APPENDIX A -- DISCUSSION OF MAJOR ISSUES

This appendix contains additional discussion and elaboration of the strategic issues described in the January 1997 report to MSBE. The numbering of these comments corresponds to the numbers identifying the issues in the body of the report. Section II, which describes the four assessment options, has been divided into two distinct portions. Section II A describes the advantages, disadvantages, and constraints of each design option overall without addressing differences between content areas. Section II B provides subject-specific comments from College Board/ETS staff and the Maryland content teams on these four design options. In addition, illustrative items which may be used in each of the design options are presented.

I. STANDARDS AND THE USES OF THE HSA

A. Proposed Uses of the HSA

Use of the High School Assessment (HSA) for individual student accountability received the most attention during the several public engagement activities. It is clear that a substantial proportion of educators and parents have not yet been persuaded that state-mandated assessments are an effective and fair means of raising expectations and standards. A sustained effort to make the case for this use will be required over the next several years. The arguments for and against this use are complex; they are briefly summarized in the report.

Accountability is an important ingredient of education and a responsibility from which policymakers cannot shirk. Without objective and consistent data, people are unable to make informed decisions about their schools and their students. Educators will be unable to determine the degree of student competency with respect to important learning goals and educational indicators. Educators will also be unable to compare student performance across the nation, over time, and from school to school.

Nearly all participants in the public engagement activities strongly supported **"higher standards."** Most participants believe students in Maryland can do better. And all participants desire more for today's students and tomorrow's citizens.

But most participants in the public engagement activities questioned whether all students can meet and should be held to **the same high standards.** As one participant noted, we can all think of several former classmates who could never have met those standards, but we should not deny them a diploma if they pass their courses. Many of these participants questioned the meaning of terms such as "world-class standards" and were very apprehensive about whether MSBE's definition of high standards would be consistent with their own. Many individuals addressed this specific issue in their remarks, but few if any supported the use of the "Ten High Hurdle" race, while over 90% spoke against various elements of that model. The majority of those individuals addressing the use of the HSA suggested other approaches such as differentiated diplomas (reward high-performing students rather than punish low-performing ones) or school accountability uses. Others suggested phasing in tests, standards, and high stakes until data are

available to the public and educators about the implications for students (e.g., failure rates, remediation options).

The exact height of the bar used to establish the standards is an important issue given the variability found across student performance in nearly all schools and districts in the nation. There are significant legal risks in holding all students to the same high standards when students differ so greatly in their performance on all current measures (e.g., grades, GPA, level and difficulty of courses, SAT and ACT scores, NAEP results). That is, despite standards and high-stakes tests, individual students differ in performance, often dramatically, within the same district, school, and classroom.

Policymakers and the media typically deal with aggregate (group-level) test scores, but rarely examine scores for individual students. Aggregate test results condition us to recognize that there are sizable variations in student performance across schools, districts, states, and different subgroups of students (differences with respect to ethnicity, language, geographical region, urban/rural/suburban residence). However, the variation in students' test scores within a school almost always exceeds the mean variation of scores across schools and districts. That is, often there is a substantially greater difference between the individual scores of students at the 25th and 75th percentiles than there is between the mean scores for schools at the 25th and 75th percentiles. When high stakes are associated with individual students rather than schools, as with the HSA, significant negative consequences result for those students. It is one thing to improve the mean score for a school by 10%; it is substantially more difficult to increase the score for each student in that school by 10%. When this does not occur, it is the individual student that will be impacted. Individual differences among a group of students account for much more variation than differences among schools -- which will likely result in a sizable proportion of students not meeting any moderate to high standard for a prolonged period of time.

At the school and district levels, there are substantial differences in curriculum, instruction, and available resources. Even with significant educational reform, extensive curriculum enhancements, and honors programs, student variation in performance exists. School- and district-level variations present a second type of risk to the assessment. Before implementing the final performance standards, MSBE needs to understand the extend of such variations:

- how many students, in which districts, complete Algebra and Geometry at each grade? How many students complete these subjects in discrete versus integrated courses? Which Core Learning Goals (CLGs) are covered in which courses, when, across districts? How many students are enrolled in half-credit courses for each content area? These issues may be more relevant in science where variations in curriculum, student completion, and integration of subjects and CLGs present serious concerns for the design of assessments.
- how many students in which schools are completing middle and high school in the various block formats?
- do students have multiple opportunities to learn each CLG? When and in which courses ?
- do students have prior exposure to the skills and processes necessary to master the CLGs at earlier grades; if so, when and where ?
- do schools have adequate plans and resources for implementing remedial programs for students who fail the HSAs?

- what additional variations across schools and districts pose legal and practical threats to the assessment system (certified teachers, staff development, etc.)?

MSBE should undertake efforts to collect additional data on these and other variations across districts and document that all students in Maryland have multiple and adequate opportunities to learn the CLGs which will be contained in the HSAs. There must be opportunities and resources available for students who initially fail a test and require retesting and/or remediation. Similarly, additional staff development will be required for teachers across all grade levels in the four content areas (for teachers in K-6, who may not be involved in delivering courses corresponding to the HSA, as well as 8-12) to ensure integrated curricular and student learning consistent with the CLGs.

Finally, because these are end-of-course assessments, there will likely be significant political risks, and perhaps legal risks, that arise as a result of students who pass a course (perhaps many courses) with grades of A, B, and C, only to fail the state test. If such a discrepancy between course grades and test scores were found in a school or district, parents and policymakers may feel that those students have not been provided with an adequate education. The fact that a sizable percentage of students who receive grades of A or B still fail the test may add support to such claims. Because of these and many additional concerns, we strongly recommend that MSBE commission a legal review to determine the risks associated with any plan that will be implemented to make meeting high standards on the HSA a state graduation requirement.

There are some additional professional and ethical concerns surrounding the proposed plan to base graduation requirements solely on test performance. The Standards for Educational and Psychological Testing (AERA, APA, and NCME, 1985) and other professional standards and guidelines concerning testing raise serious cautions regarding the use of and reliance upon test scores in decision making. Generally, it is widely accepted that decisions with important consequences for individuals should not be made on the basis of a test score alone. Determining who will receive a high school diploma is an enormously important decision, and information beyond a student's performance on the HSA should be considered in that process. Student performance in courses, grades attained, and the level or difficulty of the courses completed all contribute important information about student achievement. These sources of information would add valuable information to such decisions. These are the same cautions and advice that are provided to employers in selecting workers and colleges in admitting students by experts in measurement and testing.

The consequences for individual students and groups of students could be severe. First, as illustrated in the examples above, individuals students who do not meet the standards will not graduate. These students require information early in their education, prior to high school, about expectations and about the opportunities available to demonstrate competency and receive remediation when needed. Much more serious attention must be devoted to issues of remediation and instruction before the HSA is implemented.

Second, there was relatively little discussion of the use of the HSA to evaluate programs and schools. Most comments cited the beneficial impact of the use of MSPAP results for this

purpose, citing the power of such measures of institutional accountability to bring about improved curriculum and instruction. Educators were primarily concerned about the **perceived differences** between the HSA and MSPAP in:

- use in accountability (student vs. school)
- perceived emphasis on content and processes
- student accountability (none with MSPAP, significant with HSA)
- focus (student performance, school performance)
- emphasis (student graduation vs. educational reform and improvement)

These and other real and perceived differences in the state's testing programs will need to continue to be addressed for successful implementation of the HSA by 1999-2000.

Third, there was little discussion about the proposed uses of the HSA by the higher education system in Maryland except among representatives from higher education themselves. There was a general sense that results from some of the assessments might apply to placement decisions, e.g., a passing score on HSA English 3 should enable a student to bypass the remedial writing class when entering college if assessments are rigorous and at the appropriate level. However, because each institution presently makes its own placement decisions (based on its unique courses and course sequences) the operational details required to use the HSA will be significant. For example, a five-point proficiency scale may not be adequately sensitive (i.e., have enough scale points) to permit appropriate placement across Maryland's higher educational institutions. In addition to a proficiency scale (used by secondary schools) scaled scores (e.g. a 30-point scale) may also be required for higher educational uses. Each college would then need to conduct a validity study to demonstrate the precision of its unique cut scores for placement decisions.

There was little enthusiasm for the use of HSA as an admissions test. Representatives of higher education and other participants in public engagement opposed basing admissions decisions solely on any one measure and emphasized the value of using multiple sources of information in admitting an incoming freshmen class. Performance on HSA could serve as an additional source of information for colleges in making admissions decisions, yet research is needed to determine the incremental validity such scores would provide over current measures of academic performance in high school (e.g., GPA, course grades, courses taken, class rank, admissions tests, etc.).

Some policymakers have posited that students could be required to attain a "minimum score" on the HSA in order to be eligible for admissions. The use of such a cut score, even if it were set at a minimal level, could have severe consequences for individuals students and groups within the state. If course grades, performance on national admissions tests, and extracurricular service and community involvement can not compensate for low performance on state assessments then the HSA will serve as an absolute barrier to admissions for some students. Such a policy will disadvantage some groups to a much greater degree than others, raising significant legal, professional, and ethical issues for Maryland educators.

During Phase I of the test design work, we have had little interaction with representatives from higher education. MSDE must ensure far greater involvement by higher education in all future

HSA activities if assessment results are to be meaningfully incorporated into admissions and/or placement decisions. The feasibility of alternative models for using the HSA in higher education should be evaluated by individuals with expertise in college admissions, measurement and statistics, test development and test validation, curriculum and instruction, and the law before development activities proceed in this area.

B. Differentiated Diplomas

A number of participants in the public engagement activities cited their personal experiences of attending school in New York State in championing the idea of using the HSA results to provide a higher-level endorsement for the diploma. The additional prestige and the recognition accorded the Regents Diploma by the higher education community was cited as sufficient motivation for most students to attempt the High School Assessments. The Regents program differs from the planned HSA in many important ways. The most substantive difference is that the Regents' program provides prescriptive curricula (across schools and districts) that correspond to the assessments, while Maryland strives to retain more local determination of curricula while instituting common end-of-course assessments across districts. The Regents examinations are also not used to deny students a diploma but to reward them with a distinguished diploma. These examinations are still important, but the individual stakes associated with the tests are substantially lower than those intended by Maryland.

Two models of differentiated diplomas are discussed in the report. The first model would permit students to graduate based on current graduation requirements, with the HSA serving the role performed by the current Regents Exams as a higher-level, state-endorsed diploma. The HSA diploma could be required for consideration for admissions at the most competitive state colleges, be used to certify that students can enter a college without remediation (in content areas selected by faculty within each college), or be associated with some state-sponsored scholarship or tuition reduction as a reward. Of course, if the HSA were to be used for admissions or to exempt students from remedial courses, appropriate validity studies would be required, and differentiated performance levels may be needed to accommodate the requirements of different state institutions. As noted above, there are substantial intermediate steps that would be required before the HSA could be operationally used for higher education decisions and representatives from K-16 must be involved in all efforts to determine the feasibility of various approaches.

The second model requires two or more cut scores on the HSA. Lower levels of performance would still be required for graduation, while higher levels would be used for one or more of the above purposes. This model does not alleviate many of the problems with uniform high standards and high-stakes tests, but also creates additional burdens because of the need to validate two separate uses with one score scale. The model of multiple levels of diplomas assumes that the assessment instruments provide precise measurement at the several cut points which may contradict the intended use of the HSAs for high school graduation. The psychometric issues concerning multiple uses for the same test are quite complex and will require substantial additional work by MSDE if they are to be meaningfully pursued.

C. How High Is High Enough?

The proposed standards and uses of the HSA ***require all students to successfully complete each of (or pass) ten separate end-of-course assessments to receive a Maryland state high school diploma (referred to as "Ten High Hurdle"). Further, these assessments will be based on high standards that correspond to the Core Learning Goals, already approved by MSBE.***

Such proposed standards and uses raise many significant issues and risks for MSBE.

It is impossible to provide a precise estimate of the passing rate if MSBE required ten, eight or even four separate assessments because the intercorrelation among these tests and the tests themselves are not know. If ten independent measures (that is, ten tests which have no relationship to each other) were required a very good student with a .9 probability of passing any one of the ten assessments (9 out of 10 chances) has about one chance in three of passing all ten tests (the probability of passing all ten assessments assuming complete independence of these assessments is about .35). That is, although he or she has nine chances out of ten of passing each particular assessment, the odds are about 1 in 3 of passing all ten. For poorer students whose likelihood of passing one assessment is lower, the odds of successfully passing all ten become very long. This is illustrated in the following chart, using a probability of .90 and .75:

If a student's probability of passing each test is:		
p =	.90	.75
the probability of passing		
One test	.90	.75
Two tests	.81	.56
Three tests	.73	.42
Four tests	.66	.31
Five tests	.59	.24
Six tests	.53	.18
Seven tests	.48	.13
Eight tests	.43	.10
Nine tests	.39	.08
Ten tests	.35	.06

If student performance on MSPAP is used, the probability of passing a single high school assessment will be closer to .40 or .50 than the probabilities (.90 and .75) illustrated above. However, this illustration is overly pessimistic because the ten tests will certainly be related to some extent and will not be independent events. What does remain constant is that the probability of passing multiple tests (ten, eight or four) is reduced as the number of tests increases. Compensatory models will result in higher passing rates.

