Psychometric Analyses of the 2006 MCAS High School Science and Technology/Engineering Tests[1,2]

Ronald Hambleton, Yue Zhao, Zachary Smith, Wendy Lam, Nina Deng
University of Massachusetts Amherst

In the spring of 2007, our team of psychometricians at the University of Massachusetts studied the psychometric qualities of the 2006 high school (grades 9 and 10) Science and Technology/Engineering (STE) tests—Biology, Chemistry, Introductory Physics, and Technology/Engineering. Our primary goals were to (1) determine the psychometric similarities and differences among the four tests, and (2) provide worthwhile psychometric data on each test that might help in the evaluation and ongoing development of these tests. Our goals were consistent with the No Child Left Behind [NCLB] legislation that requires states to "have implemented a set of high quality, yearly student assessments…in science" (NCLB, 2001) with the focus on the psychometric quality of the tests.

In addition to this summary of our findings, we have prepared reports for each test providing the full set of analyses we completed (Deng & Hambleton, 2008, Lam & Hambleton, 2008, Smith & Hambleton, 2008, Zhao & Hambleton, 2008). The four reports are approximately parallel. Some variations are due to the particular interests of the persons who completed the analyses, and others are due to the characteristics of the tests. For example, because of the smaller number of students who took the Technology/Engineering test, we used a small-sample approach for identification of DIF; and because of special interests on the part of some of the researchers, the model fit analysis included some extra analyses. Each report, therefore, contains the main analyses as well as some special analyses of lesser importance that were of interest to the researchers.

The basic structure of each test is described in the reports: 45 questions, including 40 multiple-choice items and five polytomously-scored items. Information about test contents (strands and learning standards) is readily available at the Massachusetts Department of Education Web site (www.doe.mass.edu/mcas). Before beginning the psychometric analyses, we attempted to rectify any problems that may have existed in the data files we were working with. We are pleased to report that the

[1] The authors want to acknowledge Michael Nering from Measured Progress for his helpful direction and advice, and for his assistance with one of the IRT model fit analyses. That being said, neither he nor Measured Progress and the Massachusetts Department of Education who supported the study are responsible for any gaps or errors in our analyses. The authors are solely responsible for any errors that may be contained in the report. The authors also benefited from the considerable assistance of April Zenisky in completing the DIF analyses. Stephen Jirka and Christine Lewis from the University of Massachusetts also participated in the discussions of the analyses and findings, and so we are grateful, too, for the assistance they gave us.

[2] *Center for Educational Assessment Research Report No. 649.* Amherst, MA: University of Massachusetts, Center for Educational Assessment.

files were in excellent shape, and the only editing we did was to eliminate from the psychometric analyses those students who omitted every question or scored zero on every question. Had we been reporting scores, these students would have been included, but for psychometric analyses with the focus on the tests and test items, the data from these students would have simply added systematic error to the analyses and would have inflated several of the statistics, such as item discrimination indices, the percent of variance accounted for by the first factor in our dimensionality analyses, and more, and so these students were deleted from all of the psychometric and statistical analyses. The actual number of students deleted was very small, less than 5% of the total sample.

In this report we will present the evidence we compiled to address the psychometric quality of the four 2006 MCAS high school science and technology/engineering tests and how they compared:

1. **Item Analyses**. In this section we reported on our efforts to determine item quality as judged by a classical item analysis—an investigation of item difficulties and item discrimination indices. Also, with three of the four tests we also carried out analyses of the distracters.

2. **Basic Test Statistics and Reliability**. Means and standard deviations of science and technology/engineering test scores were reported. We looked, too, at the internal consistency of the total test scores as well as the internal consistency of the multiple-choice component and the performance component of student test scores.

3. **Test Dimensionality**. Central to the IRT models we applied to the test data (and which are used by Massachusetts in equating and scoring) is the assumption that the tests have a strong first factor that might simply be called "the competence or proficiency measured by the test." The validity of the assumption was investigated using two approaches: Eigenvalue plots and a structural equation modeling analysis.

4. **Item Calibration and Model Fit**. This was a major component of our work. First, the three-parameter logistic model was fit to the binary scored items, and the graded response model was fit to the five polytomously scored items (possible item scores ranged from 0 to 4). Item statistics were compiled along with the plots of the IRT curves. Second, regarding model fit, chi-square statistics for fit at the item level were compiled and reported, followed by detailed analyses of both the item residuals and standard residuals. And finally, fit at the test level was addressed by comparing the actual and predicted test score distributions assuming the IRT model to be true and using the best fitting model parameter estimates to generate an expected test score distribution.

5. **Test Information and Conditional Standard Errors**. These statistics were compiled because they addressed the level of score precision the tests provided over the score reporting scales.

6. **Identification of Differentially Functioning Test Items.** A standard analysis of tests today is an investigation of the extent to which male and female students, and Black, Hispanic, Asians, and White students, who are matched on science and technology/engineering proficiency, perform differentially on each test item. When they do perform differently, and in consistently different ways across the score scale (0 to 60), these items are labeled as "DIF" and are worthy of further investigation by studying the item format and content, and any other peculiarities (e.g., item position in the test) to see if an explanation can be found for the differences. Because the analyses we carried out were based on actual state level test data (as opposed to pilot test data), the most important advantage of these analyses at this time is what can be learned and used in any future item writing and item selection for the test. In our analyses, attention was limited to locating potentially problematic items. No attempt was made to try to explain the DIF. Our focus was on DIF identification and comparison of findings across the four tests.

An Analysis of the Backgrounds of Students Taking the Science and Technology/Engineering Tests

Students who took the science and technology/engineering tests in 2006 were either 9th or 10th graders. We were able to obtain the 2004 or 2005 grade 8 Science and Mathematics test scores for many of those students. A summary of our findings for students who were 9th graders in 2006, organized by the Science or Technology/Engineering test they took, appears in Table 1 and Figures 1 and 3. The findings for students taking a Science or Technology/Engineering test in the 10th grade are organized in Table 2 and Figures 2 and 4. We were able to match up 8^{th} grade scores to 9^{th} or 10^{th} grade scores for between 82.2% and 88.3% of the students. These results are, as follows:

1. Of the students whose high school scores could be matched to their 8th grade MCAS scores, over 80% of the students took Biology and nearly 100% of the students took the Chemistry test in the 10th grade; whereas 85% of the Introductory Physics students and 70% of the Technology/Engineering students took the test in the 9th grade.

2. The small number of students taking the Chemistry test in the 9th grade did noticeably poorer on their 8th grade Science and Mathematics tests than students taking the other Science and Technology/Engineering tests. In contrast, students taking the Chemistry test in the 10th grade did significantly better on the 8th grade Science and Mathematics tests than students taking the other Science and Technology/Engineering courses.

3. Students taking the Biology test in the 9th grade performed noticeably better on the MCAS 8th grade Science and Mathematics tests than students taking the other Science and Technology/Engineering courses.

We do not attach much significance to these findings however because 2006 was the first year for the tests, and it remains to be seen how schools will organize their courses to match up with the tests in the future. Very different patterns may be observed in 2007 and beyond.

