score of the test containing both MC and CR items is .981. The correlation between the total score of the test containing only CR items and the total score of the test containing both MC and CR items are .765. The correlation

between the total score of the test containing only MC items and the total score of the test containing only CR items is .625.

Table5. Correlation between MC scores, CR scores and total scores

EngE07	**Correlation**
Corr(MC, CR)	.625
Corr(MC,Total)	.981
Corr(CR,Total)	.765

The reliability of the test decreases when CR items are removed from the test and only MC items remain. The reliability of the test containing both CR and MC items is .901 with unconditional SEM equaling 3.03, and the reliability of the test containing only MC items is .884 with unconditional SEM equaling 2.71. 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, Spearman Brown Prophecy Formula is employed to calculate reliability for the new test into which some items are hypothesized to be added so that the points of these items added are equal to the number of points lost from dropping the CR items. The new reliability for the new lengthened English test is .909. We can see that the reliability is slightly higher than the original test. Therefore, we believe that increasing the number of MC items (at least making the test without CR items have the same number of points with the test with CR items) may counter the effect of decreasing reliabilities by removing CR items in the English test.

Table6. Reliability comparison when with CR and without CR

Eng E07	with CR	without CR
Reliability (Cronbach's Alpha)	.901	.884
Unconditional standard error of measurement	3.03	2.71

The descriptive statistics are shown in table 7. We can see from the figure 5, 6 and 7 that the pattern of mean scores for different races remains unchanged when CR items are removed. Asian/pacific islander always ranks highest closely followed by White, and African Americans always ranks lowest.

Table 7. Mean and SD for English form E07

	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

Figure5. Mean score plot for different ethnicity

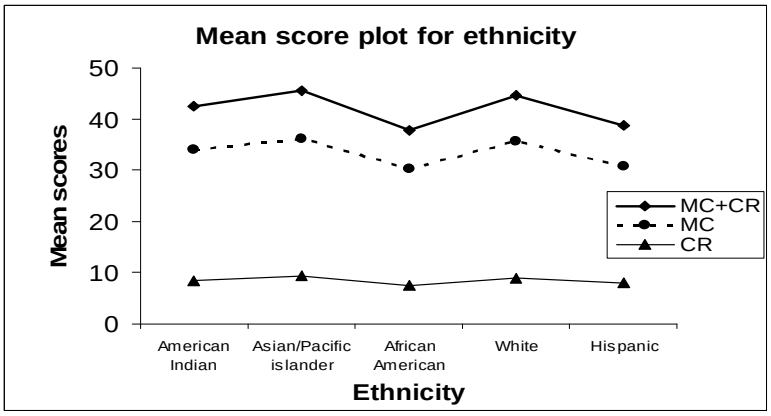


Figure6. Boxplot for MC+CR scores for Ethnicity

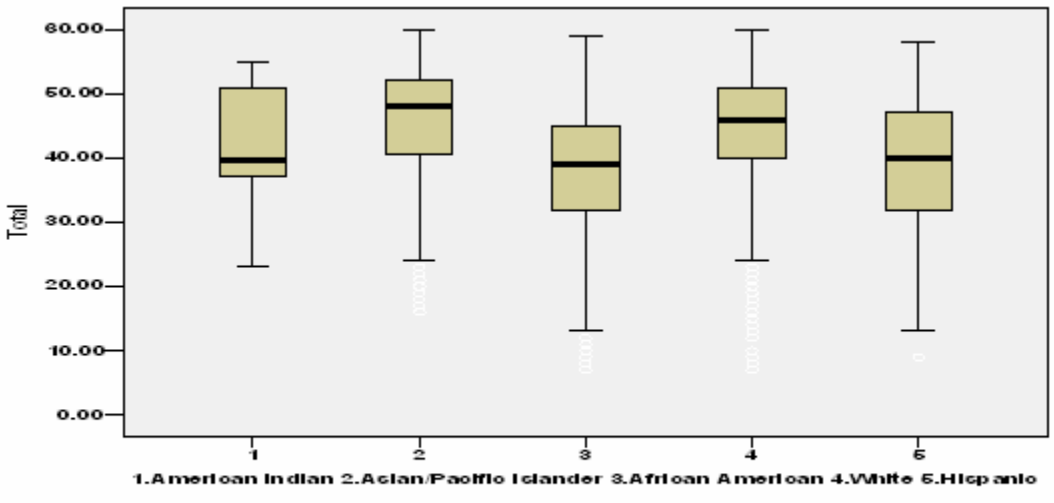
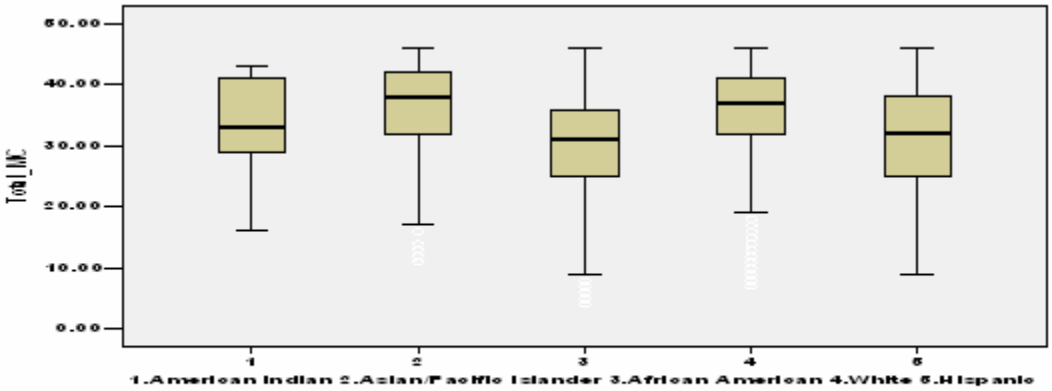