Three compensatory models were briefly described in the body of the report. Each of these presents additional challenges. Psychometrically, composite scores called for in the compensatory models represents, implicitly, weighted combinations of the scores making up the composite. If

one thinks of the composite score as differentiating between students at different levels, the component instruments having the highest variances have more influence on that differentiation. Conversely, instruments having very low variance (for example, almost all students receive scores of "1" or "2") will have very little influence on the differentiation between students on the composite. That is, any assessment where student performance varies greatly (where many students get scores across the score distribution) will have greater weight than tests where most students score at the same level. Such instruments have less compensatory effect. This may not constitute a problem for the High School Assessment, but it should be considered in the design and scaling of each assessment. Similarly, the relationship (correlation) among the several instruments included in a composite will affect the effective weight of each in the composite score.

Also, as stated in the report, the compensatory models create difficulties in advising and counseling students who do not initially pass an assessment (or when a summative score is needed at the end of high school) regarding remediation or retesting. Perhaps the most educationally sound advice would be for students in the early years of high school to seek further instruction and to retake the assessment at a later date. Students in their senior year might rely on the compensating effect to meet the overall requirement *if* they have achieved a better-than-satisfactory average (including scores earned on retest) on the assessments taken during the first three years of high school.

The following illustrations describe the possible standards that would be used in the three compensatory models as well as the possible outcomes for four students. The actual standards could be changed (lowered or raised) in each example, yet the variations in results and student outcomes would remain similar. Each illustration is based on students completing ten assessments which are scored from 1 to 5 (5 the highest score, 1 the lowest). Let's also assume that all four are exceptionally good students with very high grades who have been accepted into college.

The Low-Hurdle Decathlon - Students must attain a minimum score of **"2"** on each of the ten tests, but have an overall score of **"26"** across all tests. In this example, Susan and Jon would meet this standard but not Gail and Mike.

	Math 1	Math 2	Eng. 1	Eng. 2	Eng. 3	Soc. Stud. 1	Soc. Stud. 2	Soc. Stud. 3	Sci. 1	Sci. 2	Sum Score
Susan	2	2	2	2	2	5	4	3	2	2	26
Jon	3	3	3	3	3	2	2	2	3	2	26
Mike	5	5	5	4	4	2	2	**1**	5	5	38
Gail	2	2	2	2	2	3	3	3	3	3	**25**

In this illustration Susan shows marginal performance across all tests, but her high scores in Social Studies put her above the standard for graduation. Jon has marginal to average scores across all subjects with no apparent strengths nor weaknesses; he too meets the standard. Mike, who has the best overall performance, would not meet the overall standard because he has difficulty with Social Studies and has been unable to pass the final Social Studies course. His exemplary

performance in Math, English, and Science courses cannot compensate for his low performance in Social Studies. Gail's performance is similar to that of Jon. She shows marginal to average performance across all subjects, but receives one fewer score of "3" than Jon; thus she too does not meet the standard. Gail and Jon would not really know whether to retake a test in order to meet the summative standard (26) until the end of their senior year when they completed their last test. Both had hoped that they would get at least a "3" on their final English exam and never saw the need to retake a test or undergo remediation, and school officials were not certain whether or not they would meet the standard. Jon was able to get a "3" on his senior English assessment so he will receive a diploma, but Gail, who received a "2" days before graduation, now must wait until August to retake the test, and her college acceptance is in jeopardy. Teachers, counselors, and school administrators have difficulty advising these students and their parents on their best options. Remediation and advising for Mike may be easiest. He simply needs to get his score up on one Science test and not worry about the others. He might retake the test immediately, with or without substantial remediation since he only needs to go from a "1" to a "2."

The Four-Event Competition - Here, students must achieve a satisfactory score in each of the four content areas, but scores on individual tests are not of concern. In Math and Science, students would need a total score of 5 (out of 10) because there are two tests required. In English and Social Studies, students would be required to attain a score of 7 (out of 15) since there are three tests in these subjects. In this example, Susan and Jon would again meet the standard, but not Gail and Mike.

	Mth. 1	Mth. 2	Mth. Sum	Eng. 1	Eng. 2	Eng. 3	Eng. Sum	Soc. S. 1	Soc. S. 2	Soc. S. 3	S.S. Sum	Sci. 1	Sci. 2	Sci. Sum
Susan	3	2	5	3	2	2	7	3	2	2	7	3	2	5
Jon	4	1	5	4	2	1	7	4	2	1	7	3	2	5
Mike	5	5	10	5	4	4	13	2	1	1	4	5	5	10
Gail	2	2	4	2	2	2	6	2	2	2	6	2	2	4

In this illustration Susan and Gail have moderate scores on all the tests, yet because Susan received a score of "3" on one test in each subject area (while Gail attained consistent scores of "2"), Susan meets all subject standards and Gail meets no subject standards. Gail would need to retake four tests, but might not know this until she had completed the last test in each subject. Students such as Gail might be difficult to advise and reluctant to retake tests (where they achieved a "2") or undertake remediation; there is always the hope that they could get a "3" on the next test. Jon meets all the standards, but clearly performed significantly better in the early grades than in the latter. His high scores in the entry-level courses across content areas have "carried him." Similar to the first example, Mike has exceedingly high scores in all areas (4-5) except Social Studies, which presents a significant obstacle. He would need to dramatically improve his performance on two or more Social Studies tests to graduate, and his achievements in other content areas would not compensate for this weakness. Again, this model presents difficulties for advising and counseling students who perform marginally (2) on initial tests. Should they retake the test, undertake remediation, or assume they will get a "3" on the next test and meet the standard?

The Decathlon - A very similar standard would require a total composite score of , say, "26" without any minimum scores specified on individual tests or within content areas.

	Math 1	Math 2	Eng. 1	Eng. 2	Eng. 3	S.S. 1	S.S. 2	S.S. 3	Sci. 1	Sci. 2	Sum
Susan	2	2	2	2	2	5	4	3	2	2	26
Jon	3	3	3	3	3	2	2	2	3	2	26
Mike	5	5	5	4	4	2	2	1	5	5	38
Gail	2	2	2	2	2	3	3	3	3	3	25

In this final illustration (using the same data as the first), Susan, Jon, and Mike all meet the standard. Mike's unsatisfactory performance in Government does not keep him from graduating. Gail, however, still needs one higher test score. She has been reluctant to retake any test or undergo remediation because she continued to believe she would get a "3" on her senior English test. Now two weeks before graduation she realizes that she will not receive her diploma and she needs to retake any exam of her choice.

Few students have equal strengths in all areas. Some students may have strengths and interests in writing, literature, social studies, and humanities, others may have these strengths and interests in math and/or science. Still others may excel in yet different areas. The compensatory models appear to be fairer to students because they do not have to meet multiple independent standards across ten courses, and they allow for relative strengths and weaknesses among students. Yet, counseling students and planning for remediation and retesting are difficult with the compensatory models. If students do not know if their performance on the 9th grade English test is good enough until they take the 10th and 11th grade English tests, it is difficult to plan courses, remediation, and retesting.

Psychometric considerations associated with scaling and scoring are very similar across all decision models discussed above. Any decision rules adopted must be designed to provide results that meet the objectives of the assessment. Cut points that define what differentiates, for example, a 3 from a 4, (or for example proficiencies such as -- a 'working toward competency' from 'demonstrates competency') must be established such that decision validity and reliability are achieved. This can best be achieved by having highly valid and highly discriminating items measuring at or near the cut points. This becomes more difficult if multiple cut points or decision rules are required (as with differentiated diplomas which attempt to make 3 or 4 different distinctions). Also, generalizability will be more easily achieved with compensatory models because the task by person interaction will be less important.

D. Implications of Ten Separate Tests

This issue was adequately addressed in the report and there is no additional discussion required here.

E. Timing or Phase-in of the Assessments/Graduation

Phasing in high-stakes use of the assessments, by extending the no-fault period and/or introducing a lower individual standard for students, would appear to be a very promising approach.

The benefits of a no-fault period are many:

- permits MSBE, districts, schools, and policymakers to examine student performance and model various pass/failure standards to determine probable pass/fail rates across the state, districts, and schools.
- provides educators with additional time to gain familiarity with the CLGs and how they are evaluated in the assessments.
- reduces or eliminates the legal risks associated with high-stakes tests until adequate validity and impact data can be assembled under operational conditions.
- ensures that the tests meet the highest psychometric standards and reduces potential risks associated with low reliability and the inability to report scores because of equating problems across forms.
- provides a mechanism to "try out" new procedures that might reduce scoring costs and turnaround time without having high stakes associated with this effort.

After field-testing but before all the assessments are required for graduation, MSBE could use the HSAs exclusively for school accountability purposes. There would still be high stakes associated with the assessments, but the level of psychometric rigor required and the burden on districts and states would be greatly reduced. If MSBE still wishes to have high stakes for individual students associated with the HSA (either high school graduation or models described under a differentiated diploma), one or more tests could be introduced for this purpose each year. Such a phase-in approach could reduce the burden and risks for the state. However, as noted above, phasing in the number of tests alone would not eliminate the substantial risks and probable negative consequences for students in the long term.

In summary, the use of the HSA with high standards as a graduation requirement will result in serious risks and probable negative consequences for substantial proportions of students in the state. If MSBE desires to use the HSA for individual graduation decisions, the following elements could reduce the risks to the state and districts and minimize the negative consequences for students:

1. An extended no-risk (to individual students) period to pilot test and familiarize schools and students with the assessments and planned uses to ensure that the assessments meet the highest professional standards available.
2. A phase-in of assessments and performance standards to ensure that large groups of students are not adversely affected by the tests.
3. A compensatory model that would permit high performance in some areas to somewhat compensate for lower performance in other areas.
4. Reduce the actual number of separate tests students must pass.
5. Consideration of additional forms of evidence (e.g., performance in honors courses, course grades, projects) combined with test scores to make individual graduation decisions.

II (PART A). AN EVALUATION OF THE DESIGN OPTIONS

A. Overview

The four designs options under consideration differ from each other in a number of respects although all assume that 120-150 minutes will be available for the administration of each assessment and all will integrate particular Skills for Success (SFS) as appropriate to the content. The first three options all include a mixture of Selected Response (SR) and Constructed Response (CR) questions which will require scoring by people who are knowledgeable in the subject and the Core Learning Goals. The distinguishing features of the four options are reviewed below, and some of the strengths and weaknesses of each are listed. The criteria that were used to systematically evaluate the four options are briefly described and a matrix summarizing the judgments about each against the criteria appear in Table A1. This overview is then followed by a section on each subject area which provides illustrative assessment questions linked to particular Core Learning Goals, Expectations, and Indicators, along with comments on the implications of each option for the subject area.

B. Description of Item Types

Each of the four design options described in this report makes use of one or more of the following broad types of items. Within each type there may be considerable variation in the skills that are evoked and the depth of knowledge and reasoning that is required. The choice among items types depends on the purpose(s) of the assessment and the kinds of inferences one wants to make about the response. Even more important than the choice among item types or formats is the skill which is used in creating the items to ensure that they measure the cognitive skills and knowledge that are of interest.

1. Extended Constructed Response - This item type provides the opportunity for a student to generate an extended response to a question or other stimulus. For purposes of HSA, it is assumed that a student would have 30 minutes to prepare and present a response. This item type can be used in a number of ways:

- to assess process, e.g., impromptu writing, by asking a general question which draws on the student's own experience and thoughts.
- to assess a student's command of specific facts and the ability to analyze and reason about those facts, e.g., compare and contrast two literary figures. This variation might be used with any of the first three design options.
- to assess a student's ability to comprehend multiple stimuli and to integrate those stimuli in constructing a response to a probing question, e.g., reading excerpts from a number of historical documents and using them to answer a question about a historical trend. This variation is cited in the description of the "Combination" design.
- to assess a student's ability to work with a complex problem and to generate a model or other solution to a particular problem or issue, e.g., a mathematical application to a "real-life" problem. This variation is cited in the description of the "Prep Plus" design where students would be given a common exercise about a complex problem in preparation to the "on demand" assessment activity.

2. Brief Constructed Response -
This item type provides students with the opportunity to generate brief answers to a question. This item type can be used to:

- assess factual knowledge
- assess a student's ability to provide an explanation
- assess a student's ability to solve a multi-step problem
- assess a student's ability to display responses/solutions to intermediate steps in a complex problem
- assess a student's ability to complete an idea or statement implicit in the stem of the question.

3. Machine-Scorable Constructed Response - This item type, sometimes called "grid-ins," can be used in mathematics or sciences for questions, the answers to which are numeric or simple algebraic expressions. These numeric or simple algebraic answers can be gridded on a machine-scorable answer sheet which has been tailored to the particular assessment or test.

4. Selected Response - This item type can be used to test a wide range of knowledge, application, and reasoning skills by asking a student to discriminate among a variety of nuanced alternatives and to identify the "best" or most appropriate alternative in response to the question. Although selected response format can be used to assess factual recall, it is more appropriately used to test a student's ability to make inferences, to judge logical relationships, or in other ways to go beyond the information presented. The most common form of this item type is the 4 or 5-choice question but it can also take the form of matching two sets of information or of selecting the correct response from a lengthy set of alternatives. Frequently, a set of selected response questions can be organized to probe a single stimulus such as a table of data, a graphical figure, or a literary passage.

5. Implications for Development of Performance Tasks - Three of the design options incorporate performance tasks. Performance assessments are often defined by *what they are not* - they are not multiple-choice items, they are not dependent primarily on recall, they do not emphasize content over skills and processes. Unfortunately, such definitions falsely imply that multiple-choice items do rely on recall and content alone. Performance assessments can also be defined by positive terms - they are said to be more complex, more authentic, less structured, more suited to alternative approaches to a solution, require physical manipulation of objects, and stress process over final solutions. The performance assessments used in education vary in the extent to which they have these characteristics. Some of the more open types of performance tasks are not appropriate for high stakes tests such as the HSA. However, many other types of performance tasks, ranging from multi-step, complex extended tasks to shorter constructed response items can be used in the HSA with care. Test specifications teams will be created by MSDE and work with contractors and department staff to design detailed specifications for each assessment during the second Phase of the design project. Traditional tests development guidelines for large-scale testing programs must be modified to incorporate performance assessments while maintaining standards of quality and fairness. There are several different principles to be observed in designing authentic performance tasks:

- Model Construction Principle - Does the task create the need for a model to be constructed, modified, extended or refined? Does the task involve constructing, explaining, manipulating,

predicting or controlling a system? Is attention focused on embedded patterns and regularities, rather than surface features?