What we did see in our analyses in 2006 was that the students enrolling in the Science and Technology/Engineering courses in 2006 were not equal in their backgrounds with respect to grade 8 performance on the MCAS Science and Mathematics tests, and in addition there were differences in the backgrounds of students taking the Science and Technology/Engineering courses in the 9th and 10th grades. These differences complicate the comparisons of the psychometric properties of the tests because the item and test statistics to some extent are dependent on the characteristics of the samples of students to whom the tests were administered. We will return to this point in the last section of the report.

Tables 3 and 4 contain means and standard deviations of scores, and correlations between 8th grade Science and Mathematics test scores and high school Science and Technology/Engineering scores. One observation is that though students in the high school Science and Technology/Engineering courses differed somewhat in their abilities (as measured by their 8^{th} grade Science and Mathematics test performance), the high school Science and Technology/Engineering tests were proving to be similar in difficulty for those taking them—the means varied from about 43% to 53% and standard deviations were relatively high for a 60 point test (about 12.3, on the average). (9^{th} grade Chemistry results were excluded from this analysis because of the very small sample size.)

Table 3 contains correlations involving scores from 2005 and 2006; Table 4 contains correlations involving test scores from 2004 and 2006. These correlations serve as predictive validity coefficients. Not surprisingly, correlations involving 2004 8th grade scores tend to be a bit lower (the interval between testing was two years compared to one year when the predictive validities were coming from 2005 8^{th} grade scores), but all the correlations (except for one) show high correlations with Science and Technology/Engineering test performance in high school. In addition, the grade 8 Mathematics and Science test scores tended to be highly correlated, with none of the correlations below 0.77, and with a range between 0.77 and 0.87.

Table 1. 2006 Grade 9 Science and Technology/Engineering Scores and 2005 Background Test Scores

Test	Test Score		
	N	$\bar{X}$	SD
Biology, 2006	8,347	32.7	12.7
- Grade 8 Science, 2005	8,347	33.9	9.7
- Grade 8 Mathematics, 2005	8,347	34.5	12.3
Chemistry, 2006	72	23.9	15.2
- Grade 8 Science, 2005	72	27.8	10.8
- Grade 8 Mathematics, 2005	72	28.5	15.1
Introductory Physics, 2006	11,561	30.5	12.1
- Grade 8 Science, 2005	11,561	31.4	10.3
- Grade 8 Mathematics, 2005	11,561	31.8	12.5
Technology/Engineering, 2006	1,538	29.0	10.2
- Grade 8 Science, 2005	1,538	31.7	9.3
- Grade 8 Mathematics, 2005	1,538	30.7	11.9

Table 2. 2006 Grade 10 Science and Technology/Engineering Scores and 2004 Background Test Scores

Test	Test Score		
	N	$\bar{X}$	**SD**
Biology, 2006	40,800	29.5	12.2
- Grade 8 Science, 2004	40,800	30.6	9.4
- Grade 8 Mathematics, 2004	40,800	31.2	11.0
Chemistry, 2006	12,838	30.8	12.9
- Grade 8 Science, 2004	12,838	34.9	9.4
- Grade 8 Mathematics, 2004	12,838	37.6	10.9
Introductory Physics, 2006	2,254	32.1	13.5
- Grade 8 Science, 2004	2,254	30.5	10.4
- Grade 8 Mathematics, 2004	2,254	31.8	12.5
Technology/Engineering, 2006	633	26.3	10.9
- Grade 8 Science, 2004	633	29.6	10.4
- Grade 8 Mathematics, 2004	633	28.7	11.9

Table 3. 2005-2006 Correlations: Grade 9 vs. Grade 8 Science and Mathematics Scores

Test	Variable Pair	Correlation
Biology	Biology vs. Grade 8 Science	0.80
	Biology vs. Grade 8 Mathematics	0.77
	Grade 8 Science vs. Grade 8 Mathematics	0.83
Chemistry	Chemistry vs. Grade 8 Science	0.83
	Chemistry vs. Grade 8 Mathematics	0.86
	Grade 8 Science vs. Grade 8 Mathematics	0.87
Introductory Physics	Introductory Physics vs. Grade 8 Science	0.78
	Introductory Physics vs. Grade 8 Mathematics	0.76
	Grade 8 Science vs. Grade 8 Mathematics	0.83
Technology/Engineering	TE vs. Grade 8 Science	0.72
	TE vs. Grade 8 Mathematics	0.66
	Grade 8 Science vs. Grade 8 Mathematics	0.81

Table 4. 2004-2006 Correlations: Grade 10 vs. Grade 8 Science and Mathematics Scores

Test	Variable Pair	Correlation
Biology	Biology vs. Grade 8 Science	0.72
	Biology vs. Grade 8 Mathematics	0.70
	Grade 8 Science vs. Grade 8 Mathematics	0.77
Chemistry	Chemistry vs. Grade 8 Science	0.72
	Chemistry vs. Grade 8 Mathematics	0.74
	Grade 8 Science vs. Grade 8 Mathematics	0.78
Introductory Physics	Introductory Physics vs. Grade 8 Science	0.76
	Introductory Physics vs. Grade 8 Mathematics	0.80
	Grade 8 Science vs. Grade 8 Mathematics	0.81

Test	Variable Pair	Correlation
Technology/Engineering	TE vs. Grade 8 Science	0.71
	TE vs. Grade 8 Mathematics	0.65
	Grade 8 Science vs. Grade 8 Mathematics	0.80

Figure 1. 2005 Grade 8 MCAS Science Score Means Reported by 9th Grade Science Course

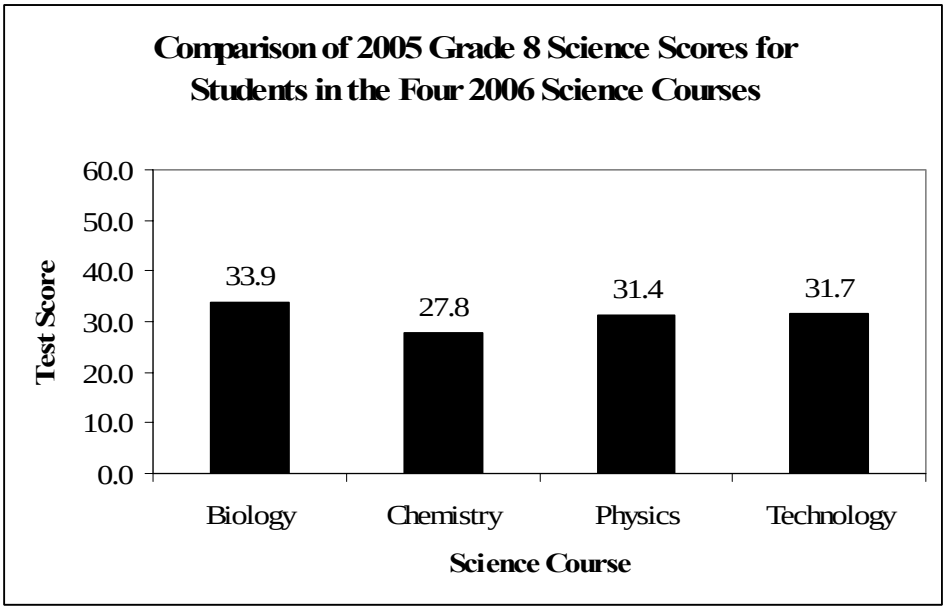


Figure 2. 2005 Grade 8 MCAS Mathematics Score Means Reported by 9th Grade Science Course