Figure 7.Boxplot for only MC scores for Ethnicity

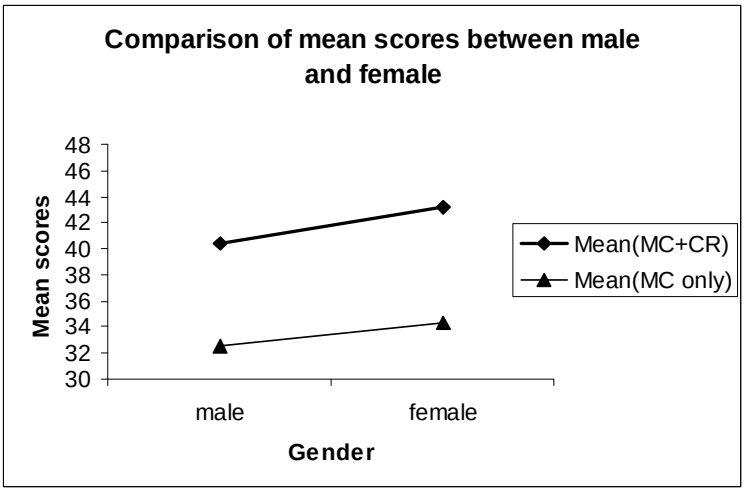


Female have higher mean scores than males whether CR items are included in the test or not.

Table 8. Gender difference in mean scores with and without CR items

	Mean_MC+CR (SD)	t test	Mean_MC (SD)	t test
Male	40.39 (9.93)	t = - 13.53	32.53(8.22)	t = -10.16
Female	43.17 (9.16)	P<.05	34.27(7.62)	P<.05

Figure 8. Comparison of mean scores in gender with and without CR



What's more, principal component analysis was conducted. Three factors were extracted based on a scree plot in principal component analysis, and varimax rotation was used. All CR items tend to load highly on the 3rd factor and no MC items load highly on this factor (See appendix B for the results).

2007 Maryland HSA Biology

The correlation between the total score of the test containing only MC items with all CR items removed and the total score of the test containing both MC and CR items is .976. The correlation between the total score of the test containing only CR items and the total score of the test containing both MC and CR items is .896. The correlation between the total score of the test containing only MC items and the total score of the test containing only CR items is .777.

Table 9. Correlation between MC scores, CR scores and total scores

Bio E07	**Correlation**
Corr(MC, CR)	.777
Corr(MC, Total)	**.976**
Corr(CR, Total)	.896

The reliability of the test decreases when CR items are removed from the test and only MC items remain in the test. The reliability of the test containing both CR and MC items is .925 with unconditional SEM equaling 3.45, and the reliability of the test containing only MC items is .891 with unconditional SEM equaling 2.93. 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, Spearman Brown

Prophecy Formula is employed to calculate reliability for the new test into which some items are hypothesized to be added so that the points of these items added are equal to the number of points lost from dropping the CR items. The new reliability for the new lengthened biology test is .928. We can see that the reliability is slightly higher than the original test. Therefore, we believe that increasing the number of MC items (at least making the test without CR items have the same number of points with the test with CR items) may counter the effect of decreasing reliabilities by removing CR items in the biology test.

Table10. Reliability comparison when with CR and without CR

Bio E07	with CR	without CR
Reliability (Cronbach's Alpha)	.925	.891
Unconditional standard error of measurement	3.45	2.93

The descriptive statistics are shown in table 11. We can see from figures 9, 10 and 11 that the pattern of mean scores for different races remains unchanged when CR are removed. Asian/pacific islander always ranks highest followed by White. And African American always ranks lowest.