- Simple Prototype Principle - Is the situation as simple as possible, while still creating the need for a significant model?
- Model-Documentation Principle - Will the response require students to explicitly reveal how they are thinking about the situation? Will it elicit evidence of the kind of system (mathematical objects, relations, operations, patterns, regularities) they are thinking about?
- The Self-Evaluation Principle - Are the criteria for assessing the usefulness of alternative responses clear? Will students be able to judge for themselves when their responses are good enough?
- The Model Generalization Principle - Does the model that is constructed apply to only a particular situation, or can it be applied to a broader range of situations?
- The Reality Principle - Could the events described actually happen in a real life situation? Will students be encouraged to make meaning based on extensions of their own personal knowledge and experiences? Will students' ideas be seriously considered of must they conform to rigid scoring rubrics about a limited number of correct responses?

These and other principles can be used to constrain the construction of performance assessment tasks. Designing such tasks presents a number of unique challenges. One challenge is to strike a balance between structuring tasks so that they are open-ended enough to require a student to generate a solution while, at the same time, providing sufficient direction about how to go about the task. Another challenge is to formulate questions or procedures that lead students to document what they are doing. Students generally enjoy the doing and not the writing with any extended tasks; thus it can be difficult to motivate students to document steps and procedures. Preliminary scoring guides are needed before develop proceeds very far. Developing these scoring guides is as time-consuming as developing the tasks themselves, but they are critical to the success of the task. Careful consideration of what students are expected to produce, the scoring rubrics that will be used, and the scoring systems and methods (1 rater vs. 2 raters) must be incorporated into the test development procedures and not left until it is time for scoring.[1]

C. Design Options

*1. **Portfolio Plus*** is based on the premise that a great deal of information about a student's knowledge of and ability to do the things specified in the Core Learning Goals is available in the body of work done throughout the year. This option proposes that such information be collected in a portfolio that would be scored by the classroom teacher. The information derived from the portfolio would be combined with a student's performance on a timed assessment given under standardized conditions. This option provides for a wide sample of a student's work/performance to be used in evaluating achievement of the CLGs; it also provides an avenue for the classroom teacher's judgment to affect the final pass/fail decision about a student. It is anticipated that the timed portion of this assessment would include two Extended Constructed Response (ECR) questions, a number of Brief Constructed Response (BCR) questions and a substantial number of

[1] Adopted with permission from Enright, Mary (1993) *Approaches to Performance Assessment Task Development,* Center for Performance Assessment Working Seminar Report (WS1), Educational Testing Service.

Selected Response or multiple-choice questions. A basic difficulty with portfolios is getting comparable examples of student work into the portfolio to start with.

PORTFOLIO PLUS	
PRO	CON
• Samples student performance at different times • Reinforces curriculum reform objectives • Provides for assessing CLGs through different styles of student work, e.g., prepared vs. on-demand • Encourages student reflection on own learning • Wide opportunity to demonstrate SFS • Portfolio can be used as stimulus for part of on-demand assessment • Heightened capacity to deal with presentations or stimuli in a variety of media • Permits some teacher/student choice • Scoring portfolio helps teacher focus on key expectations • Provides opportunity for student to show a variety of works/performances and to demonstrate growth • SR permits test to sample widely among goals and expectations • Assesses complex goals • High content validity • Shows assessment is ongoing and not a single event • Portfolio component may provide early warning • Portfolio component may provide direct evidence of opportunity to learn	• Impossible to standardize conditions for creating portfolio portion • Need to develop a model/paradigm to structure a portfolio in each discipline • Need to develop scoring tools to guide portfolio scoring • Logistically difficult for teachers to manage • Takes time away from direct instruction • Requires training of teachers to score portfolios • Requires substantial teacher time to score portfolios • Expensive to score on-demand test • Long turnaround time • Complex process to combine portfolio and test information into single score • Very difficult to achieve needed reliability of scores • Need to build evidence of validity of portfolio scoring process • Difficult to accommodate modular design

2. ***Preparation Plus*** is based on the idea that students should share a common learning experience prior to assessment and that the learning experience can provide the substantive basis for a portion of the assessment. The common assignment or learning experience would model the kinds of instructional strategies that are advocated for the particular subject area. This common learning experience makes it possible for the timed, standardized assessment to draw on more complex stimuli (e.g., a "messy" real-world mathematics application) than would otherwise be possible within the time limits of the assessment administration. It is anticipated that the timed portion of this assessment would include an Extended Constructed Response question based on the preparatory assignment, a second Extended Constructed Response question, a number of

Brief Constructed Response questions, and a substantial number of Selected Response or multiple-choice questions.

PREPARATION PLUS	
PRO	CON
• Provides common learning experience as rich stimulus for assessment • Reinforces curriculum reform objectives • Provides opportunity for student to construct a response in addition to recognizing the correct response • Permits students to make a "considered" response to ECR because of preparation • Provides heightened capacity to deal with presentations or stimuli in a variety of media • SR permits test to sample widely among goals and expectations • Preparatory activity models desired type of learning experience • Assesses some complex goals • High content validity	• Requires class time for preparation in addition to "testing" time • Difficulty in ensuring equal preparation across classrooms • Expensive to score on-demand test • Long turnaround time • Difficult to achieve adequate reliability of scores • Questionable whether possible to assess "group" work as part of requirement • Increases difficulty of dealing with absenteeism • May require new equipment/material for preparatory activity • Difficult to accommodate modular design • May need to demonstrate opportunity to learn

*3. **Combination*** reflects a commonly used measurement technology that combines Constructed Response questions with a significant number of Selected Response questions to create a reliable assessment instrument. Within the time limits of the assessment administration, this option would include an Extended Constructed Response question that draws on a number of brief documents, laboratory results, or other authentic stimuli plus a second Extended Constructed Response question as well as a number of Brief Constructed Response questions, perhaps based on the documents, and a substantial number of Selected Response or multiple-choice questions.

COMBINATION	
PRO	CON
• Provides opportunity for student to construct a response in addition to recognizing the correct response • Reinforces curriculum reform efforts • SR permits test to sample widely among goals and expectations • Testing time is only impact on school/class schedule • Assesses some complex goals • High content validity	• Limited stimulus for ECRs due to time constraints--although this may make the assessment more accessible to some students • Expensive to score on-demand test depending on mixture of item formats • Long turnaround time • Difficult to achieve adequate reliability of scores depending on mixture of item formats • Difficult to accommodate modular design • May need to demonstrate opportunity to learn

*4. **Limited Combination*** is an entirely machine-scorable assessment. Challenging multiple-choice questions would be the predominant item type, although in the sciences and in mathematics a machine-scorable Constructed Response item format would also be used.

LIMITED COMBINATION	
PRO	CON
• SR permits test to sample widely among goals and expectations • Moderate scoring costs • Moderate turnaround time • Testing time is only impact on school/class schedule • Readily achieve adequate reliability of scores • Can accommodate modular design • Readily adaptable to computer-based testing	• Limits skills and performances that can be tested • Less supportive of reform curriculum • Emphasizes recognition over construction • Content validity limited to certain aspects of CLGs • May need to demonstrate opportunity to learn

D. Scoring

The first three options include performance assessments in design. Scoring these questions presents a myriad of questions and concerns. First, there is a tension between the desire to identify and report varied, meaningful information, and the pressure to reduce information to a single score to serve policy needs. The successful use of performance assessments for policy-making or decision-making purposes requires developing ways to present complex assessment results in manageable and simple ways. Unfortunately, many of the benefits associated with performance assessments can not co-exist with high stakes uses. For example, large-scale performance assessments, like most large-scale assessments, typically fail to provide much diagnostic information to teachers and students (Gearhart, 1995) and rarely include more opened ended performance tasks that can be solved by many alternative solutions. Second, it is much more difficult, time-consuming and expensive to reliably score complex performances than to score short, simple responses. The more varied the set of possible responses, the more difficult it is to establish scoring rubrics and to train scorers. For example, in the field tests for the NAEP science assessments, nearly as much time was spent in training scorers as in actually scoring the assessment. This is equally true of other assessment programs where reliable scoring is required because of high stakes uses (e.g., HSA). Finally, a tension between reliable scoring and validity can develop. Those factors that increase reliability (shorter responses, convergent answers, highly concrete or specific scoring rubrics) may limit the types of performances that can be assessed. Thus the validity of the assessment as a measure of deeper understanding and more complex and varied skills will often be compromised because of the need to obtain reliable scoring. As one would expect, the proposed use of the HSA for individual, high stakes decisions, requires the most stringent levels of psychometric integrity for the assessments and the highest levels of reliability that can be obtained for performance assessments. [2]

[2] Adopted with permission from Enright, Mary (1993) *Approaches to Performance Assessment Task Development*, Center for Performance Assessment Working Seminar Report (WS1), Educational Testing Service

The Extended Constructed Response items require a student to produce work, such as extended essays, respond to multiple parts of a question that may be interrelated, and/or create graphics (e.g., tables, figures) along with written responses to problems. Usually two or more trained scorers are required to review each item. Currently, such items require substantial time and resources to ensure accurate scoring. Many separate steps are required to score such item types, each step involving several smaller steps, with quality control procedures to ensure accuracy of scoring. Some of the steps include:

- Collecting the papers and transporting them to a central location.
- Having a small group of scorers or scoring leaders review a sample of papers to establish and revise the scoring standards and to identify exemplar responses at each of the scoring levels.
- Copying and collating papers, organizing scoring (which scorers will review which papers), and incorporating additional quality control procedures.
- Training the scorers and reviewing the quality of the scoring throughout the process, coaching individual scorers as required.
- Actually reading and scoring all the papers (probably two readers will be required at a minimum for each paper).
- Collecting the score sheets and incorporating scores into a database.
- Translating the scoring levels to proficiency levels (1-5) which will be reported to students and schools.
- Equating different forms of each exam to ensure they represent the same level of difficulty across students and years.
- Generating and distributing score reports to students and schools.

With over 60,000 ninth grade students enrolled across Maryland, the ninth grade English test alone would result in 120,000 essays that would need to be read by two readers. Multiplying the number of students (across grades and subjects) by the number of different tests required, one might have well over one million Extended Response items to score each year. Currently, major testing programs which include such item types (e.g., Advanced Placement, MSPAP, NAEP) require anywhere from 7 to 15 weeks for score turnaround. Each of the first three designs also includes a number of shorter Constructed Response items that must follow generally the same scoring process and would add to the overall scoring burden. These designs also include a significant section of multiple-choice items that can be machine scored.

The scoring of the volume of assessments projected under the HSA will place a large burden on MSDE. The need for exceedingly high standards for the quality of scoring (since high-stake individual decisions will be based on the results and legal risks will emerge) and economical affordability further complicates the task.

A phase-in of the assessments may be required simply to deal with the volume of constructed response questions to be scored. It is questionable whether MSDE, or any large contractor they may enlist, could, in the first year or two, successfully score the volume and variety of performance assessments generated across the 12 assessments with a sufficient degree of quality and reliability. The operational requirements for such large-scale scorings strongly support a phase-in strategy until the processes are well-tuned and refined. The scoring costs associated with any of the first three designs will be substantial given the requirements for quality. Once MSBE arrives at a decision regarding the general design of the HSA, MSDE should explore

various options for accommodating the scoring demands that will be imposed. A feasibility study should examine:

- whether any contractor currently has the capability of scoring this volume of assessments within the proposed timetable,
- establishment of a Maryland State Scoring Institute located at one or more higher educational institution, and
- whether there is any viable model for involving Maryland teachers in the scoring, given the constraints (e.g., released time, costs, time commitment for quality scoring) and timetable (if scores are provided in the summer this option would appear more feasible than if scoring must be done during the school year).
-

Throughout the public engagement process, Maryland educators stressed the benefits of having teachers involving in the reading and scoring of the HSA. Many educators believe that teacher scoring is essential for their "buy-in" and the eventual success of the program. One reason for this is that many educators believe that the training that is incorporated in the scoring provide teachers with some of the most rewarding and valuable staff development opportunities available. There is also evidence that as teachers work through the scoring they gain additional insight into the content and performance expectations that improve their instruction. Teachers recognize the characteristics of student performance that are most important, and the curriculum's goals and indicators are instantiated during the scoring process.

The fourth design option is comprised exclusively of items that can be machine scored (optically scanned) such as various types of multiple-choice items and grid-in items. This option would likely require a minimum of four to six weeks for score turnaround, since there are still many steps involved in collecting and scanning the tests, quality assurance, equating forms, creating a database of scores, and producing/disseminating individual and school score reports.

E. Successful Use of Performance Assessments for High Stakes Purposes

Research suggests that many performance assessments as presently designed do not produce valid student level scores on a large-scale basis. It may be best to recognize that single assessments, either norm-referenced multiple choice or performance based, do not well serve multiple, high stakes needs (Gearhart, 1995). However, during the past five years research has increased our knowledge of potential problems in using such assessments for high stakes purposes. Successful strategies have been developed such as using well-designed rubrics, more specific scoring procedures to measure students skills (Rogosa, 1995). All designs proposed above combine performance tasks with other forms of test items that are generally more objectively scored and thus can increase the overall reliability of the assessment score. Such designs may not capture the true flavor and benefits that proponents associate with assessments which are exclusively performance based, but such assessments can not yet be used reliably for high stakes decisions.