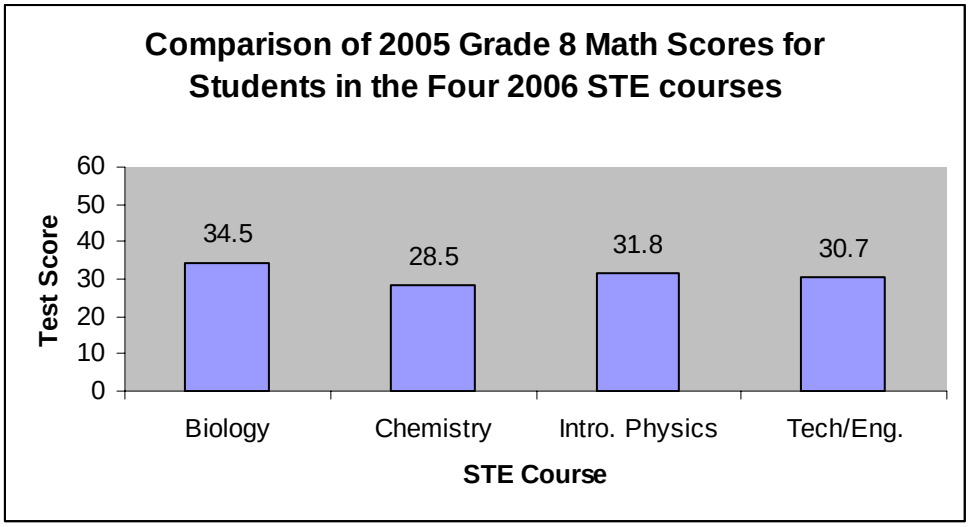


Figure 3. 2004 Grade 8 MCAS Science Score Means Reported by 9th Grade Science Course

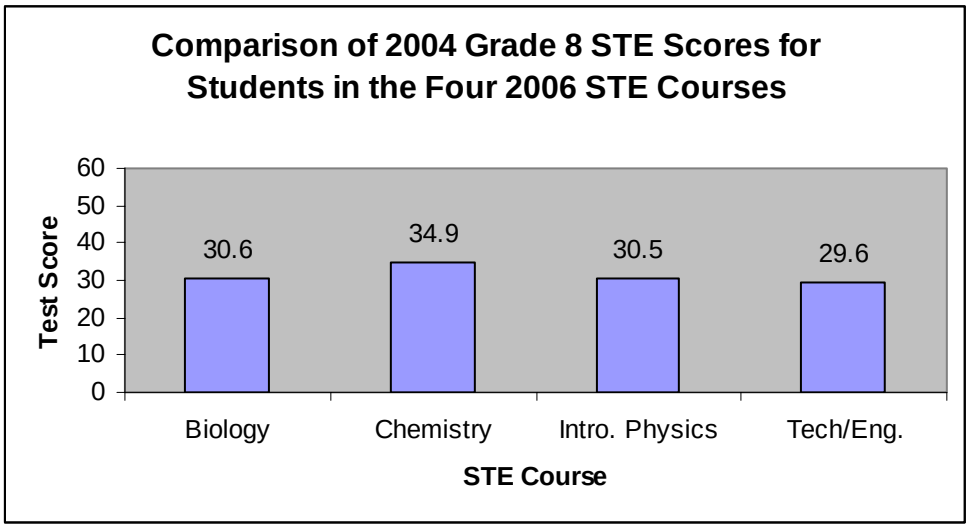
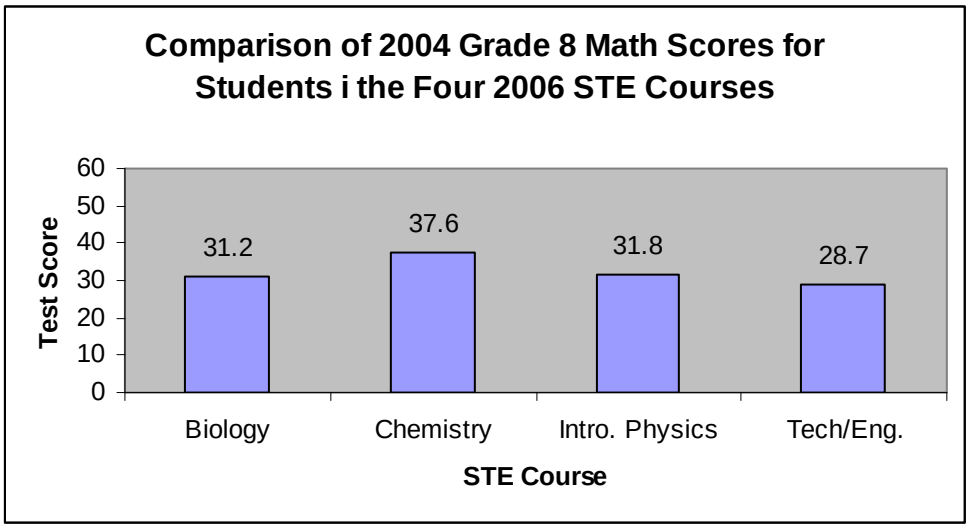


Figure 4. 2004 Grade 8 MCAS Mathematics Score Mean Reported by 9th Grade Science Course



Basic Descriptive Information about the Tests

All of the STE tests administered at grades 9 and 10 in 2006 included 40 multiple-choice items, scored 0 to1, and five polytomously-scored items, scored 0 to 4. The range of total test scores is from 0 to 60. Test statistics appear in Table 5. Complete content information on the tests is available at the Department of Education's Web site (www.doe.mass.edu/MCAS and will not be repeated here.)

Table 5. Test Score Statistics

Test	N	Mean	SD	Min	Max
Biology	55,673	29.3	12.5	1	60
Chemistry	14,997	30.1	12.9	1	60
Introductory Physics	15,762	30.1	12.6	1	60
Technology/Engineering	2,641	27.5	10.6	1	60

An interesting finding in Table 5 is that all of the high school science tests had test score means about 50%, and the standard deviations were similar, though one (Technology/Engineering) was a bit lower. The score distribution for Technology/Engineering reflected a somewhat lower standard deviation than we observed with the other three tests. Also, students taking the Technology/Engineering course did not do quite as well on the test they were given (achieving about 46% of the available score points). Of course, the groups taking the four tests are not necessarily comparable, and so it is not known whether the Technology/Engineering test was harder, or the students taking the test were less capable. We suspect that both possibilities may be true and we base this evaluation on the results reported in Tables 1 and 2. Certainly there is some evidence suggesting that the 10th grade students who took the Technology/Engineering course in 2006 (about 600 of the 2100 students) were a little less prepared than students taking the other three STE courses. This was not the case though with 9th grade students.