Table 11. Mean and SD for Biology form E07

	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

Figure 9. Mean score plot for different ethnicity

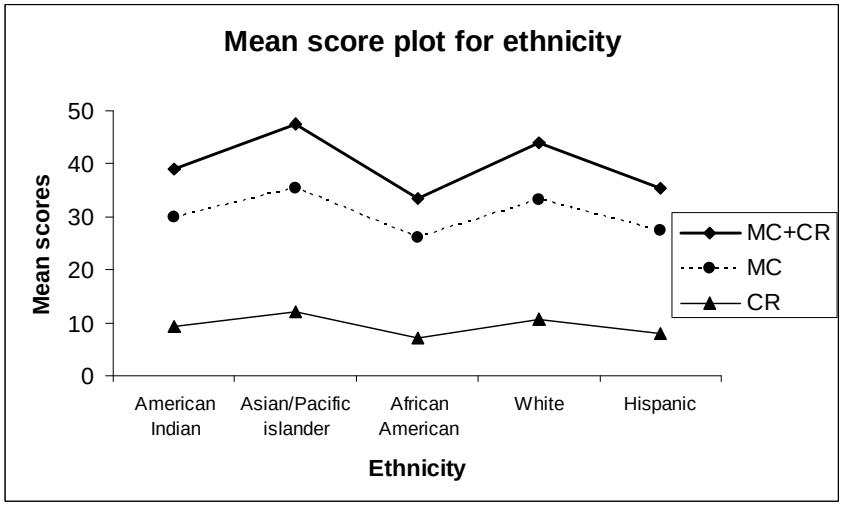


Figure 10. Boxplot for MC+CR scores for Ethnicity

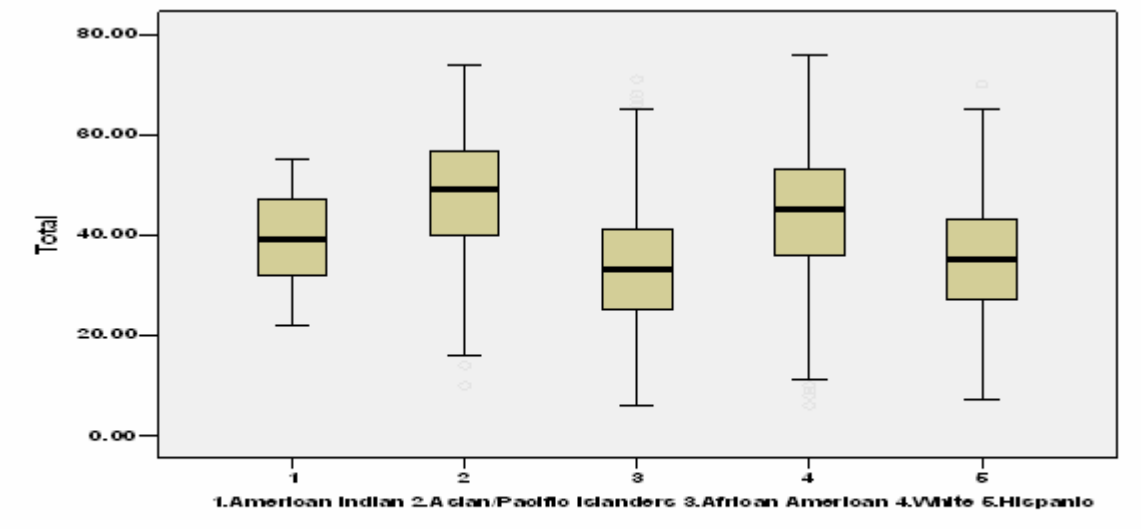
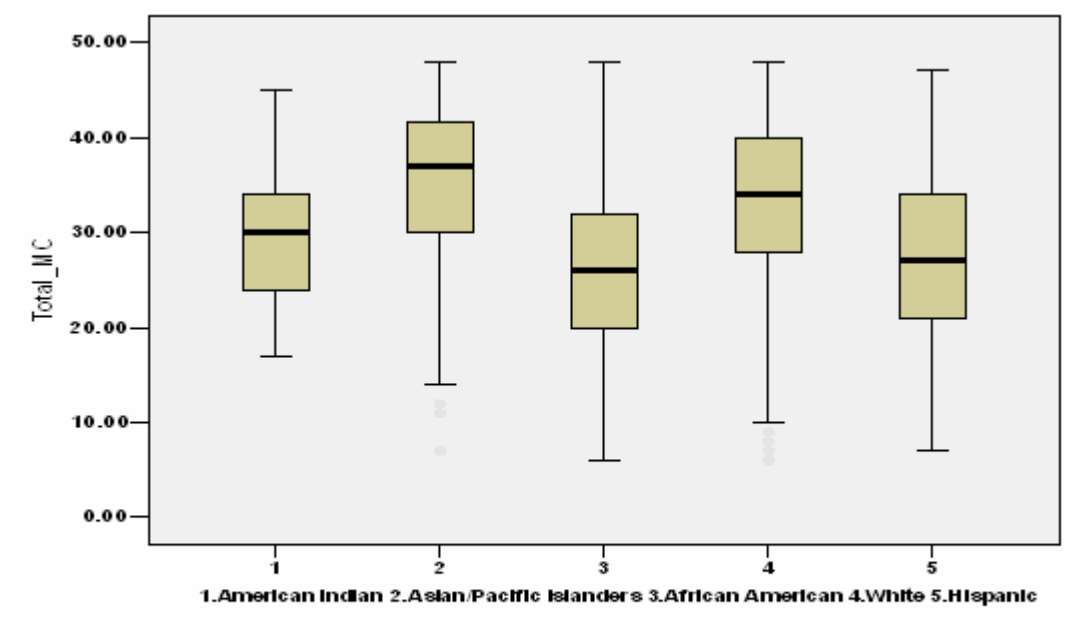