Just as there are unique strengths associated with performance assessments, there are some unique criticisms of these assessments. For example, a small but vocal group of participants in public engagement activities objected to the use of performance tasks on any high stakes testing program used for accountability. Similar concerns and misunderstandings about performance tasks have

been expressed in other state assessment programs. For example, Kirst and Mazzeo (1996) recently cited the criticism and rumors from conservative groups and parents about performance tasks as one of the reasons for the demise of the California Learning Assessment System. They noted that rumors of objectionable content, subjective scoring standards, and criticism of the content selected for the assessments all contributed to the unraveling of the assessment. To overcome such misperceptions, greater participation of diverse groups, including parents, must be part of the test development process. Sample test items and performance tasks might be released both prior to the assessment and at the time that results are reported. Demonstrating connections between performance tasks and multiple choice items may also help the public to link the scores to student strengths and weakness in the knowledge and skills called for by the CLGs (Weigle and Wakai, 1995). Releasing all items is particularly problematic for the security of performance assessments since entirely new forms would be required. This would be prohibitively expensive and would also constrain psychometric procedures to equate and pre-test tasks for new test forms. Instead, a public review panel can be created who would review all items and tasks prior to their use. If the group includes appropriate stakeholder groups it may help to alleviate concerns about the content of the tasks. This strategy, as well as releasing sample items and illustrations of student performance at various proficiency levels (and across subjects) can help communicate the results from performance assessments in appropriate manners while minimizing negative consequences to the testing program.

Often it is the divergent priorities and tensions between political and technical requirements that can undermine a large-scale assessment program. In California, there was agreement among policymakers and testing experts on the benefits of performance-based testing, yet policy makers subordinated the technical realities to political demands in establishing an overall optimistic timeline for implementation. Because the traditional needs for validity and reliability are more problematic with performance assessments, using such assessments for high stakes purposes raises the ante considerably on the technical and cost issues (Kirst and Mazzeo, 1996). Establishing innovative and high stakes state assessment systems typically require expending substantial amounts of political capital. There are always a range of opponents ready to seize upon any perceived or actual problems with the assessments. Many state assessment systems do not survive multiple challenges or several large and visible missteps. Large-scale assessments, particularly those involving performance tasks and other innovative components require substantial time to get things right. MSBE should consider the demands placed on the HSA when discussing the phase-in of the program as discussed in Section I, e of the report.

F. Criteria For Evaluation

*1. **Academically Rigorous*** - The goal of the Maryland School Improvement Program is to raise the expectations of students, teachers, school, family, and community. The High School Assessments are a critical element in conveying the "new" expectations, and their rigor must be consistent with this purpose. Each design option can be evaluated in terms of the support it provides for the Core Learning Goals, desired pedagogy, demonstrating Skills for Success, and encouraging the kinds of teaching and learning that will develop the knowledge and the skills of application envisioned by the School Improvement Program.

*2. **Professionally Acceptable*** - Any assessment that is used for making high-stakes (life-changing) decisions about individuals must meet certain widely accepted psychometric standards. These are documented in *Standards for Educational and Psychological Testing* (1985). Each assessment must be evaluated in terms of its "validity," i.e., does it measure the Core Learning Goals and Skills for Success it purports to measure; and its "reliability," i.e., does it measure a student's ability with enough precision and consistency that one would get the same results if the student were tested on another day or with another form. In addition, because it is anticipated that the HSA will involve multiple forms of each assessment, the forms must be capable of being linked statistically (calibrated and/or equated) so that the reported scores from one form mean the same thing as the reported scores from another.

The scoring of the Constructed Response questions must be done in a professionally responsible manner that minimizes variation among readers and provides means for monitoring the consistency of the scores assigned.

*3. **Practically Doable*** - The success of any assessment program depends on the practicality of the myriad details connected with moving an assessment from the minds of its creators to the actual administration of the assessment to students to the scoring and reporting of the results in a timely and useful fashion.

- Pre-Administration Processing - MSDE and/or the contractor must be able to manage anticipated volumes by school and by assessment, produce and print the assessments and any ancillary materials, and procure and distribute any required nonprint materials or equipment.
- Impact on the School - Each school needs to consider steps such as the maintenance of test security, disruption of schedules, impact on teaching time, availability of teachers to administer the assessments, accommodations for students with special needs, make-up administrations, adequate equipment, and retrieval and return of all assessment materials in a timely fashion.
- Post-Administration Processing - MSDE and/or the contractor must make provisions to receive and account for all assessment materials, score all scannable documents, arrange and manage the scoring of Constructed Response questions, control the quality of the resulting data, combine machine-scored and human-scored data with student and school identifications, perform the necessary equating, scaling, and standard-setting operations, and produce both individual and aggregated reports.
- Long-Term Records - MSDE and/or the contractor must set up a database for maintaining student records over the several years during which a student is in high school--a period of 5-6 years minimum. This database must be able to track retests, identify subjects in which requirements have been satisfied or not, and provide a transcript service to the high schools that students attend during their senior year.

*4. **Legally Defensible*** - Whenever assessment instruments are used to make high-stakes decisions about individuals and some of those individuals are denied a social good, such as a diploma, on the basis of their assessment results, litigation can be expected. The question of legal defensibility has two elements: the assessment itself and the opportunity to learn the content and abilities measured by the assessment. These both impinge on the High School Assessment program.

It is important to be able to demonstrate that each assessment was developed and analyzed using generally accepted procedures and that steps were taken to eliminate performance variation due to anything other than the construct(s) being measured. The professional standards described above are important evidence in defending the soundness of the High School Assessment.

It is equally important to be able to show that the plaintiff (the student) had the opportunity to acquire the knowledge and the ability to perform that which is included in the Core Learning Goals measured by the particular instrument. This burden will fall most heavily on the school and district that the student attended.

*5. **Economically Affordable*** - Any public activity must be designed so that it can be sustained financially over a mid- to long-term implementation. The costs for HSA can be analyzed in several categories:

- Start-Up Costs - These are one-time or infrequent costs such as the creation of a processing system and a student database, initial activities to inform districts, schools, students, parents, etc.
- Annual Costs (Assessment) - These include the costs of developing one or more forms of the assessment each year, the development of scoring guidelines and training tools specific to each form, the statistical work needed to equate/scale/and set the standard for each new form, the distribution of assessment descriptions and sample questions, special training for teachers, etc. The total of this category would be the sum of the annual costs associated with each of the 12 assessments. Because these are recurring costs, they must be budgeted for each subsequent year.
- Annual Costs (Student) - These are the variable costs incurred because of each additional student who takes a particular assessment during the year. They include production, printing, shipping, scoring, and reporting costs. The total annual cost in this category would be the cost per student for each assessment multiplied by the number of students taking that assessment and summed across the 12 different assessments. These costs recur each year but will vary with the number of students who take each assessment in a particular year. **These costs are discussed in Section II of the report.**

In addition to the costs incurred by the State for the functions described above, there will be both "out-of-pocket" and opportunity costs incurred at each school in administering the HSA.

TABLE A1 ANALYSIS OF GENERIC DESIGN OPTIONS

	PORTFOLIO PLUS	PREP PLUS	COMBINATION	LIMITED COMBINATION
Academically Rigorous	--Provides very strong support of reform efforts, including pedagogy, CLGs, and the development of SFS --Models integration of assessment with instruction --Maximizes teacher input to curriculum and outcomes --Provides a variety of ways for students to demonstrate competence on CLGs and SFS	--Provides strong support for reform efforts, CLGs, and the development of SFS --Provides a limited model of integration of assessment with instruction --Provides a number of ways for students to demonstrate competence on CLGs and SFS	--Supports CLGs and SFS --Provides a number of ways for students to demonstrate competence on CLGs and SFS	--Supports certain aspects of CLGs --Provides a limited array of ways for students to demonstrate competence on CLGs
Professionally Acceptable	--Requires development of new training tools for portfolio use and scoring --Combining such diverse sources of information creates many psychometric challenges--If new statistical processes are developed, still will have no assurance that the results will be comparable over administrations and forms	--Will require substantial work to ensure adequate reliability --Training and monitoring of scorers is critical	--Will require substantial work to ensure adequate reliability --Training and monitoring of scorers is critical	-SRs can be scored with high reliability; enough SRs on test will produce an adequately reliable score -Content validity will be limited to certain aspects of the CLGs
Practically Doable	--Use of portfolio will impact the entire course and will create difficulties for absentees or transferees --Managing portfolios can be substantial task for teacher --Issues of plagiarism difficult to monitor with portfolio --Student information from portfolio will need to be matched with results of timed assessment --Security of timed assessment can be maintained prior to administration	--Requires several class periods for preparatory activity --Preparatory activity complicates coping with absentees or transferees --Security of timed assessment can be maintained prior to administration	--Only assessment period will impact school schedule --Security of timed assessment can be maintained prior to administration	--Only assessment period will impact school schedule --Security of timed assessment can be maintained prior to administration

	PORTFOLIO PLUS	PREP PLUS	COMBINATION	LIMITED COMBINATION
Economically Affordable	--Substantial local costs for teachers to be trained in using & scoring portfolios --Substantial local costs for teachers to score portfolios --May require purchase of new equipment, e.g., calculators	--May require purchase of new equipment, e.g., calculators	--May require purchase of new equipment, e.g., calculators	--May require purchase of new equipment, e.g., calculators

II (PART B). THE DESIGN OPTIONS: SUBJECT-SPECIFIC COMMENTS, EVALUATION, and ILLUSTRATIVE QUESTIONS

In this section, the "Comments and Evaluation" identify for each subject area where the general comments made above about each design option do not apply to the particular subject or where there are additional constraints or considerations imposed by the nature of the Core Learning Goals for the subject. These "Comments and Evaluation" should be read as exceptions or additions to the general comments.

The "Illustrative Questions" provide examples of the kinds of questions referred to above to be used in the design options for each of the subject areas. Note that they are taken from a variety of sources and would be used as models during the development of detailed specifications for each assessment.

A. ENGLISH

Comments and Evaluation

*1. **Portfolio Plus*** - This model allows the maximum flexibility for districts and teachers, yet it is very difficult to implement for high-stakes assessments. Logistics, scorer reliability, issues of out-of-class work, the lengthy scoring process, and variations across students make this design problematic for any high-stakes assessment where individual student-based decisions are required. In order to implement a portfolio system for individual high-stakes decisions, one would need to highly constrain the tasks, projects, and time line for completion and scoring, or move to a certification program that would permit students to demonstrate competency on CLGs in ways that could not be equated (to derive comparable scores) across districts (e.g., alternative forms of evidence of accomplishment that are not psychometrically linked).

*2. **Preparation Plus*** - Members of the content team prefer this model because it combines authentic tasks based on prior preparation and more objective item types that can result in higher levels of reliability for high-stakes uses. Prior preparation will reinforce elements from the CLGs and the national and statewide reform efforts in English. The preparation activity must be carefully scripted so all students receive the same preparation. The preparatory work should be done either individually or as a whole class, not in small groups. The time for preparation should be specified and be the same for all schools and classrooms. There are still cost, reliability, and logistical concerns associated with the design if used for individual high-stakes decisions. This model most closely resembles the College Board's Pacesetter English course and assessments which are used in several districts across the state; examples are provided below from the Pacesetter materials.

*3. **Combination*** - Members of the content team find this model acceptable. This model most closely resembles the Advanced Placement Examinations which combine constructed response items (essays in the case of the English Literature and Language examinations) with multiple-choice items to result in one overall grade for students. The model presents many of the advantages of Preparation Plus, but still carries substantial costs and scoring burdens.

4. Limited combination - There is no grid-in response available for English. This option would involve either a straight multiple-choice test or, possibly, adaptation of PRAXIS reading comprehension item types (currently machine scorable only in computer-delivered exams). The content team was strongly opposed to this design model because it is inconsistent with the direction of national and statewide reform in English and would not be acceptable to the majority of English teachers across the state who would see this as a major step backward.

Illustrative Questions:

1. Title: Portfolio (English)
1.1. Item Text- See below for some ways of organizing portfolios.

There have been enormous problems in scoring portfolios reliably in the statewide testing program in Vermont. Reliability in scoring may be improved by clarifying the scoring system (in Pacesetter, the scoring is based on dimensions, which are reprinted below) or including common elements, tasks, and curriculum experiences (common tasks in Pacesetter include writing exercises such as creating a director's notebook or a shot list or conducting and recording an interview; domain projects are emphasized in Arts PROpel).

1.2. This item links to the following Goal-Expectation-Indicators:

All of the CLGs could be accommodated within a portfolio, depending on what assignments teachers give and what exercises by students are included in the portfolio. The portfolio would allow students to include evidence of the reading, writing, and speaking processes (Goal 1.1.1-4, 2.2.1-6) in a much fuller way than the other item types through such exercises as reading logs and journals and revision and reflection on writing and oral presentations. Portfolios allow for more extensive research (1.3.1-5) and use of technology than the other item types, and would probably allow for fuller evidence of experience with visual media. Other goals not fully covered in the other item types, such as those emphasizing personal response and use of particular literary approaches (e.g., 1.2.5, 1.2.6, 1.3.2), could be targeted by specific assignments.

Those goals having to do with connotations of words and with dialects (3.2.2-5) could also be dealt with in greater depth through targeted assignments. Goal 4 throughout, which demands reflection on the student's own writing and transformations of that writing, is probably best addressed by untimed assignments such as those that would go into a portfolio.