Item Analyses

Table 6 provides a summary of the item analysis findings compiled for each test. These statistics confirm that the tests were relatively difficult for the students (students were averaging about 50% of the score points), which means that the tests were definitely on the difficult side—multiple-choice items were somewhat easier than the performance tasks (between 10% and 16% across the four tests) and discrimination indices were high (highest for the polytomous response items) and generally of very high quality. We felt that these classical item indices showed the four tests to be generally excellent in technical quality, and the high quality is consistent across the four tests. Items in the Technology/Engineering test appeared noticeably harder than items in the other three tests, and discrimination levels were noticeably lower, though still in the highly acceptable range.

Table 6. Summary of the Science and Technology/Engineering Test Item Analyses

	Item Difficulty				Item Discrimination			
	Multiple-Choice Items		Performance Items		Multiple-Choice Items		Performance Items	
Test	$\bar{p}$	**SD(*p*)**	$\overline{X}$	**SD(X)**	$\bar{r}$	**SD(*r*)**	$\bar{r}$	**SD(*r*)**
Biology	0.53	0.14	0.40	0.12	0.41	0.07	0.70	0.03
Chemistry	0.56	0.15	0.40	0.07	0.42	0.08	0.76	0.04
Introductory Physics	0.54	0.15	0.43	0.15	0.36	0.08	0.72	0.06
Technology/Engineering	0.51	0.14	0.36	0.11	0.32	0.09	0.55	0.02

* There are 40 multiple-choice items and five performance items in each test. $\overline{X}$ and SD(X) were rescaled (by a factor of 4) so that they could be compared to the $\bar{p}$ and SD(p) obtained with the multiple-choice items.

Test Reliabilities

Coefficient alpha for the STE tests (and for multiple-choice and performance items separately) are reported in Table 7. They are consistently high, with the statistics ranging from 0.87 to 0.92 for the total test scores. Again, as with the item analysis findings, statistics for the Technology/Engineering test are a bit lower than for the other three tests. It is not clear whether this is a reflection of the tests themselves or the homogeneity of the test scores compared to the other three tests. But, in sum, all of the reliability statistics are high and acceptable by current standards.

Table 7 Test Score Reliabilities

Test	Portion of the Test	Coefficient α
Biology	All Items	.91
	MCQ only	.88
	Performance Items	.81
Chemistry	All Items	.92
	MCQ only	.89
	Performance Items	.88
Introductory Physics	All Items	.90
	MCQ Items	.87
	Performance Items	.83
Technology/Engineering	All Items	.87
	MCQ Items	.84
	Performance Items	.75

Additional breakouts of scores of students by gender and ethnicity are reported in Deng & Hambleton (2008), Lam & Hambleton (2008), Smith & Hambleton (2008), and Zhao & Hambleton (2008). Gender differences were very small (see Table 11). Ethnic groups differed, typically, by about one standard deviation, except for the Asian students who typically performed about 0.25 standard deviations above the white students (see Table 12). These findings are similar to those reported with other MCAS tests in grades 3 to 8, and 10..

Test Dimensionality

At this point in our analyses, we drew a random sample of approximately 5,000 students from the available student data for each test to continue with our investigations. The exception was Technology/Engineering. With this test we had data from approximately 2,700 students, so we retained all of the students in our subsequent analyses. Sampling was done to simply make the analyses more manageable. Five thousand students is a more than sufficient sample size for all of the analyses we wanted to complete, except for DIF analyses, where we continued to use all the available data because of the need to have the largest possible number of Black, Hispanic, and Asian students.

The eigenvalue plots revealed the strong first factor associated with each test (see Figures 5, 6, 7, and 8). The first factor with Biology, Chemistry, and Introductory Physics accounted for 31%, 34%, and 30%, respectively, of the variance and a trivially small and non-significant second factor (as indicated by the parallel analysis we carried out). Also, see Table 8 below for a summary of the findings. In practice, it is common to argue for a strong first factor when it accounts for more than 20% of the variability, and the ratio of the first eigenvalue to the second is at least five. These minimums were far exceeded with the first three tests and met for the fourth one. Clearly then, there was a strong first factor with all four STE tests. The Technology/Engineering test was the only one of the four that showed a second factor worthy of follow-up investigation and this factor was quite small.

Table 8 Test Dimensionality Analysis

Test	% Variance on First Factor	Eigenvalue (Largest Four)			
		1	2	3	4
Biology	31%	13.8	1.7	1.4	1.2
Chemistry	34%	15.3	1.8	1.2	1.1
Introductory Physics	30%	13.5	1.7	1.4	1.3
Technology /Engineering	24%	10.8	2.0	1.4	1.4

Our efforts to use structural equating modeling as an analytic model for investigating test dimensionality were successful, and these findings also suggested a strong first factor. By conventional standards, these findings, obtained from two different

methodologies, are sufficient to consider using unidimensional item response theory (IRT) models in the test analysis (Hambleton, Swaminathan, & Rogers, 1991), assuming of course that an IRT model can be found that actually fits the unidimensional test data. Model fit is addressed next.