Figure 11.Boxplot for only MC scores for

ethnicity

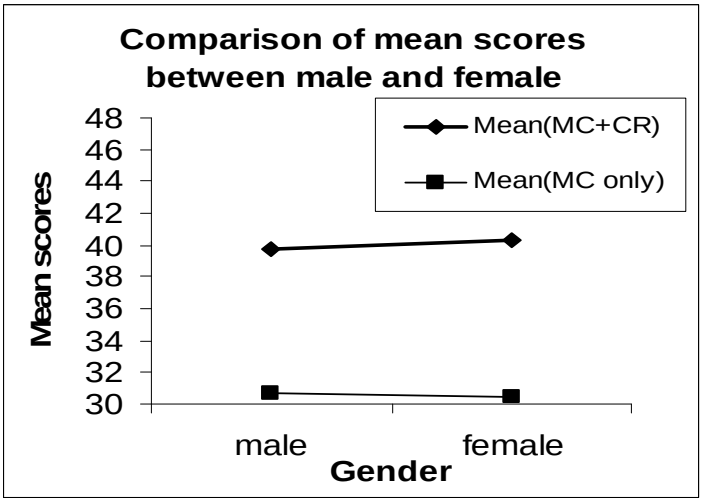


There are significant differences in mean scores between male and female when CR items are included in the test. However, there is no significant difference in mean scores between male and female when CR items are removed in the test. When CR items are included, the mean score for males is lower than the mean for females. When CR items are removed, the mean scores for males and female are almost the sames.

Table 12. Gender difference in mean scores with and without CR items

	Mean_MC+CR (SD)	t test	Mean_MC (SD)	t test
Male	39.71 (12.98)	t = -2.05	30.64(9.30)	t = 1.15
Female	40.26 (12.21)	P<.05	30.42(8.44)	P>.05

Figure 12. Comparison of mean scores in gender with and without CR



In the Biology test, principal component analysis was conducted. Two factors were extracted based on a scree plot in principal component analysis, and varimax rotation was used. All CR items tend to load highly on the 1st factor but there are several MC items which also load highly on this factor (See Appendix C for these results).

2007 Maryland HSA Government

The correlation between the total score of the test containing only MC items with CR items removed and the total score of the test containing both MC and CR items is .968. The correlation between the total score of the test containing only CR items and the total score of the test containing both MC and CR items is .894. The correlation between the total score of the test containing only MC items and the total score of the test containing only CR items is .753.

Table13. Correlation between MC scores, CR scores and total scores

GovE07	**Correlation**
Corr(MC, CR)	.753
Corr(MC, Total)	.968
Corr(CR, Total)	.894

The reliability of the test reduces when CR items are removed from the test and only MC items remain. The reliability of the test containing both CR and MC items is .939 with unconditional SEM equaling 3.61, and the reliability of the test containing only MC items is .913 with unconditional SEM equaling 2.94. 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, Spearman Brown Prophecy Formula is employed to calculate reliability for the new test into which some items are hypothesized to be added so that the points of these items added are equal to the number of points lost from dropping the CR items. The new reliability for the new lengthened government test

is .945. We can see that the reliability is slightly higher than the original test. Therefore, we believe that increasing the number of MC items (at least making the test without CR items have the same number of points with the test with CR items) may counter the effect of decreasing reliabilities by removing CR items in the government test.

Table 14. Reliability comparison when with CR and without CR

GovE07	with CR	without CR
Reliability (Cronbach's Alpha)	.939	.913
Unconditional standard error of measurement	3.61	2.94

The descriptive statistics are shown in table 15. We can see from the figure 13, 14 and 15 that the pattern of mean scores for different races remains unchanged when CR items are removed except Hispanic and American Indian. Asian/pacific islander always ranks highest followed by White. And African American always ranks lowest.