1.3. This item is appropriate to the following Skills for Success:
Virtually all of the Skills for Success in Learning, Thinking, and Communication can be tested through a portfolio. Those for Interpersonal Skills (except probably 3.3 through 3.6) can be tested if group work is used extensively and if students work with several groups in the course of the year. Of the Technology Skills, 2.1 through 2.3, 3.1, and 3.2 are the likeliest to be addressed by assignments in an English course.

1.4. Item's source:

Reference to Pacesetter and Arts PROpel portfolios

Scoring guide: See Pacesetter portfolio dimensions below, on which scoring guide for portfolios is based.

Pacesetter Course Dimensions
(Note: not published/released; do not circulate or copy)

Dimension 1: Making Meaning from Texts: ***Students understand written, oral, and visual texts from a variety of times and cultures in a variety of media and genres.***

Aspects of Making Meaning:

- Respond to texts: Students respond to texts in terms of their own cultural backgrounds and personal experiences. They present their own impressions, opinions about, and predictions related to texts--characters, events, ideas, views, emotions, and language.

- Interpret and analyze texts: Students interpret the meaning of text. They analyze the effect of the voices, literary elements (such as form, organization, imagery, word choice, language and details), and film techniques (such as staging, choice of image, music, sequence of shots and lighting). They evaluate the effectiveness of others' texts.

- Put texts in context: Students make connections between the text and other texts, fictional characters, real people, current events, cultures, and recurring themes. Students examine historical, cultural, and geographical influences on authors and their texts, as well as the setting of texts, and explain how this information helps with the understanding of texts.

- Reflect on and evaluate processes for making meaning: Students reflect on and evaluate strategies (such as reading aloud, discussing, taking notes, underlining, looking up or figuring out the meaning of unfamiliar words, etc.) they use to explore texts, to develop their own ideas about texts, and to understand texts. They set goals for improving how they make meaning from texts.

- Work with others: Students collaborate with others to understand texts. They participate in group decision making, taking on different roles at different times and planning so that all group members are actively involved. Students accept responsibility for their agreed-upon parts of projects, and help the group stay on task and meet schedules. They critique texts of their peers in terms of both content and technique and make constructive suggestions for improvement. They encourage group members to share divergent views and listen thoughtfully to the suggestions and ideas of others.

- Demonstrate growth in making meaning: Students demonstrate increasing willingness and ability to respond to texts thoughtfully, to analyze texts in ways that contribute to understanding, to interpret the meaning of texts, to place texts in various contexts that

enhance meaning, to evaluate and reflect on their own work in making meaning, and to work collaboratively with others.

Dimension 2: Creating and Presenting Texts: ***Students communicate ideas through oral, visual, and written texts, in both informal and formal modes of presentation.***

Aspects of Creating and Presenting Texts:

- Use their own voices: Students communicate in a variety of their own voices, appropriate for a range of purposes, reflecting both their own culture(s) and unique points of view.

- Develop and present texts: Students communicate in a variety of genres, media, and forms. They develop texts using a variety of strategies such as use of resources, anecdotes, examples, reasons, quotations, and questions. They use language effectively. They create focused and coherent texts and tailor presentations for various purposes.

- Demonstrate technical command: Students use oral and written language effectively and precisely. Their texts employ grammatical usage, sentence and paragraph structure, spelling, and punctuation that are appropriate for the intended purpose. Some texts are polished to meet the standards and expectations of academic and public audiences.

- Reflect on and evaluate how their own texts are created and presented: Students reflect on and evaluate how they develop oral, visual, and written texts and the effectiveness of their presentations. They set goals to improve both the texts they present and the processes they use to create them.

- Work with others: Students collaborate with others to design, develop, present, or perform both individual and group texts. They participate in group decision making, taking on different roles at different times and planning so that all group members are actively involved. Students accept responsibility for their agreed-upon parts of projects, and help the group stay on task and meet schedules. They encourage group members to share divergent views and listen thoughtfully to the suggestions and ideas of others.

- Demonstrate growth in creating and presenting texts: Students demonstrate increasing skill and confidence in the processes they use to develop texts and the strategies they employ to communicate their ideas. They develop a range of voices, including those used in school and in other settings, that they use appropriately depending on the purpose of the text. They demonstrate increasing willingness and ability to reflect on and evaluate the creating and presenting of their own texts, and to work collaboratively with others. Their communications are increasingly mature and technically sound.

2. Title: Pacesetter Parts One and Two (for Prep Plus) (English)
[NOTE: Modeled on an unpublished Pacesetter examination. DO NOT CIRCULATE.]

2.1 Item Text: (see attached): This item links to the following Goal-Expectation-Indicators:

1.1.1 through 1.1.5
1.2.1 through 1.2.6
1.3.1, 1.3.3 through 1.3.6

2.1.1 through 2.3.3
2.2.1 (possibly 2.2.5)
2.3.4 and 2.3.5

3.1.1, 3.1.5 through 3.1.7, 3.1.9
3.2.1, 3.2.3, 3.2.4
3.3.1 and 3.3.2

4.1.1 and 4.1.2
4.2.2 and 4.2.4
4.3.2 and 4.3.4

2.2. This item is appropriate to the following Skills for Success:
Learning:
1.2.3, 1.3.2, 1.3.5, and 1.3.6 (limited), 1.4.3, 1.4.5, 3.2.3, 3.3.2, 5.4.2, 5.4.3

Thinking:
1.1.4-1.1.6 (1.1.7 limited)
1.2.3 through 1.2.6
2.1.1 through 2.1.3, 2.1.6
3.1.1 through 3.1.5 (tested in Part Two only)
3.2.2
4.3.4

Technology:
possibly 2.2.1

Interpersonal:
1.1 through 1.3 all, 3.2 all
If group work included: 2.1 all, 2.4 through 2.6 all

Communications:
1.1 all
1.2.1 and 1.2.4
1.3.3 through 1.3.5
2.4.5 and 2.4.6
2.5.1 through 2.5.5
4.1.2 and 4.1.3

2.3. Item's source: Pacesetter. Draft scoring guides included.
(Pacesetter model for PREP PLUS)

PART ONE

Preparatory Portion

Read a selection from *Lord of the Flies* (for example, the early scene on the island when the boys introduce themselves and elect a leader) [would be provided]. Then watch the film versions of that scene from clips you will be shown of the Peter Brook film from the 1960's, in black and white, and the more recent color film (from the 1980's).

Take notes on the selection from the novel, underlining key words or phrases. Which images convey the setting? Which images create a mood?

Now watch each of the film selections three times. Do a shot list for each that shows: visual elements (camera angle [long shot, close-up, etc.]), length of shot, other visual elements (such as lighting, color), sound effects (such as voice-over, music), and other sound elements such as tone of voice.

(Preparation section would be an expansion of this basic outline, probably directing students to look at the films first with the sound off and then with the sound on and to work in groups to do a shot list, and would provide leading discussion questions on the selection from the novel.)

[NOTE: Even with the preparatory portion, this presumes preparation in earlier courses in the study of film techniques. Also, note that permission for using film material is likely to be costly and time-consuming to obtain.]

Timed Portion (40 minutes)

1. Describe two important visual effects in the film. Explain why they are important. (5 minutes)

2. Describe two important sound effects in the film. Explain why they are important. (5 minutes)

3. Discuss an image or phrase that is repeated in the selection from the novel. Why is it important and what is the effect of the repetition? (10 minutes)

4. Which of the two film versions do you think is a more accurate portrayal of the novel and why? Which version (film or printed text) is more effective, in your view? (20 minutes)

SCORING: Each question to be graded on a 4-point scale. For example, a scoring guide for Question 1 might be generally:
4 = Two important visual effects identified, with appropriate explanation.
3 = Two effects identified, but explanations are sketchy or one explanation is missing.
2 = One effect with good explanation, or two effects with no explanation.
1 = One effect with no explanation.
0 = No relevant information.

PART TWO

Preparatory Portion

Almost all of us have had the experience of being a "stranger in a strange land," a newcomer to a place that seems unfamiliar to us. That new place may be a new country, town, school, or even a new experience or new group of friends.

Read the selections by writers describing the experience of immigration to this country [here would be provided 4-6 one-page selections from writers such as Abraham Cahan, Gloria Anzaldua, etc.; could include fiction and nonfiction, poetry, and photography].

Discuss the selections in groups, focusing on the following questions:

- what language does the writer use to convey his or her experience?
- what relation does the writer see between his or her new country and old country?
- what is gained and what lost by moving to a new country?

Think of an experience you have had this year that you recorded or expressed in a written, spoken, or performed work, when you were a "stranger in a strange land." Some examples might be:

- an essay, poem, or story about an experience you had of being a stranger (you may have written a story based on your experience in which you do not appear)

- a play you performed in which you played a character who was a stranger to you

- a new experience you had (such as joining an athletic team, making a speech before a large audience, or acting for the first time)

Find a relevant example from your portfolio and reflect on how you changed as a result of this experience, and how you expressed it. You may wish to write notes on your portfolio piece to underline specific words or phrases that illustrate your points.

(Note: Preparatory portion would be expanded from this, including some more discussion questions on each piece and possibly some work in groups. This exercise assumes that students have been keeping portfolios and reviewing them regularly.)

Timed Portion (40 minutes)

In an essay, describe an experience you had when you were a "stranger in a strange land." How were you changed by this experience? How did you express this experience? Compare your experience and your way of expressing it to the work of one of the writers you read in the preparatory portion.

SCORING: 6-point scoring guide

(Modified version of scoring guide for writing used in Pacesetter in 1995; this adaptation addresses reflective and interpretive as well as writing skills):

6 Effective use of language in conveying ideas. Reflects clear understanding of own work and that of writer discussed. Makes connection between own work and writer discussed. Uses supporting details effectively. Paper is coherent and in strong control of sentence-level features.

5 Appropriate use of language in conveying ideas. Reflects clear understanding of own work and that of writer discussed. Uses supporting details clearly. Paper is coherent and in good control of sentence-level features.

4 Adequate use of language in conveying ideas. Reflects clear understanding of own work and some understanding of writer discussed. Uses supporting details, though they may not be discussed at length. Paper is generally coherent and control of sentence-level features is adequate.

3 Somewhat limited use of language in conveying ideas. Reflects relatively clear understanding of own work, but limited or superficial understanding of writer discussed. Provides few or no supporting details. Errors in sentence-level features are not pervasive and do not interfere seriously with meaning.

2 Very limited use of language to convey ideas. Understanding of own work is conveyed briefly or superficially, and interpretation of other writer may be lacking or largely erroneous. Provides one or no supporting details. Errors in sentence-level features are frequent and interfere with meaning.

1 Seriously limited use of language to convey ideas. As a result, little reflection on own work or that of other writer. No supporting details. Frequent errors in sentence-level features interfere seriously with meaning.

3. Title: Critical Thinking Task (English)

3.1. Item Text: (see attached)

For exercise on Native American mythology: Begin by stating a Native American creation myth in a paragraph or so, then include three brief (page-long) creation myths from other cultures, as well as a painting or sculpture depicting the myth or a character from it. Then provide 5-6 "note cards" (each no more than a paragraph) that provide interpretations or additional information.

Alternative 1: Use same structure as for task on creation myths for a myth or legend that has been transformed through time, such as Helen in Egypt. Begin with a brief selection from the Odyssey, then include Yeats's "Leda and the Swan" and a brief selection from h.d., plus a contemporary prose version.

Alternative 2 (Based on Pacesetter): Use a text that has multiple versions, such as Conrad's *Heart of Darkness*. Use a selection from the beginning which introduces the narrative (e.g., the

first two pages), as well as a brief selection from Conrad's *Congo Diary*, and (if administrative constraints permit) a brief clip from *Apocalypse Now* (if not, maybe the corresponding selection from the screenplay). Also include Nigerian writer Chinua Achebe's essay on Conrad (which severely criticizes the colonial perspective in *Heart of Darkness*) (35 minutes for reading and viewing)

Questions: (perhaps 5-10 multiple-choice questions, 5 minutes)
Achebe's principal objection to Conrad's book is that
* it presents Africa entirely from a European perspective

In the selection given from *Heart of Darkness*, the point of view is that of:
Joseph Conrad
* the unnamed narrator
the character Marlow

The principal images in the selection from *Heart of Darkness* describe
* the river and the ocean
mind and heart
Africa and Europe

Short answer: Name one similarity and one difference you see between Conrad's *Congo Diary* and *Heart of Darkness*. (10 minutes)

Long essay (40 minutes): Francis Ford Coppola set his film *Apocalypse Now* in Vietnam at the height of the Vietnam War (late 1960's). What techniques does he use that echo Conrad's? Why do you think he set his movie version of *Heart of Darkness* in the time and place he did?