Figure 5. Biology Test Eigenvalue Plot

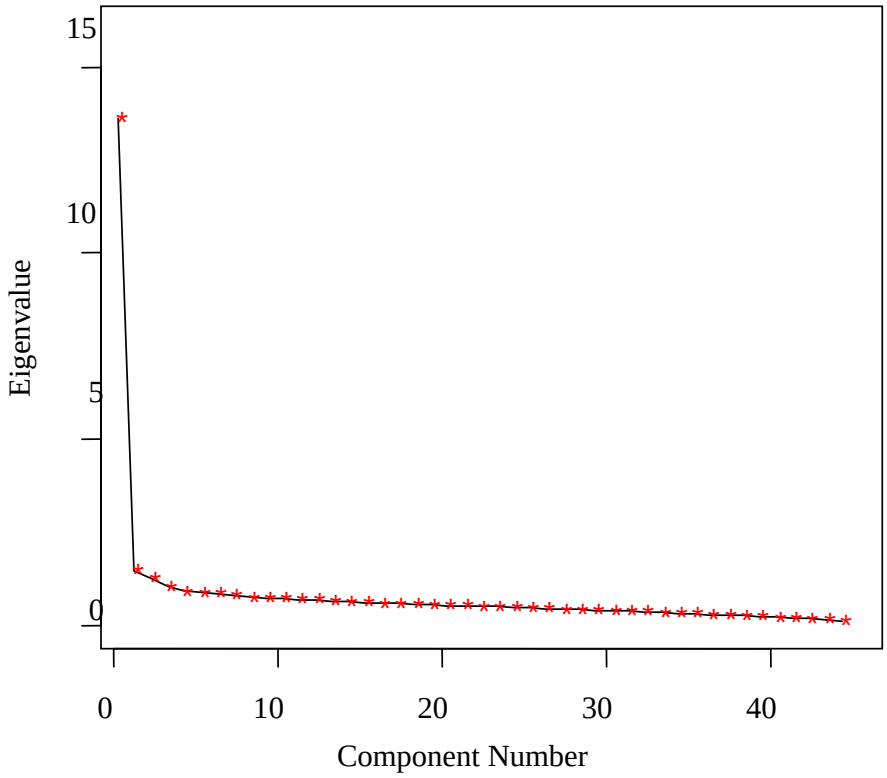


Figure 6. Chemistry Test Eigenvalue Plot

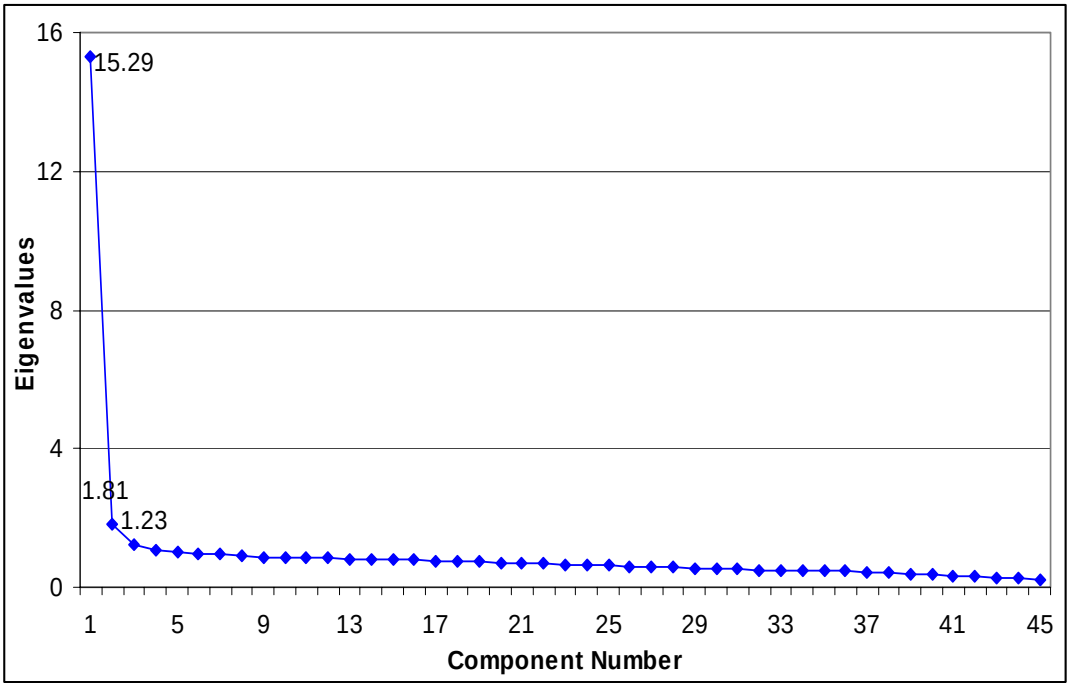


Figure 7. Introductory Physics Test Eigenvalue Plot

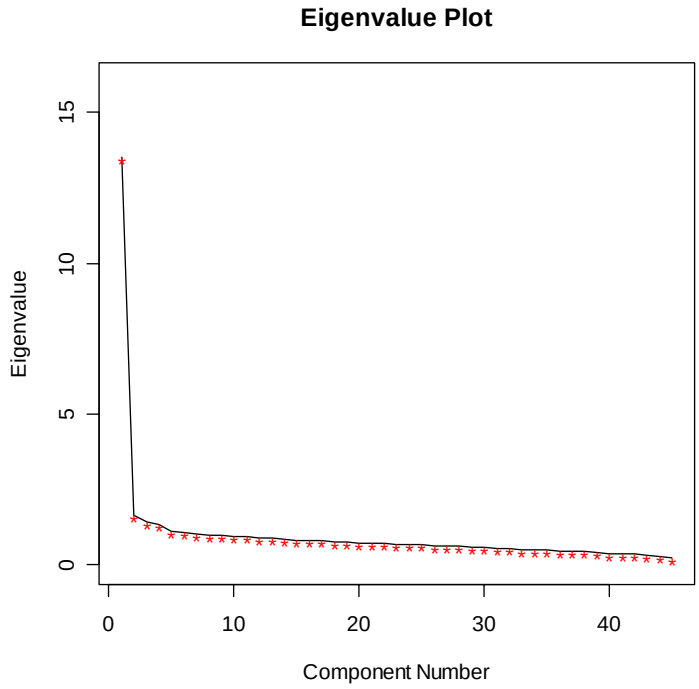
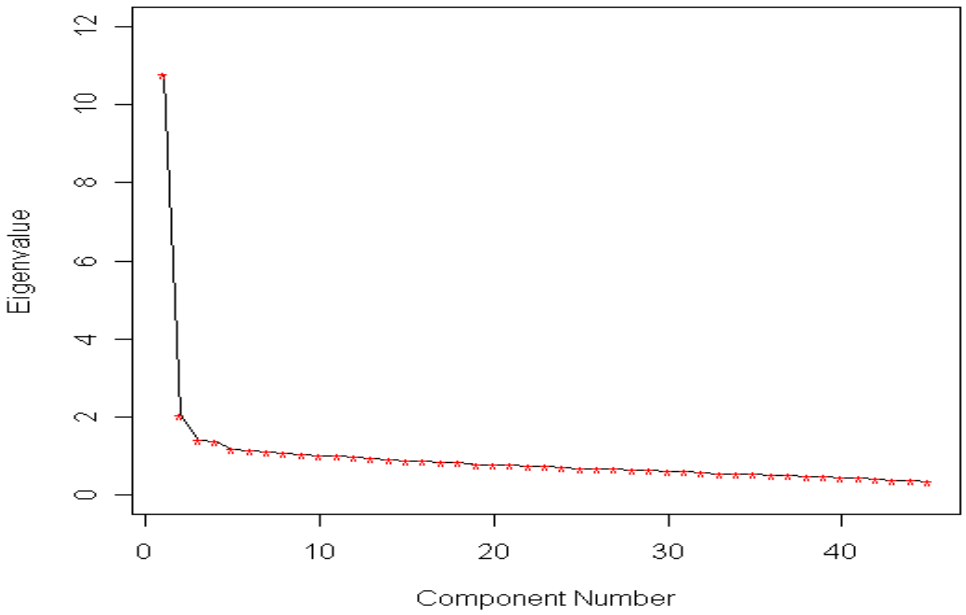


Figure 8. Technology/Engineering Test Eigenvalue Plot



Item Calibration and Model Fit

We were easily able to fit the three-parameter logistic model to the multiple-choice items, and the graded response model to the polytomous-response items, using a software program called "Parscale." Table 9 provides a summary of the item statistics. They confirmed again that the test items tended to be difficult for the students, and item discrimination levels were high. Again, the Technology/Engineering Test was a bit more difficult and the items a little less discriminating.