Table 15. Mean and SD for Gov form E07

	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

Figure 13. Mean score plot for different ethnicity

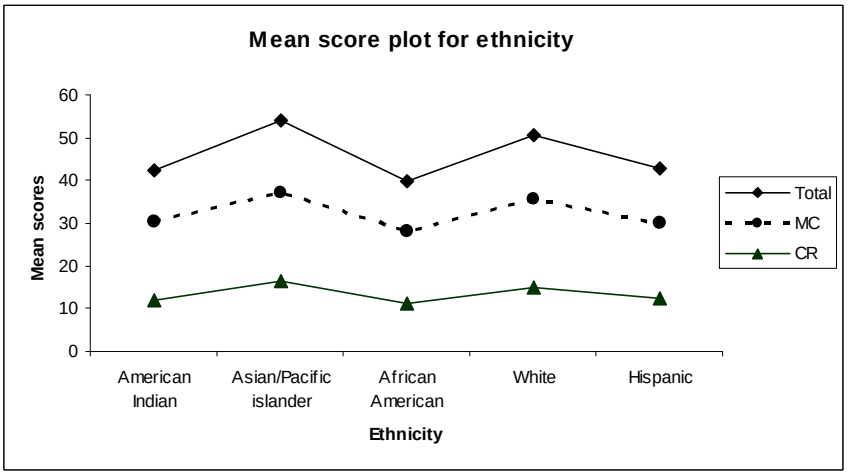


Figure 14. Boxplot for MC+CR scores for Ethnicity

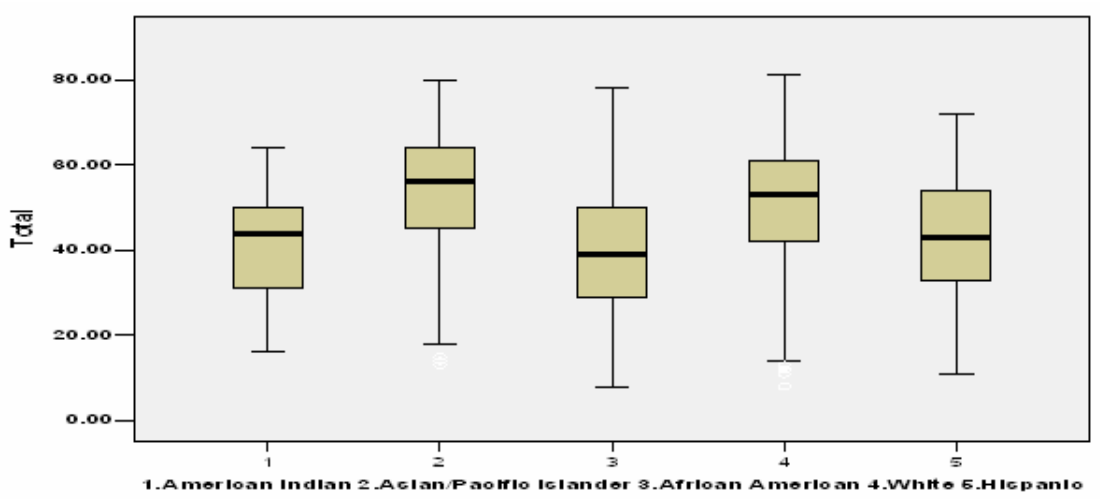
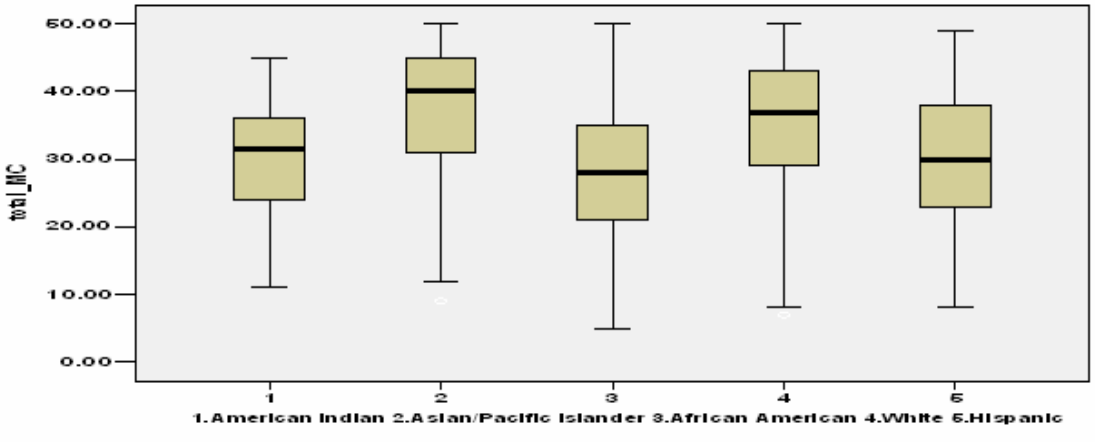


Figure 15.Boxplot for only MC scores for Ethnicity

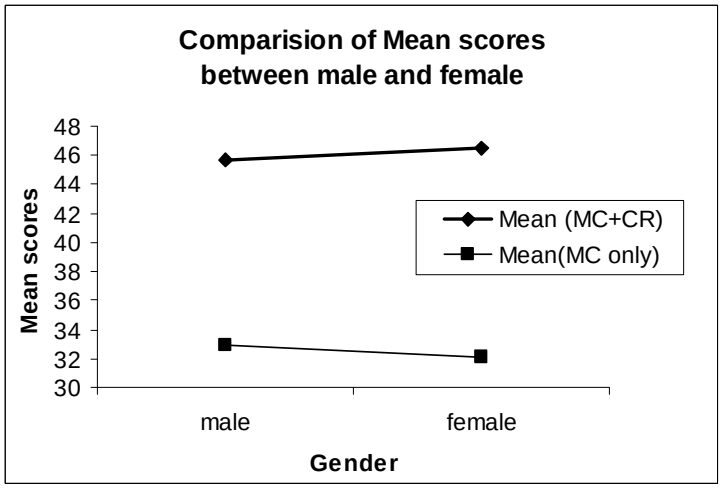


There are significant differences in mean scores between males and females whether CR items are included in the test or not. When CR items are included, the mean score for males is lower than the mean for females. When CR items are removed the mean score for males is higher than for females, though the mean difference in gender is not as much as that when CR items are included in the test. Figure 16 shows the pattern of mean change for both genders.

Table 16. Gender difference in mean scores with and without CR items

	Mean_MC+CR (SD)	t test	Mean_MC (SD)	t test
Male	45.70 (14.91)	t = -2.52	32.88(10.32)	t = 3.80
Female	46.44 (14.31)	P<.05	32.11(9.58)	P<.05

Figure 16. Comparison of mean scores in gender with and without CR



In the government test, principal component analysis was conducted. Three factors were extracted based on a scree plot in principal component analysis, and varimax rotation was used. All CR items tended to load highly on the 3rd factor and no other MC items loaded highly on this factor. See Appendix D for these results.

Discussion

The correlations between the total scores with CR items and the total scores without CR items for Algebra, Government, Biology, and English tests are relatively high, at .961, .968, .976 and .981, respectively. These results suggest that dropping the CR items will not do any great harm to the conclusions that result from the test without CR items. The correlations between the total scores without CR items and the total scores for only CR items for English, Government, Biology, and Algebra tests are .625, .753, .777 and .796 respectively.