Alternative 3: Same as above, but text is a story or essay by a contemporary Native American writer (e.g., Simon Ortiz, "The Power of Language"). Accompany it with documents that illustrate the point about the way language determines identity:
- editorial by Native American writer about romanticization of Native Americans
- photograph of Native American schoolchildren from early 20th century
- letter from Native American schoolchild to his or her parents
- a Native American myth referred to by Ortiz

Multiple-choice questions: ask about specific images in the Ortiz essay
Short answer: compare Ortiz's account of myth with the original myth
Essay: Agree or disagree with Ortiz's view of his education, using two or more of the accompanying documents

3.3. This item links to the following Goal-Expectation-Indicators:

1.1.1, 1.1.3, possibly 1.1.4 and 1.1.5 (if visual media [movies, photographs, other pictures] are used as stimuli)
1.2.1 through 1.2.4 and possibly 1.2.5 and 1.2.6 (if visual stimuli used)

1.3.1, 1.3.3, 1.3.5 and possibly 1.3.6
2.1.1 or 2.1.2, and 2.1.3 or 2.1.4
2.3.4 and 2.3.5
3.1.1, 3.2.1, and 3.3.1
4.1.1 and possibly 4.1.2, 4.3.4

3.4. This item is appropriate to the following Skills for Success:

Learning:
1.4.3, 3.2.3, and 5.3.2

Thinking:
1.2.3, 2.2.1-4 (especially if research-like task is emphasized), 4.1.1 through 4.1.3, 4.3.4

Technology:
possibly 2.2.1

Communication: 2.1.1 through 2.1.3, 2.3.2 through 2.3.5, 2.4.5 and 2.4.6

3.5 Item's Source: Tasks in Critical Thinking, Pacesetter, NAEP Reading

See scoring rubric below from Tasks in Critical Thinking. Note: Core scoring used for Tasks in Critical Thinking. The "core" (passing or proficient score of 4) is defined first; the other scores are seen as exceeding or falling short of a 4 in various degrees. Responses are judged on up to three dimensions: inquiry, analysis, and communication. Each of the alternatives proposed for this task would address inquiry (especially if focus is on research), analysis, and communication. A 6-point scale would be used for all Extended Constructed Responses, as shown in the scoring guide below.

RESEARCH TASK (Note: not for circulation. Adapted from unpublished Task in Critical Thinking)

For this task you are to assume the role of someone preparing to write a research paper for class. In such a paper, a writer begins by finding reliable sources of information related to the topic of the paper. Notes taken from these sources provide the writer with material to develop and support his or her ideas. As you follow the steps of the task, you will be asked to

> read the note cards included with the task. The note cards are in a packet in your envelope. They contain the information you will make use of in your report. You will:
>
> organize the note cards into groups;
> decide on the central idea that you will develop in your report; and
> write a first draft of your report.

Be sure to pace yourself to allow enough time for each step in the process, from studying the note cards to organizing the information on them to writing the draft. Write all your responses so that they can be read and understood easily by someone who does not know you or your handwriting.

Assignment: You have been assigned to write a paper dealing with creation myths. As you look through several books in the library, you repeatedly run across references to creation myths, including those of Native Americans and those from Greece, China, and Japan. You decide to focus on Native American creation myths.

After doing some preliminary research, you have prepared a brief bibliography of sources containing useful information (see page -- of this booklet for that bibliography). The notes you have taken on the information in these books appear on the note cards included with this task. The information is not yet organized, and you are not yet sure what the focus of your paper will be. You need to look through your notes and organize them, with a view to deciding on the central idea about myth that your paper will develop.

READ THROUGH ALL THE MATERIAL ON THE NOTE CARDS. These notes contain the kinds of information a researcher might write down as he or she reads books and articles about a particular topic. The names of authors and page references on the note cards refer to the works listed in the bibliography on page --. As you read, look for common themes or subjects. You may mark the note cards or make notes on them for your own use.

Question 1

Organize the material on the note cards.

- Look back over the note cards and sort them into at least three groups. Each group should have a common topic. These topics should be ones that might be useful for your paper. (You need not use all the cards.)
- Give each group a heading that clearly indicates a major topic you want to cover in your paper. In the space provided below and on the next page, write the headings you have decided on and the numbers of the cards that belong under those headings. **You should give at least three headings.**

Question 2

Read note cards 19 and 20 again. Note that the information contained on these two note cards is inconsistent. Explain what the inconsistency is between the two note cards. Indicate which of these two note cards contains information consistent with information on note cards 7 through 9.

Question 3

What is the important central idea about Native American creation myths that you will develop in your research paper? The idea should be based on your reading of the documents and the headings you provided in **Question 1.** In a complete sentence, write the idea you will develop below.

Question 4

Now write a draft of your research paper. In organizing your draft, use the headings you provided in **Question 1**. Make sure the information you are providing develops the central idea you stated in **Question 3.** The draft will be judged on how well you organize and express your ideas and how fully you use the information on the note cards.

- You must use information from the note cards to support your ideas and you should indicate your source, as you would in any research paper. Write the number of a note card in parentheses after a sentence that contains information from that note card.

- Your draft should contain more than one paragraph. It should have a clear beginning, in which you introduce your subject; it should offer evidence to support your ideas; and it should have a definite conclusion.

- Remember to write all your responses so that they can be read and understood easily by someone who does not know you or your handwriting.

SCORE DESCRIPTIONS

INQUIRY AND ANALYSIS:
Fully proficient (4) 8 The Core Score means that the questions were understood and the responses were correct and complete; students met all basic requirements.

Exceeds Requirements (5) 9-10 Students met all the basic requirements AND provided some expansion or extension--citing evidence, providing additional information, or in some other way going beyond what was required.

Superior Performance (6) 11-12 All basic requirements were met AND expanded upon and, IN ADDITION, these students presented ideas, interpretations, relationships, or examples that showed originality and/or insight.

Some Proficiency (3) 5-6 Scores at this level mean that, although students understood the questions, the basic requirements were not met. Responses were vague, incomplete, and/or inappropriate.

Limited Proficiency (2) 3-4 The basic requirements were not met, and responses were very brief, inappropriate, and/or incorrect. Students may have addressed the questions but their responses were vaguely expressed or inaccurate.

Not Proficient (1) 1-2 A response was attempted but students scoring at this level either did not understand the questions or their explanations were erroneous, illogical, or totally unrelated to the requirements.

(There is a similar scoring guide for Communication.)

NOTE: Task scores can be reported in two ways--the actual score given by each Reader/Scorer which is on a 1-6 scale, OR the sum of the scores given by the two Readers who scored each question, in which case the score range is 1-12. The number in ( ), for example (6), is the actual score; the two number range, 11-12, for example, is the sum. A score of 6 and a score of 11-12 mean the same thing.

4. Title: Persuasive or descriptive essay (English)

4.1. Item text: See topics below from SAT-II Writing and AP English Language:

SAT-II Writing:
Consider carefully the following quotation and the assignment below it. Then plan and write your essay as directed.

"Any advance involves some loss."

Assignment: Choose a specific example from personal experience, current events, or from your reading in history, literature, or other subjects and use this example as the basis for an essay in which you agree or disagree with the statement above. Be sure to be specific. (20 minutes)

AP English Language:
Our perceptions of people often differ according to our attitudes and circumstances. Describe in a vivid and concrete way one person seen at two different times (or in two different situations) so that the reader understands the difference in your attitude. (40 minutes)

(Note: AP English Language Essay is 40 minutes; SAT-II Writing essay is 20 minutes. Other kinds of analytic and persuasive writing, often based on brief reading passages, are included in AP English Language. SAT-II Writing sometimes uses a descriptive or analytic rather than a persuasive essay.)

4.2. This item links to the following Goal-Expectation-Indicators:

SAT-II (persuasive writing):
2.1.1, 2.1.3, 2.1.4, 3.1.5-3.1.7, 3.2.1

AP English Language (descriptive writing):
2.1.1, 2.1.2, 2.1.3, 3.1.5-3.1.7, 3.2.1 (could incorporate more of goals 1 and 4 if response to a literary text was part of the assignment)

4.3. This item is appropriate to the following Skills for Success:

SAT-II (persuasive writing): Thinking 1.2.3, 2.4.1 and 2.4.5, 2.6.4; Communication 2.5.4-2.5.7

AP English Language (descriptive writing): Thinking 1.2.3, 2.1.2 and 2.1.3, 3.2.2; Communication 2.5.4-2.5.7

4.4. Item's source: SAT-II Writing (persuasive essay), AP English Language (descriptive essay)

Scoring guide for SAT-II Writing below (holistic guide)
For AP English Language, scoring guide is more analytic than holistic.
Also included below is NAEP Writing scoring guide to persuasive writing; NAEP has similar guides for narrative and informative writing.

SAT-II Writing

Score of 6: A paper in this category demonstrates *clear and consistent competence* though it may have occasional errors. Such a paper:
-effectively and insightfully addresses the writing task
-is well organized and fully developed, using clearly appropriate examples to support ideas
-displays consistent facility in the use of language, demonstrating variety in sentence structure and range of vocabulary

Score of 5: A paper in this category demonstrates *reasonably consistent competence* though it will have occasional errors or lapses in quality. Such a paper:
-effectively addresses the writing task
-is generally well organized and adequately developed, using appropriate examples to support ideas
-displays facility in the use of language, demonstrating some syntactic variety and range of vocabulary

Score of 4: A paper in this category demonstrates *adequate competence* with occasional errors and lapses in quality. Such a paper:
-addresses the writing task
-is organized and somewhat developed, using examples to support ideas
-displays adequate but inconsistent facility in the use of language, presenting some errors in grammar or diction
-presents minimal sentence variety

Score of 3: A paper in this category demonstrates *developing competence.* Such a paper may contain one or more of the following weaknesses:
-inadequate organization or development
-inappropriate or insufficient details to support ideas
-an accumulation of errors in grammar, diction, or sentence structure

Score of 2: A paper in this category demonstrates *some incompetence.* Such a paper is flawed by one or more of the following weaknesses:
-poor organization

-thin development
-little or inappropriate detail to support ideas
-frequent errors in grammar, diction, and sentence structure

Score of 1: A paper in this category demonstrates *incompetence.* Such a paper is seriously flawed by one or more of the following weaknesses:
-very poor organization
-very thin development
-usage and syntactical errors so severe that meaning is somewhat obscured

Essays that appear to be off topic or that pose unusual challenges in handwriting or other areas should be given to the Table Leader.

NAEP PERSUASIVE SCORING GUIDE

1 Opinion. Paper is a statement of opinion, but no reasons are given to support the opinion, or the reasons given are inconsistent or unrelated to the opinion.

2 Extended Opinion. Paper states opinion and gives reasons to support the opinion, but the reasons are not explained or the explanations given are incoherent.

3 Partially Developed Argument. Paper states opinion and gives reasons to support the opinion, plus attempts to develop the opinion with further explanation. However, the explanations are given but not developed or elaborated. May contain a brief reference to the opposite point of view.

4 Developed Argument. Paper states opinion, gives reasons to support the opinion, plus explanations, with at least one explanation developed through the use of rhetorical devices (such as sequence of events, cause and effect, comparison/contrast, classification, problem/solution, point of view, drawing conclusions). May contain a brief summary of the opposite point of view.

5 Partially Developed Refutation. Paper states opinion, gives reasons to support opinion, explanations, plus attempts to discuss and/or refute the opposite point of view. Contains an adequate summary of the opposite point of view.

6 Developed Refutation. Paper states opinion, gives reasons to support opinion, explanations, plus a discussion and/or refutation of opposing point of view. Refutation is clear and explicit--summarizes opposite point of view and discusses why it is limited or incorrect.

5. Title: Open-ended literary essay (English)

5.1. Item text: See below (from AP English Literature; 40 minute question). Note: list would be modified to include more works associated with Core Learning Goals (and/or more accessible works)

5.2. This item links to the following Goal-Expectation-Indicators:
1.1.3, 1.2.1 and 1.2.2, 1.2.4 and 1.2.5, possibly 1.3.1, 1.3.2 through 1.3.6
2.1.1 through 2.1.4, 3.1.5 through 3.1.7, 3.2.1, 3.3.1, 4.1.1 and 4.1.2

5.3. This item is appropriate to the following Skills for Success:
Thinking: 1.2.3, 2.1.3, 2.3.1, and 2.6.4
Communication: 1.1.2, 2.5.4 through 2.5.7

5.4. Item's source: AP English Literature. Scoring rubric is included below. Note: suggest, for this purpose, that scoring scale be modified to a 6-point scale by collapsing 6-7/8-9 categories.

Item Text:
Choose a novel or play that depicts a conflict between a parent (or a parental figure) and a son or daughter. Write an essay in which you analyze the sources of the conflict and explain how the conflict contributes to the meaning of the work. Avoid plot summary. You may base your essay on *one* of the following works or choose another of comparable literary quality.

Aeschylus, *The Oresteia*
Austen, *Persuasion*
Baldwin, *Go Tell It on the Mountain*
Brontë, *Wuthering Heights*
Dickens, *Hard Times, Our Mutual Friend*
Dostoevski, *The Brothers Karamazov*
Eliot, *The Mill on the Floss*
Faulkner, *As I Lay Dying*
Fielding, *Tom Jones*
Hansberry, *A Raisin in the Sun*
Hellman, *The Little Foxes*
Hurston, *Their Eyes Were Watching God*
James, *Washington Square*
Lawrence, *Sons and Lovers*
Miller, *All My Sons*
Morrison, *Beloved*
O'Neill, *Long Day's Journey into Night*
Pinter, *The Homecoming*
Shakespeare, *King Lear, Henry IV, Romeo and Juliet*
Shaw, *Mrs. Warren's Profession*
Sophocles, *Antigone*
Spark, *The Prime of Miss Jean Brodie*
Turgenev, *Fathers and Sons*
Williams, *The Glass Menagerie*

Scoring Guidelines (for AP English Literature question on conflicts with parents; here modified from 9-point to 6-point scale)
General Directions: Scores assigned should reflect the quality of the essay as a whole. Reward the writers for what they do well. The score for a particularly well-written essay may be raised by one point from the score otherwise appropriate. In no case may a poorly written essay be scored higher than 3.

6-7 Well-written essay on a novel or a play that depicts a conflict between a parent (or parental figure) and a son or daughter; discusses in accurate and specific detail the sources of the conflict; and discusses the meaning of the work and how the parent-child conflict contributes to it. May contain flaws, but demonstrates an ability to discuss a literary work with insight and understanding and to control the elements of effective writing.