The fits were excellent for all of the tests with only a few misfitting items. To carry out the item level fit analyses, we considered item residuals and standardized item residuals. If the models fit perfectly to the data, the means of the standardized residuals would be approximately zero and the standard deviations would be about one (see Hambleton, et al., 1991). The results shown in Table 10 are close to those values. For Introductory Physics there was a bit less model fit (see the SD of 1.22).

Table 9. Summary of Item Parameter Estimates

Test	*A*		*b*		*C*	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Biology	0.98	0.21	0.35	0.70	0.20	0.08
Chemistry	1.16	0.34	0.25	0.70	0.23	0.07
Introductory Physics	0.96	0.34	0.29	0.73	0.19	0.07
Technology/Engineering	0.83	0.25	0.55	0.76	0.20	0.09

Table 10. Summary Statistics of Standardized Residuals (SR)

Test	Mean	SD
Biology	-0.06	1.05
Chemistry	-0.17	1.04
Introductory Physics	-0.10	1.22
Technology/Engineering	-0.07	0.92

Model fit at the test level was assessed using a graphical procedure to check whether the observed test score distribution was consistent with the predicted test score distribution, assuming the IRT model to be a true fit to the test (see Figures 9 to 12). The findings were that the predicted and actual test score distributions were very much in line, confirming excellent IRT model fit. This finding of excellent model fit combined with the strong evidence of test unidimensionality provides the support needed to use the IRT item statistics with MCAS science tests in test score equating.

Figure 9. Observed and Predicted Test Score Distributions – Biology. (The smooth curve is the predicted distribution.)

Figure 10. Observed and Predicted Test Score Distributions – Chemistry. (The smooth curve is the predicted distribution.)

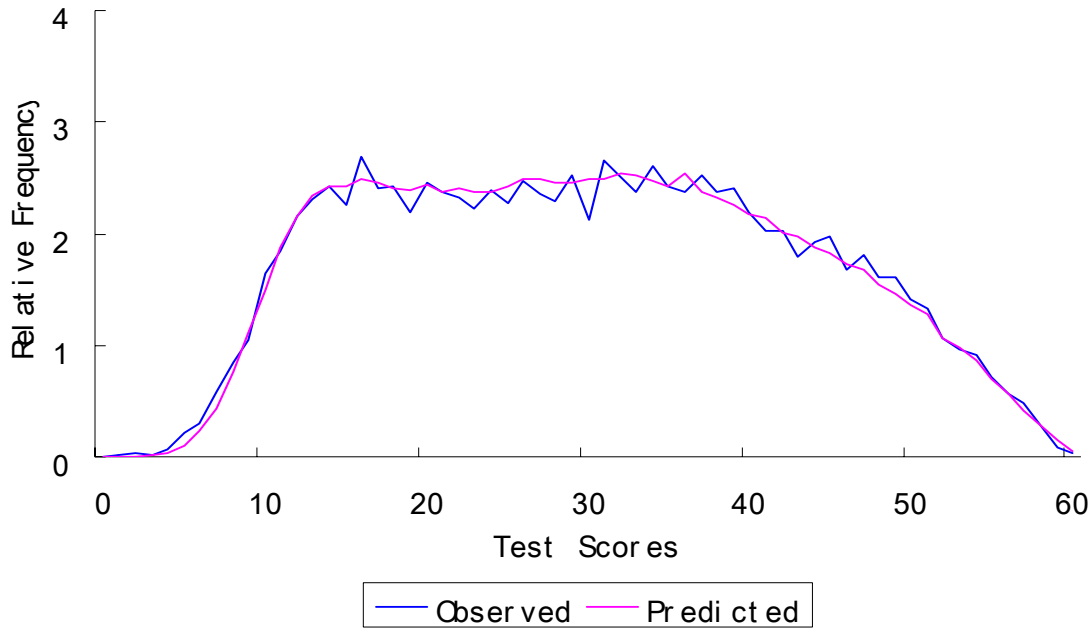


Figure 11. Observed and Predicted Test Score Distributions – Introductory Physics. (The smooth curve is the predicted distribution.)

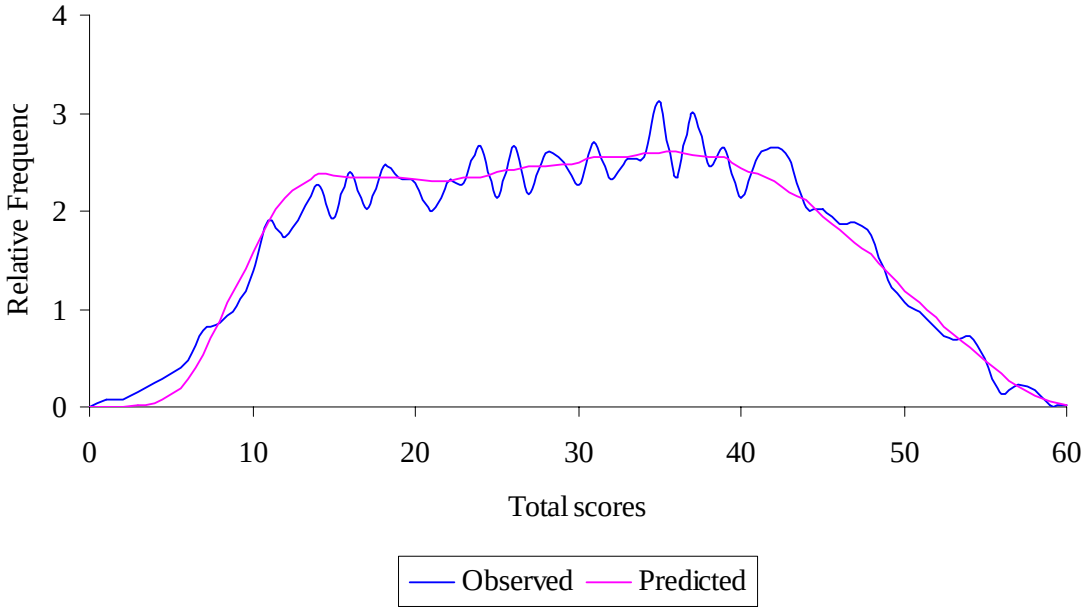
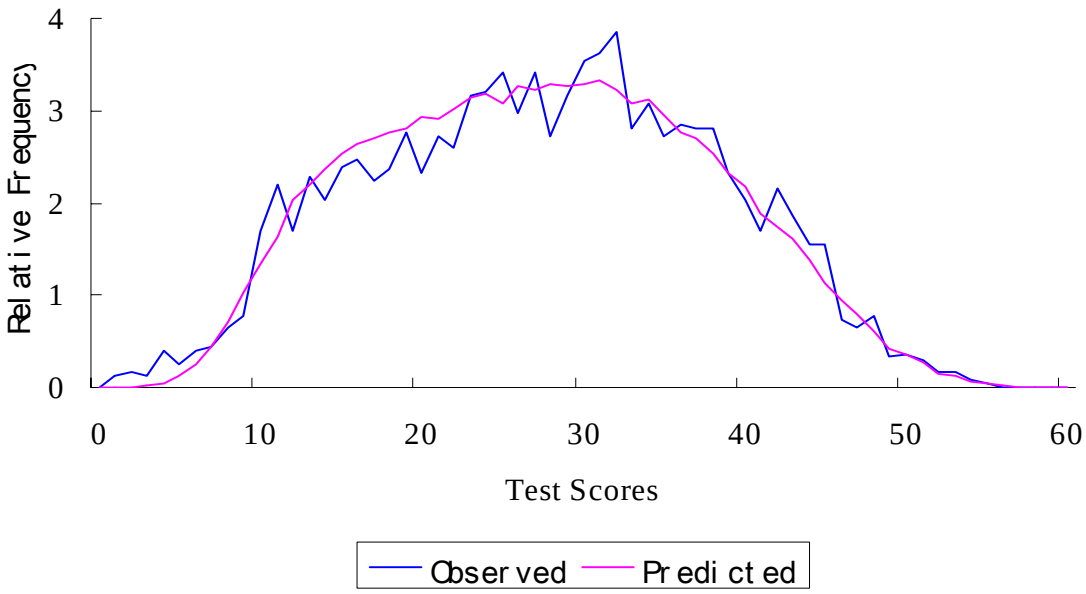