By observing the means of total scores 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 patterns, or the rank order of the scores, of the different racial/ethnic groups remains essentially unchanged for the three situations above with similar patterns of standard deviations. Asian/pacific islander always ranks highest in the means of the total scores, followed by White, and African American always rank lowest. So, again, dropping CR items will make no meaningful difference to these results.

For the government test, males tend to have a lower score than females when the test included both CR and MC items. However, when CR items are removed, males tend to have a slightly higher score than females. For the algebra test, when the test has both CR and MC items, the mean scores across gender groups are not different. However, when CR items are removed, males tend to have a slightly higher score than females. For the Biology test, males have a lower score on average than females. But when CR items are removed, there is now no mean score difference across gender groups. In the English test, females perform better than males regardless of whether the test contains CR items or not. But this score difference across gender groups decreases when CR items are removed. Based on the observations of average scores for the four tests, males tend to benefit more from MC items and females tend to benefit more from CR items. In the case of gender group differences, there is some advantage to females for retaining the CR items. Although the difference is not numerically large, the sample size for these statistical tests is large enough to detect significant effects.

For all four tests, the reliabilities (Cronbach's Alpha) decrease when the CR items are removed. The Cronbach's Alpha for the government test drops from .939 to .913, the

Alpha for the Algebra test drops from .905 to .879, the Alpha for the Biology test drops from .925 to .891 and the Alpha for the English test drops from .901 to .884. These are not large drops in reliability, although they are consistent. In order to examine whether increasing the number of MC items would counter the effect, the Spearman Brown Prophecy Formula was employed to calculate reliabilities for the four new tests into which some items are hypothesized to be added so that the points of these items added are equal to the number of points lost from dropping the CR items. The new reliabilities for the four new lengthened tests are .934, .909, .928, and .945 for Algebra, English, Biology, and Government, respectively. We can see that the reliabilities for the four new tests are slightly higher than the original tests. Therefore, we believe 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) will counter the effect of decreasing reliabilities by removing CR items.

In addition, principal component analyses with varimax rotation were conducted for all four tests. The CR items in the Biology test tends to load highly on one component with several MC items also loading at a relatively high level on this factor. The CR items in the English and Government tests tend to load highly on one component without any MC items loading at a high level on this factor. No clear pattern emerged with the components analysis for Algebra/Data Analysis HSA examination. These results, along with the effect on gender differences, seems to indicate that CR items assessed a somewhat different latent trait from the MC items in three of the tests (English, Government, and Biology). It should be noted, though, that the pattern observed for

gender differences does not exactly match that observed for the principal components analysis and the observation of separated loadings.

In summary, there are some differences in the test results that are associated with the elimination of the CR items, but these differences are not large. Overall, the effect of eliminating CR items does not appear to be critical for the State agenda.

Note 1. We would like to thank the Maryland State Department of Education for supporting the Maryland Assessment Research Center for Education Success in doing the work reported here.

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement,12*(2), 117-128.

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

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

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)

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

Frisbie, D. A,, & Cantor, N. K. (1995). The validity of scores from alternative methods of assessing spelling achievement. *Journal of Educational Measurement,* 32(I), 55-78.

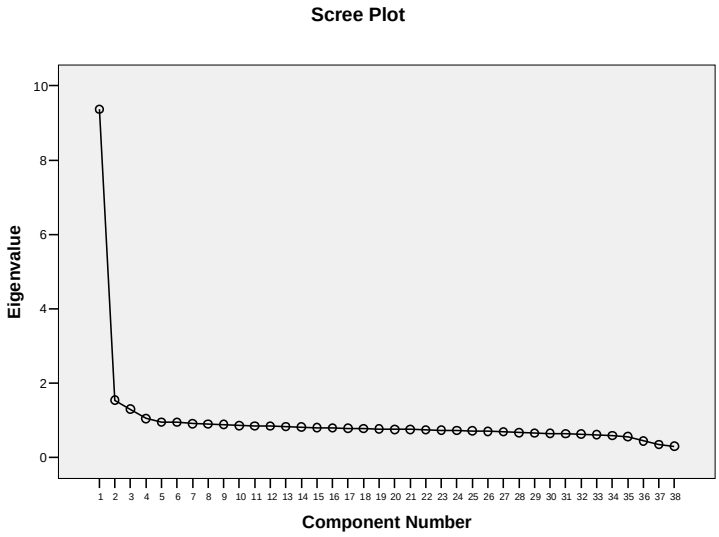
Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *Journal of Experimental Education, 62*(2). 143-157.

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.

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

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.