5 Chooses an appropriate work but deals superficially with the sources of the conflict. May allude to the meaning of the work as a whole but does not connect it clearly to the parent-child conflict. Typically, comments are overly generalized or simplistic; writing may be immature or demonstrate inconsistent control over the elements of composition. Not as well-conceived, organized, or developed as higher-level papers, but writing is sufficient to convey the writer's ideas.

3-4 May discuss an appropriate work, but analysis of the sources of conflict may be ineffective or unconvincing. May choose an inappropriate relationship, or account of the relation of the conflict to the meaning of the work as a whole may be perfunctory or omitted altogether. Often exclusively plot summaries. May allude at times to parent-child conflicts, but fails to focus on the sources of the conflict. Fails to define the meaning of the work and consequently its relation to the parent-child conflict. May convey the writer's ideas, but reveals weak control over such elements as diction, organization, syntax, or grammar. Typically contains misinterpretations or significant distortions of the work discussed; may also contain little, if any, specific or persuasive evidence to support statements.

1-2 Compounds the weaknesses of essays in the 3-4 range. May choose a work that does not contain a parent-child conflict or fail to discuss the meaning of the work or how the parent-child conflict contributes to the meaning of the work. Frequently unacceptably brief or poorly written on several counts. Although may make some attempt to answer the question, views presented have little clarity or coherence.

6. Title: Selected Response (multiple-choice) literary interpretation (English)
(For Combination or Limited Combination; in Combination, could precede short-answer questions or an essay on the same work or works)

6.1 Item Text: (see below for stimuli from AP Literature, to go with questions below, and for examples from SAT-II Literature)
[NOTE: The Keats/Frost pairing, in particular the Keats poem, may be too difficult for this purpose. Prose passages from more accessible sources might be preferable, though issue of how to test pre-20th century literature will need to be addressed.]

Multiple-choice questions on "Bright Star" and "Choose Something Like a Star"
Introduction: Keats's poem, written in 1819, served in some ways as an inspiration for Frost's poem, written in 1949.

Multiple-choice:
1. The word "staid" (line 25) includes which of the following meanings?
standing and halted
* stuffy and stable

2. Both poems share which of the following poetic forms?
sonnet form
* regular meter and rhyme

blank verse

3. The "Language we can comprehend" is language that
is highly poetic
* refers to the physical world

Could be followed by short-answer questions (5-10 minutes each) such as:

1. What is the effect of Frost's repeated use of phrases that begin with "but"? (lines 7, 12, 17)

2. What is the meaning of Frost's allusion to Keats's Eremite, line 18?

Could be followed by an essay question such as: Compare the moods of the endings of the two poems. How do they sum up the themes of the poems? (30 minutes)

6.2. This item links to the following Goal-Expectation-Indicators:
1.1.3, 1.2.2 and 1.2.3, 1.3.3

6.3. This item is appropriate to the following Skills for Success:
Does not really address any of the Skills for Success.

6.4. Item's source: SAT-II Literature, AP English
(No scoring guide for Selected Response/multiple-choice.)
Stimuli from AP English Literature (this pairing originally used as the basis of an essay):

Bright Star

Bright star! would I were steadfast as thou art-
Not in lone splendor hung aloft the night,
And watching, with eternal lids apart,
Like nature's patient, sleepless Eremite*
The moving waters at their priest-like task
Of pure ablution round earth's human shores,
Or gazing on the new soft-fallen mask
Of snow upon the mountains and the moors-
No-yet still steadfast, still unchangeable,
Pillowed upon my fair love's ripening breast,
To feel for ever its soft fall and swell,
Awake for ever in a sweet unrest,
Still, still to hear her tender-taken breath,
And so live ever-or else swoon to death.

--John Keats

*hermit

Choose Something Like a Star

O Star (the fairest one in sight),
We grant your loftiness the right
To some obscurity of cloud-
It will not do to say of night,
since dark is what brings out your light.
Some mystery becomes the proud.
But to be wholly taciturn
In your reserve is not allowed.
Say something to us we can learn
By heart and when alone repeat.
Say something! And it says, "I burn."
But say with what degree of heat.
Talk Fahrenheit, talk Centigrade.
Use Language we can comprehend.
Tell us what elements you blend.
It gives us strangely little aid,
But does tell something in the end.
And steadfast as Keats' Eremite,
Not even stooping from its sphere,
It asks a little of us here.
It asks of us a certain height,
So when at times the mob is swayed
To carry praise or blame too far,
We may choose something like a star
To stay our minds on and be staid.
--Robert Frost

Passage and items from SAT-II Literature

Kitchenette Building

We are things of dry hours and the involuntary plan,
Grayed in, and gray. "Dream" makes a giddy sound, not strong
Like "rent," "feeding a wife," "satisfying a man."

But could a dream send up through onion fumes
Its white and violet, fight with fried potatoes
And yesterday's garbage ripening in the hall,
Flutter, or sing an aria down these rooms

Even if we were willing to let it in,
Had time to warm it, keep it very clean,
Anticipate a message, let it begin?

We wonder. But not well! not for a minute!
Since Number Five is out of the bathroom now,
We think of lukewarm water, hope to get in it.
"Kitchenette Building," *The World of Gwendolyn Brooks.* Copyright (c) 1945 by Gwendolyn Brooks Blakely. By permission of Harper & Row, Publishers, Inc.

27. The best way to paraphrase "dry hours" (line 1) is
(A) summer drought
(B) fruitless existence
(C) chronic fatigue
(D) sudden misfortune
(E) orderly lives

28. The kind of paradox in the phrase "involuntary plan" (line 1) most closely resembles that in which of the following?
(A) Careful disorder
(B) Spontaneous combustion
(C) Dangerous hobby
(D) Secret agreement
(E) Irrelevant information

29. As it is used in the poem, "giddy" (line 2) can be understood in all of the following senses EXCEPT
(A) dizzy (B) flighty (C) ephemeral (D) impractical (E) raucous

Possible additional items for SAT-II Literature (original set has 9 items)
The tone of the poem can best be described as
light and informal
angry and satirical
* resigned and ironic

The use of the word "things" emphasizes
the number of possessions the speakers own
* the depersonalization of the speakers

Which of the following occurs in line 11?
* a change of pace
an extended metaphor

7. Title: Revision in Context (English)

7.1. Item text: See below
Note: This item type could be used in conjunction with an essay (in the Combination or Prep Plus), as for example:

Students answer multiple-choice questions which revise a revision-in-context passage (editorial on a subject such as wearing school uniforms)

Then, students are asked to write a 30-minute essay question giving their opinions on the same topic as the revision-in-context passage, e.g., "Do you think your school should require school uniforms?"

7.2. This item links to the following Goal-Expectation-Indicators:
3.1.1, 3.1.3, 3.1.4, 3.1.6, 3.1.8 and 3.1.9, 3.2.2, 4.3.1 (limited)

7.3. This item is appropriate to the following Skills for Success:
Communication: 1.1.1, 2.5.4, 4.1.2 (all in a somewhat indirect and limited fashion)

7.4. Item's source: SAT-II Writing

No scoring rubric (Selected Response/multiple-choice).

Directions: The following passages are early drafts of student essays. Some parts of the passages need to be rewritten. Read the passages and answer the questions that follow. Some questions are about particular sentences or parts of sentences and ask you to make decisions about sentence structure, word choice, and usage. Other questions refer to the entire essay or parts of the essay and ask you to consider organization, development, and appropriateness of language. Choose the answer that most effectively expresses the meaning and follows the requirements of standard written English. After you have chosen your answer, fill in the corresponding oval on your answer sheet.

Questions 37-42 are based on the following passage.

(1) I used to be convinced that people didn't actually win radio contests; I thought that the excited winners I heard were only actors. (2) Sure, people could win T-shirts. (3) They couldn't win anything of real value.

(4) I've always loved sports. (5) My friends fall asleep to the music of U2. (6) I listen to "Sports Night with Dave Sims." (7) His show is hardly usual fare for a teen-aged girl. (8) One night I heard Dave Sims announce this sports trivia contest with cash prizes of two thousand dollars. (9) I jumped at the chance to combine my talk-show knowledge with everything my father had taught me about sports in my almost sixteen years. (10) I sent in my self-addressed stamped envelope. (11) I forgot about the whole matter. (12) Then the questionnaire appeared in my mailbox ten days later. (13) Its arrival gave me a rude surprise. (14) Instead of sitting down and whipping through it, I trudged to libraries and spent hours digging for answers to such obscure questions as "Which NHL goalie holds the record for most career shutouts?"

(15) Finally, after days of double-checking answers, I mailed off my answer sheet, certain I would hear no more about the matter. (16) Certain, until two weeks later, I ripped open the

envelope with the NBC peacock and read "Congratulations..." (17) I was a winner, a winner of more than a T-shirt.

37. Which of the following is the best way to revise the underlined portions of sentences 2 and 3 (reproduced below) so that the two sentences are combined into one?

Sure, people could win T-shirts. They couldn't win anything of real value.

(A) T-shirts, and they couldn't win
(B) T-shirts, but they couldn't win
(C) T-shirts, but not being able to win
(D) T-shirts, so they do win
(E) T-shirts, while there was no winning

38. Which of the following sentences, if added after sentence 3, would best link the first paragraph with the rest of the essay?
(A) I have held this opinion about contests for a long time.
(B) The prizes offered did not inspire me to enter the contests.
(C) However, I recently changed my opinion about these contests.
(D) Usually the questions on these contests are really easy to answer.
(E) Sometimes my friends try to convince me to enter such contests.

B. MATHEMATICS

Comments and Evaluation:

1. ***Portfolio Plus*** - Allows for extended realistic problems and requires student reasoning and communication about and in the language of mathematics.

2. ***Preparation Plus*** - Allows for realistic problems and requires student reasoning and communication about and in the language of mathematics. Can affirm cooperative learning objectives. Similar to MSPAP experience.

3. ***Combination*** - Can tap process as well as product and teachers can be trained to score reliably with a limited rubric. The members of the content team strongly support this option.

4. ***Limited Combination*** - Does not use the language of mathematics or assess communication. The questions appear artificial and not tied to the real world. Does not follow through from MSPAP.

Illustrative Questions:

MHSA - ILLUSTRATIVE ITEM EXAMPLES
for the
DRAFT DESIGN MATRIXES

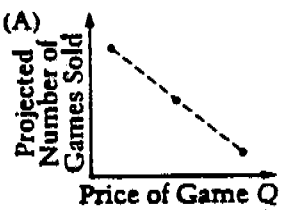
Item Text -

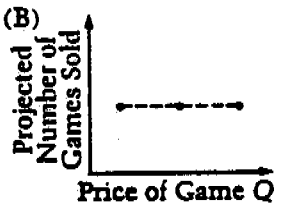
PROJECTED SALES FOR GAME Q

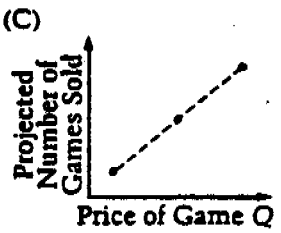
Price of Game Q	Projected Number of Games Sold
$50	50,000
$30	100,000
$10	150,000

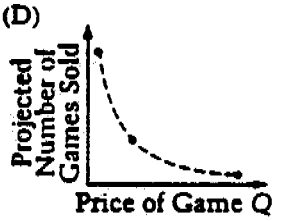
Game Q is a new game. Its manufacturers are trying to decide what the price of game Q should be. The table above shows the projected number of games sold for three different prices.

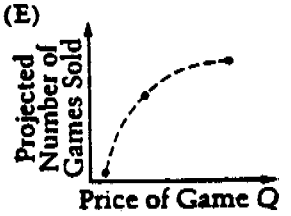
Which of the following graphs best represents the relationship between the price of game Q and the projected number of games sold, as indicated by the table?

(A)


(B)


(C)


(D)


(E)


This item links to the following Goal - Expectation - Indicators

Goal 1: Functions and Algebra
Expectation: 1.1
Indicator: 1.1.1

Goal 3: Data Analysis and Probability
Expectation: 3.2
Indicator: 3.2.2

This item is appropriate to the following Skills For Success

2.4.1, 2.4.4

Item's source

8 Real SATs

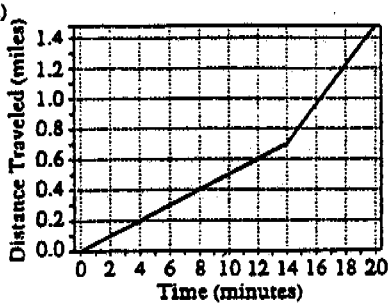
Scoring rubrics and instructions are attached when appropriate.

MHSA - ILLUSTRATIVE ITEM EXAMPLES
for the
DRAFT DESIGN MATRIXES

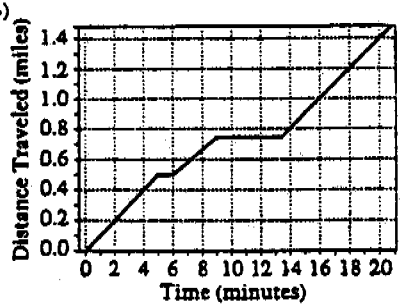
Item Text -

Andrea lives 1.5 miles from her school. On a particular day, she walked 0.5 mile, stopped for 1 minute to tie her shoelaces, then walked 0.25 mile, stopped for 3 minutes to talk to friends, then ran the remaining distance to school. Which of the following graphs could correctly represent her journey to school on this day ?