Figure 12. Observed and Predicted Test Score Distributions – Technology/Engineering. (The smooth curve is the predicted distribution.)



Test Information and Conditional Standard Errors

Figures 13 and 14, respectively, provide the test information function and the associated level of measurement error along the proficiency continuum for each of the STE tests. These figures reveal, as do other psychometric analyses, that the tests are on the difficult side. To provide better measurement in the lower half of the STE proficiency continuum, future tests will need to include some easier questions or constructed-response items for which score points are easier to achieve in order to increase the precision of STE scores for lower performing students. In the case of the Biology, Chemistry, and Introductory Physics tests, there is substantial extra information available for assessing proficiency so that some medium difficult items could be replaced by easier questions. This shift in difficulty is especially important for the Technology/Engineering test.

Figure 13. Information Functions for the Science and Technology/Engineering Tests.

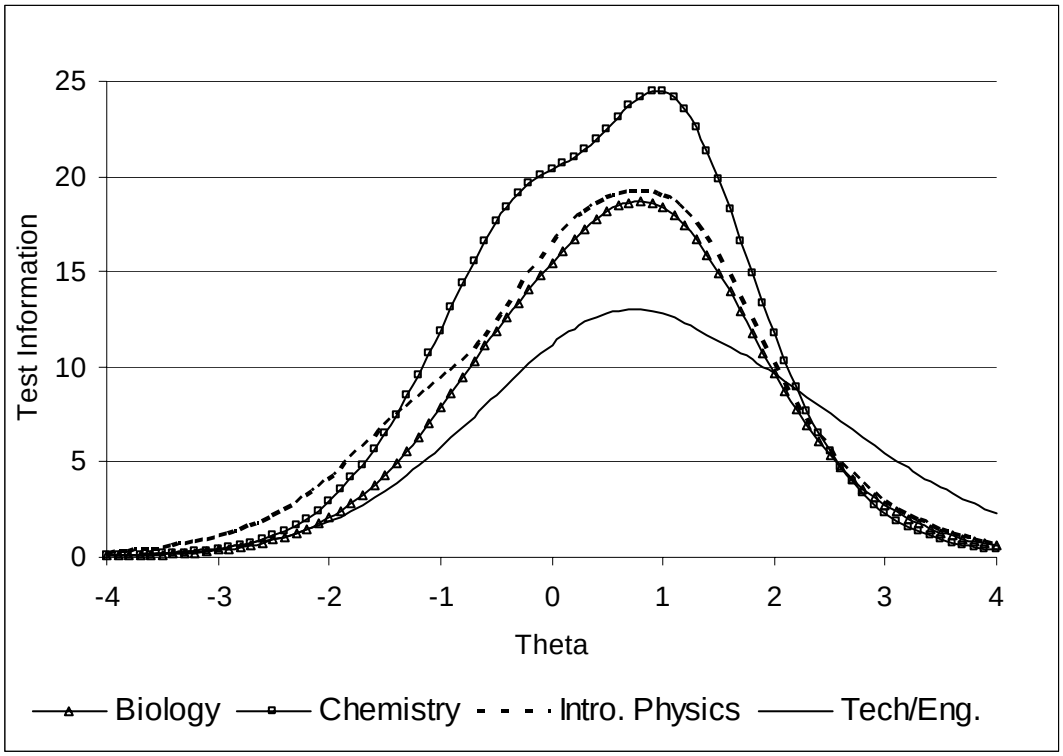
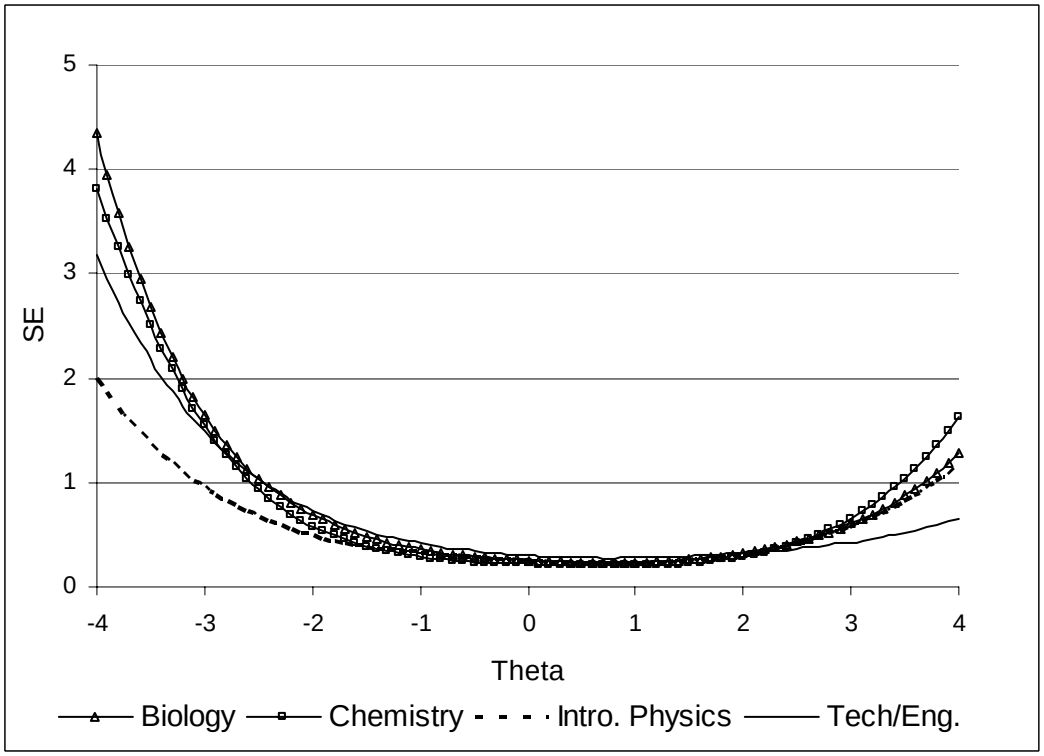


Figure 14. Plot of conditional standard errors of measurement.