Appendix A. Principal component analysis for 2007 HSA Algebra Form E



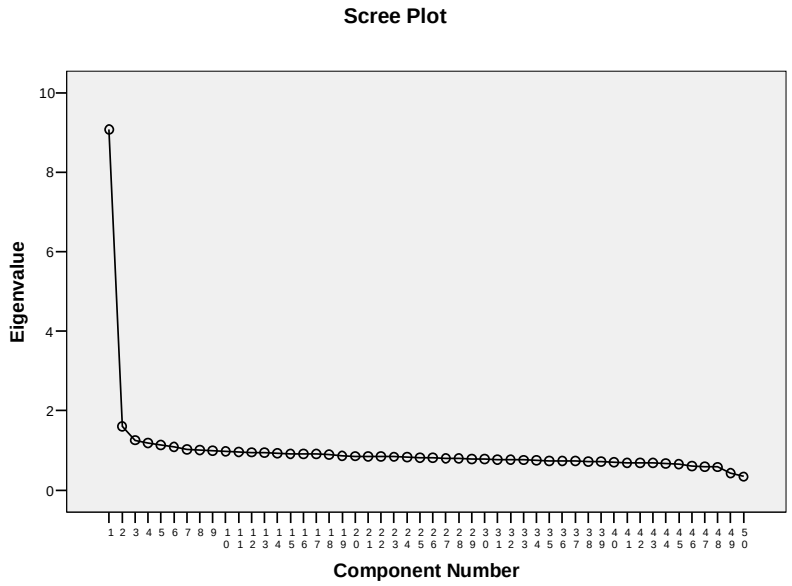
Rotated Component Matrix(a)

	Component			
	1	2	3	4
MC1	.038	.485	.125	.085
MC2	.155	.250	.458	.130
MC3	.438	-.092	.057	.134
MC4	.167	.461	.105	.078
MC5	.234	.084	.639	.139
MC7	.365	.178	.097	.070
MC8	.255	.239	.381	.073
MC9	.155	.089	.193	.340
MC10	.357	.144	.034	.232
MC13	.525	.358	.145	-.025
MC14	.510	.199	.092	.056
MC15	.511	.064	.147	.119
MC16	.139	.127	.627	.172
MC17	.281	.396	.180	-.019
MC18	-.010	.549	.079	.118
MC19	.055	.097	.720	.142
MC20	.362	.282	.230	.053
MC23	.451	.201	.347	.224
MC24	.461	.301	.106	-.042
MC25	.429	.239	.212	.065
MC27	.425	-.107	.017	.276
MC28	.410	.300	.193	.050

MC29	.383	.258	.270	.124
MC30	.157	.258	.120	-.096
MC31	.455	.196	.245	-.004
MC32	.448	-.037	.051	.220
GR2	.140	.547	.012	.078
GR3	.298	.282	.218	-.038
GR4	.134	.572	.116	.160
GR5	.353	.570	.186	.052
GR6	.411	.184	.198	.169
GR7	.019	.517	.080	.218
CR1	.496	.463	.330	.271
CR2	.479	.129	.074	.403
CR3	.402	.304	.377	.411
CR4	.139	.207	.231	.765
CR5	.486	.275	.084	.265
CR6	.297	.223	.241	.713

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 12 iterations.

Appendix B. Principal component analysis for 2007 HSA English Form E



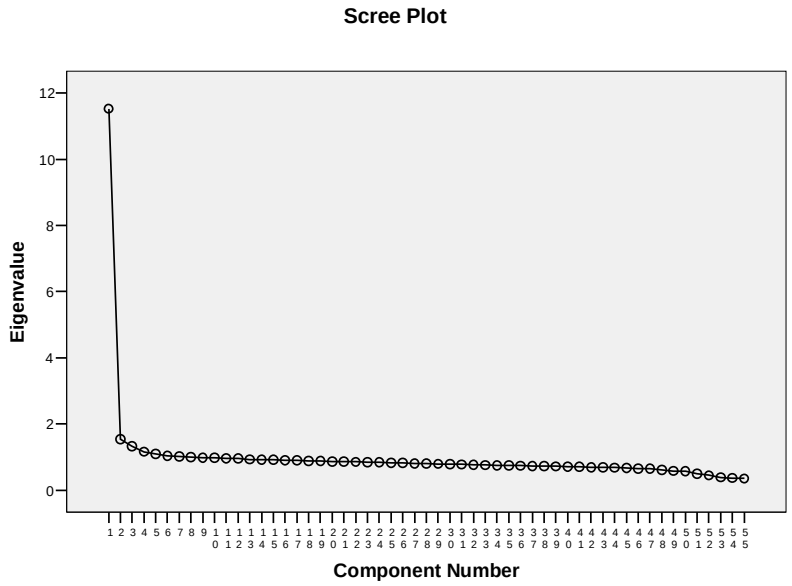
Rotated Component Matrix(a)

	Component		
	1	2	3
MC1	.244	.092	.106
MC3	.165	.123	.146
MC4	.186	.266	.174
MC5	.032	.588	.185
MC6	.492	-.040	.103
MC7	.445	.121	.167
MC8	.540	.093	.155
MC9	.318	.287	.144
MC10	.432	.099	.046
MC11	.289	.231	.076
MC12	.465	.098	.038
MC13	.605	.025	.122
MC19	.372	.198	.122
MC20	.393	.280	.060
MC21	.430	.228	-.001
MC22	.445	.295	.117
MC23	.378	.344	.072
MC24	.185	.197	.207
MC25	.143	.091	.116
MC26	.444	.174	.145
MC27	.073	.558	.120
MC28	.066	.460	.110
MC29	.174	.438	.116
MC30	.336	.199	.159

MC31	.400	.225	.082
MC32	.417	.217	.093
MC33	.325	.156	.126
MC34	.183	.364	.113
MC47	.103	.068	.235
MC48	.320	.190	.108
MC49	.406	.158	.176
MC50	.374	.199	.197
MC51	.256	.133	.258
MC52	.303	.057	.250
MC53	.374	.197	.202
MC54	.265	.318	.104
MC55	.067	.141	.114
MC56	.340	.376	.143
MC57	.190	.454	.125
MC58	.101	.368	.148
MC59	.142	.344	.062
MC60	.238	.451	.087
MC61	.326	.363	.123
MC62	.197	.447	.106
MC65	.371	.338	.143
MC66	.084	.481	.123
CR1	.103	.188	.778
CR2	.325	.315	.489
CR3	.141	.231	.792
CR4	.184	.287	.683

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 6 iterations.

Appendix C. Principal component analysis for 2007 HSA Biology Form E



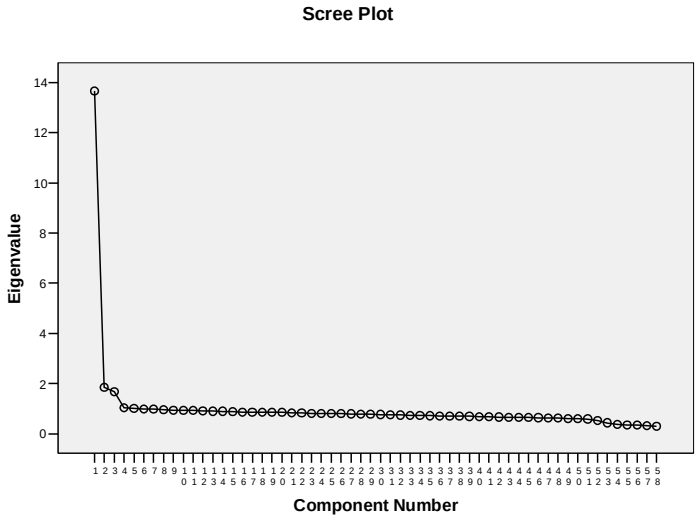
Rotated Component Matrix(a)

	Component	
	1	2
MC1	.313	.156
MC2	.316	.272
MC3	-.063	.215
MC4	.450	.149
MC5	.272	.207
MC6	.479	.040
MC7	.349	.193
MC8	.136	.105
MC13	.277	.163
MC14	.256	.498
MC15	.106	.138
MC16	.200	.389
MC17	.270	.308
MC18	.477	.315
MC19	.528	.226
MC20	.537	.069
MC21	.088	.396
MC22	.194	.476
MC26	.177	.245
MC27	.259	.274
MC28	.340	.289
MC29	.417	.268
MC30	.187	.506
MC31	.243	.530
MC32	.348	.312
MC33	.369	.330
MC37	.304	-.060

MC38	.394	.391
MC39	.178	.414
MC40	.214	.424
MC41	.305	.113
MC42	.283	.390
MC43	.304	.180
MC44	.306	.427
MC45	.203	.438
MC46	.371	.159
MC48	.319	.242
MC49	.120	.167
MC53	.067	.482
MC54	.227	.490
MC55	.455	.024
MC56	.260	.510
MC57	.377	.117
MC58	.019	.199
MC59	.123	.530
MC60	.213	.434
MC61	.385	.207
MC62	.301	.397
CR_1	.546	.307
CR_2	.631	.400
CR_3	.615	.404
CR_4	.687	.329
CR_5	.684	.271
CR_6	.575	.342
CR_7	.682	.279

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 3 iterations.

Appendix D. Principal component analysis for 2007 HSA Government Form E



Rotated Component Matrix(a)

	Component		
	1	2	3
MC1	.199	.198	.130
MC2	.211	.304	.007
MC3	.248	.524	.153
MC4	.498	.076	.219
MC5	.428	.232	.087
MC6	.295	.263	.134
MC7	.340	.156	.126
MC10	.391	.129	.078
MC11	.028	.412	.108
MC12	.284	.463	.197
MC13	.252	.383	.111
MC14	.409	.204	.146
MC15	.476	.334	.144
MC18	.527	.032	.156
MC19	.386	.222	.174
MC20	.460	.287	.173
MC21	.605	.137	.169
MC22	.356	.400	.204
MC23	.322	.440	.168
MC26	.331	.149	.050
MC27	.284	.386	.183
MC28	.391	.309	.039
MC29	.409	-.009	.191
MC30	.477	.285	.120
MC31	.381	.209	.142
MC32	.081	.555	.132
MC33	.118	.361	.139
MC34	.363	.107	.110
MC35	.375	.220	.201

MC36	.318	.026	.109
MC37	.329	.215	.064
MC38	.194	.363	.136
MC41	.228	.251	.107
MC42	-.075	.252	.004
MC43	.489	.157	.135
MC46	.151	.548	.169
MC47	.219	.482	.134
MC48	.144	.352	.111
MC49	.428	.328	.158
MC50	.315	.287	.198
MC51	.050	.558	.200
MC52	.543	.009	.162
MC53	.378	.237	.144
MC56	.067	.394	.091
MC57	.571	-.004	.171
MC58	.206	.170	.147
MC59	.181	.136	.074
MC60	.275	.490	.200
MC61	.548	.102	.178
MC62	.461	.259	.136
CR1	.297	.281	.660
CR2	.357	.288	.656
CR3	.318	.291	.693
CR4	.381	.285	.623
CR5	.405	.277	.624
CR6	.275	.268	.702
CR7	.223	.260	.764
CR8	.257	.264	.746

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 6 iterations.