Differential Item Functioning

Our efforts to identify potentially biased test items (called "differential item functioning" or "DIF") centered on comparing the item level performance of male and female students matched on science proficiency, and comparing the performance of White, Hispanic, Black, and Asian students (Zenisky, Hambleton, & Robin, 2003, 2004). Sample size information is contained in Tables 11 and 12. The Native American sample of students was too small to carry out any DIF analyses. We were only able to investigate DIF against the Asian sample on two of the tests.

Only a small amount of DIF was detected in our analyses. Of 13 DIF analyses with 45 test items each (or, in total, 585 test items) only seven test items were identified, and of those seven, four were due to gender bias. Our criterion for detection was items showing an average conditional difference at score levels of .10, or about $1/10^{th}$ of a score point. Clearly, the amount of DIF being detected was small—about 1% of the test items.

Table 11. Frequencies and Means of Total Scores by Gender

Subject	N	Male		Female	
		%	Mean	%	Mean
Biology	54,794	50	29.0	50	**30.0**
Chemistry	14,796	46	30.6	54	30.0
Introductory Physics	15,321	51	30.3	49	30.6
Technology/Engineering	2,364	69	28.1	31	27.0

Table 12. Frequencies and Means of Total Scores by Ethnicity

Subject	N	Asian		Black		Hispanic		Native		White	
		%		%		%		%		%	
Biology	54,794	4	34.2	8	21.6	9	20.8	0.3	26.7	79	31.0
Chemistry	14,796	8	35.0	8	21.4	9	20.0	0.2	27.1	74	32.1
Introductory Physics	15,321	6	34.9	13	21.3	12	20.7	0.3	28.7	69	33.3
Technology/ Engineering	2,364	2	28.4	7	19.3	10	21.4	0.4	18	79	29.5

Table 13. Number of Differentially Functioning Items by Test, and by Groups

Groups	Biology	Chemistry	Introductory Physics	Technology/ Engineering
Males, Females	0	0	1	3
Whites, Blacks	0	1	1	--
Whites, Hispanics	0	1	0	0
Whites, Asians	0	0	--	--
Whites, Native Americans	--	--	--	--

"—" means that the analysis was not carried out because at least one of the sample sizes was too small to permit a meaningful DIF analysis.

Conclusions

We carried out all the important classical and modern psychometric analyses that we thought were necessary to determine the psychometric qualities of the four Science and Technology/Engineering tests.

Our analyses showed some interesting patterns of enrolment in the four science and technology/engineering courses in the 9^{th} and 10^{th} grades. They were quite different, and these results will be of interest to policy-makers and educators. Of course, they may also be unique to 2006. From our psychometric perspective, these analyses reported in Tables 1 to 4 and Figures 1 to 4, showed that there is substantial score variability on the grades 9 and 10 tests to carry out a variety of analyses. In addition, the tests proved to be quite similar in their difficulties for students taking the tests. Test score variability was high too. In addition, it was shown that the grade 8 mathematics and science tests have similar predictive validities with the grade 9 and 10 tests, a finding which highlights the similarities of the four high school science tests. If one or more of these tests had been noticeably superior or inferior, we would have expected the predictive validities to vary too. That was clearly not the case.

The item analyses provided strong evidence for the quality of the test items—item discrimination indices were high. These analyses also revealed that for the 2006 STE students, the tests tended to be on the difficult side. Test score reliabilities, as measured by coefficient alpha, were high with values based on the total test scores being 0.88 or higher.

Our investigation of test unidimensionality, using eigenvalue plots and structural equation modeling, revealed that all four tests had strong first factors, a prerequisite for effective use of unidimensionality IRT models. After fitting the three-parameter IRT model to the binary-scored test items, and the graded response model to the polytomously-scored test items (models that are applied to other MCAS tests), we found model fit to be excellent. Both residuals and standardized residuals highlighted excellent model fit, and predictions of test score distributions were accurate. All these analyses highlighted that the IRT modeling of the data would support the use of these models in test development, test score equating, and score reporting.

Our analysis of the test information functions and standard error of measurement functions revealed that the current tests were not optimally centered in relation to the students. Also, the available information for Chemistry was especially high, and high for Biology and Introductory Physics, too. With all three tests it would be easy to substitute some easier or harder items of the same content to improve measurement precision at the extremes of the proficiency scale with little loss in the middle of the proficiency scale. The situation for Technology/Engineering is not quite so simple.

We searched for evidence of differential item functioning (DIF) in the data and turned up only a few items worthy of further study. We investigated both gender and ethnic DIF.

In sum, both our classical and modern psychometric analyses of the tests show that the current tests are technically sound and the three-parameter model and graded response models fit the test well, essential for effective IRT-based equating. The number of potentially biased items is very small, so small that type I error cannot be ruled out as an explanation. It is difficult to make a recommendation about the current test difficulties. For the 2006 samples of students, the tests were a bit on the difficult side, and this is reflected by the mean difficulties and the placement of the information functions. If the expectation is that the students will become more capable and will perform better on the tests as the STE curricula are widely adopted, then tests like those constructed for 2006 may be acceptable. If the current performance is not expected to change very much in the coming years, then it may be desirable to make some minor adjustments in the levels of test difficulty so that more measurement precision will be associated with the scores of the low-performing students.

Clearly, the four Science and Technology/Engineering tests in their current form are sound technically, and highly comparable in quality. The presence of a very small number of DIF items and some less than optimal placements of test information functions are small flaws in overall, excellent quality tests.

References

Deng, N., & Hambleton, R. K. (2008). *Psychometric analyses of the 2006 MCAS High School Introductory Physics Test* (Center for Educational Assessment Research Report No. 647). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Dorans, N. J., & Kulick, M. S.. (2006). Differential item functioning on the mini-mental state examination. *Medical Care, 44(1), Number 11 Supplement 3*, S107-S114.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Holland, P.W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. Braun. (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Lam, W., & Hambleton, R. K. (2008). *Psychometric analyses of the 2006 MCAS High School Chemistry Test* (Center for Educational Assessment Research Report No. 646). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

No Child Left Behind Act of 2001, Pub. Law No. 107-110.

Smith, Z., & Hambleton, R. K. (2008). *Psychometric analyses of the 2006 MCAS High School Biology Test* (Center for Educational Assessment Research Report No. 645). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Zenisky, A. L., & Hambleton, R. K. (2007). *Differential item functioning analyses with STDIF: User's guide* (Unpublished report). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Zenisky, A., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement*, *63*(1), 51-64.

Zenisky, A., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment, 9*(1&2), 61-78.

Zhao, Y., & Hambleton, R. K. (2008). *Psychometric analyses of the 2006 MCAS High School Technology/Engineering Test* (Center for Educational Assessment Research Report No. 648). Amherst, MA: University of Massachusetts, Center for Educational Assessment.