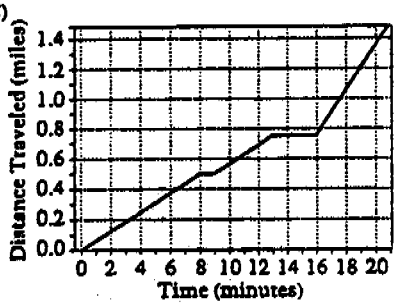
(A)



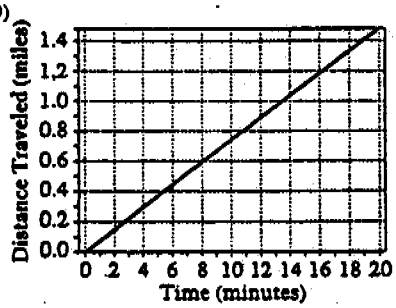
(B)



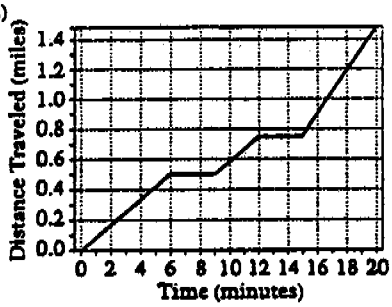
(C)



(D)



(E)



This item links to the following Goal - Expectation - Indicators

Goal 1: Functions and Algebra
Expectation: 1.2
Indicator: 1.2.4

This item is appropriate to the following Skills For Success

2.4.1, 2.4.4

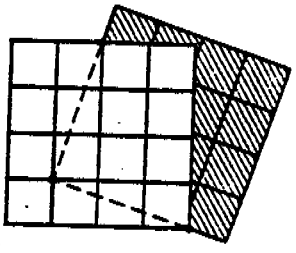
Item's source

8 Real SATs

Scoring rubrics and instructions are attached when appropriate.

MHSA - ILLUSTRATIVE ITEM EXAMPLES
for the
DRAFT DESIGN MATRIXES

Item Text -



Two squares, each composed of 16 small squares, overlap as shown in the figure above. If the area of each small square is 1, what is the area of the shaded region?

This item links to the following Goal - Expectation - Indicators

Goal 1: Functions and Algebra	Goal 2: Geometry, Measurement, and Reasoning
Expectation: 1.2	Expectation: 2.2, 2.3
Indicator: 1.2.5	Indicator: 2.2.1, 2.2.2, 2.3.2

This item is appropriate to the following Skills For Success

2.4.1, 2.4.4

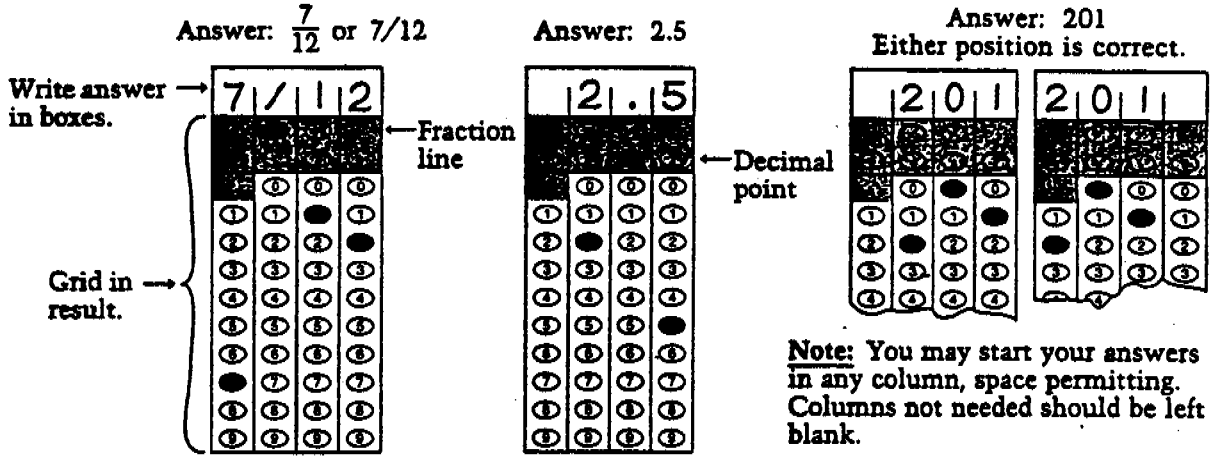
Item's source

8 Real SATs

Scoring rubrics and instructions are attached when appropriate.

Directions for Student-Produced Response Questions

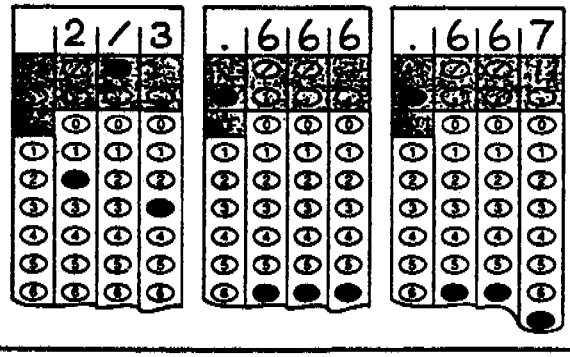
Each of the remaining 10 questions requires you to solve the problem and enter your answer by marking the ovals in the special grid, as shown in the examples below.



Note: You may start your answers in any column, space permitting. Columns not needed should be left blank.

- Mark no more than one oval in any column.
- Because the answer sheet will be machine-scored, you will receive credit only if the ovals are filled in correctly.
- Although not required, it is suggested that you write your answer in the boxes at the top of the columns to help you fill in the ovals accurately.
- Some problems may have more than one correct answer. In such cases, grid only one answer.
- No question has a negative answer.
- Mixed numbers such as $2\frac{1}{2}$ must be gridded as 2.5 or 5/2. (If 2 1/2 is gridded, it will be interpreted as $\frac{21}{2}$, not $2\frac{1}{2}$.)
- **Decimal Accuracy:** If you obtain a decimal answer, enter the most accurate value the grid will accommodate. For example, if you obtain an answer such as 0.6666 . . . , you should record the result as .666 or .667. Less accurate values such as .66 or .67 are not acceptable.

Acceptable ways to grid $\frac{2}{3}$ = .6666 . . .



MHSA - ILLUSTRATIVE ITEM EXAMPLES
for the
DRAFT DESIGN MATRIXES

Item Text -

Geraldo is saving for the compact disc (CD) player that is shown in the advertisement below. He must also pay a sales tax of 6 percent on his purchase.

CD PLAYER $119.99

Suppose Geraldo has saved *d* dollars toward his purchase and plans to earn the rest of the money by helping at the local day-care center. At the day-care center, he earns at most *w* dollars per week. Write an expression, in terms of *d* and *w*, that can be used to find the fewest number of weeks that will pass before he has enough money to buy the CD player.

This item links to the following Goal - Expectation - Indicators

Goal 1: Functions and Algebra
Expectation: 1.1
Indicator: 1.1.2

This item is appropriate to the following Skills For Success

2.4.1, 2.4.4

Item's source

NAEP

Scoring rubrics and instructions are attached when appropriate.

Solution:

$119.99(1.06) = 127.19$

$\frac{127.19 - d}{w}$ or any equivalent form

(Note: The fewest number of weeks is actually determined by the least integer greater than or equal to $\frac{127.19 - d}{w}$)

Scoring Guide:

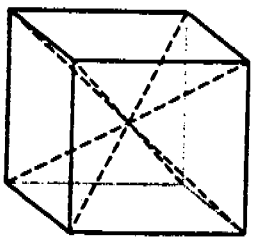
1 Incorrect response

2 $\frac{(.06)(119.99) - d}{w}$ or $\frac{127.19}{w} - d$ or any equivalent form.

3 Correct response

MHSA - ILLUSTRATIVE ITEM EXAMPLES for the DRAFT DESIGN MATRIXES

Item Text -



In the figure shown above, the center of a cube is connected by line segments to the eight vertices of the cube to form square pyramids. If the length of the edge of the cube is 10 centimeters, what is the volume of each pyramid? (The formula for the volume of a pyramid is $V = \frac{1}{3}Bh$.)

This item links to the following Goal - Expectation - Indicators

Goal 1: Functions and Algebra
Expectation: 1.2
Indicator: 1.2.5

Goal 2: Geometry, Measurement, and Reasoning
Expectation: 2.2, 2.3
Indicator: 2.2.2, 2.3.2

This item is appropriate to the following Skills For Success

2.4.1, 2.4.4

Item's source

NAEP

Scoring rubrics and instructions are attached when appropriate.

Solution:

$V_p = \frac{1}{3}(10)^2(5)$ $\qquad$ $V_p = \frac{V_c}{6} = \frac{10^3}{6}$

$= \frac{500}{3}$ $\qquad$ OR $\qquad$ $V = \frac{1,000}{6}$

$= 166\frac{2}{3}$ $\qquad$ $= 166\frac{2}{3}$

Scoring Guide:

1. Incorrect response

2. $V_p = \frac{1}{3}(10)^2(5)$ $\quad$ OR $\quad$ $V_p = \frac{10^3}{6}$ $\quad$ OR $\quad$ 166

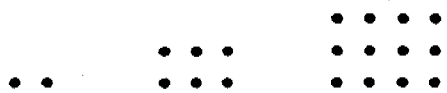
3. Correct response: $166\frac{2}{3}$ or equivalent with correct accompanying work
   OR
   $166\frac{2}{3}$ or equivalent with no work

MHSA - ILLUSTRATIVE ITEM EXAMPLES
for the
DRAFT DESIGN MATRIXES

Item Text -

A pattern of dots is shown below. At each step, more dots are added to the pattern. The number of dots added at each step is more than the number added in the previous step. The pattern continues infinitely.

(1st step) (2nd step) (3rd step)



Marcy has to determine the number of dots in the 20th step, but she does not want to draw all 20 pictures and then count the dots.

Explain or show how she could do this and give the answer that Marcy should get for the number.

This item links to the following Goal - Expectation - Indicators

Goal 1: Functions and Algebra
Expectation: 1.1
Indicator: 1.1.1

This item is appropriate to the following Skills For Success

2.4.4, 3.2.4, 3.2.5

Item's source

NAEP

Scoring rubrics and instructions are attached when appropriate.

This question requires you to show your work and explain your reasoning. You may use drawings, words, and numbers in your explanation. Your answer should be clear enough so that another person could read it and understand your thinking. It is important that you show all your work.

Solution:

Explanation should include one of the following ideas with no false statements.

a. For each successive step, the number of rows and the number of columns is increasing by 1, forming a pattern. For example, the first step forms 1 by 2 rows and columns, the next step 2 by 3, the third step 3 x 4, and so on. Continuing this pattern would mean that the 20th step has 20 x 21 or 420 dots.

b. Look at successive differences between consecutive steps. The differences 4, 6, 8, 10, . . . form a pattern. There are 19 differences forming the pattern 4, 6, 8, 10, . . . , 38, 40 and this sum is (9 x 44) + 22 or 418. However, 2 must be added for the 1st step, yielding a response of 420.

1 An attempt to generalize OR to draw all 20 pictures in the pattern (with a clear understanding of the pattern).

2 A partial (incomplete) correct explanation, i.e., does not tie together well.

3 Correct explanation of pattern but does not include or omits the correct number of dots (420).

4 Correct answer. (Must state 420; must tie step 20 back to beginning of pattern in some specific form of generalization.)

9 The work is completely incorrect, irrelevant, or off task. (Just 420 is a score of 9.)

0 = No response (blank)

MHSA - ILLUSTRATIVE ITEM EXAMPLES
for the
DRAFT DESIGN MATRIXES

Item Text -

The attached sample shows a preliminary activity and several individual student activities for an extended task.

This item links to the following Goal - Expectation - Indicators

Goal 1: Functions and Algebra	Goal 3: Data Analysis and Probability
Expectation: 1.2	Expectation: 3.2
Indicator: 1.2.1	Indicator: 3.2.1, 3.2.2

This item is appropriate to the following Skills For Success

2.2.1, 2.2.2, 2.2.6, 2.4.1, 2.4.4, 3.2.4, 3.2.5, 4.2.2

Item's source

Scoring rubrics and instructions are attached when appropriate.

Preliminary Activity

	Person 1	Person 2	Person 3	Person 4
Wrist Measurement				
Ankle Measurement				

Each student should collect a set of data from four different people (family members, friends, teachers, self, etc.) by taking two different measurements with a flexible tape measure (for accuracy) and fill in the chart above. No individual should provide his or her measurements to more than one class member. For consistency, it should be decided beforehand if the measurements are to be in centimeters or inches. All class members should use the same units of measure.

All data collected by individual class members should be pooled and shared. This will ensure that everyone has the same set of data and will provide some consistency for scoring student work. For example, if there are 20 class members, there should be 80 pairs of data.

Individual Student Activities

1. Using the data as ordered pairs (wrist measurement, ankle measurement), make a scatterplot of all the data.

2. Find a line of best fit and its equation.

3. Explain why the *y*-intercept of the line of best fit has no significance in this situation.

4. Suppose another set of data consisting of waist measurements for the same set of individuals who were measured previously was collected. Suppose also that a scatterplot was made from the ordered pairs (wrist measurement, waist measurement) and a line of best fit were found for this scatterplot. How would the slope of this second line compare to the slope of the line in 2 above? Explain why you think this is true.

5. Suppose you were given a person's wrist measurement that was not among the original data. Using the two lines of best fit (from 2 and 4 above) as predictors, do you think you could better predict the person's ankle measurement or the person's waist measurement? Give a reason for your answer.

Portfolio Number _____ First Reader _____

Check Most Appropriate Score for PUTTING MATHEMATICS TO WORK

NOT ENOUGH EVIDENCE to score this dimension ☐

BEGINNING ☐	DEVELOPING ☐	ACCOMPLISHED ☐	EXEMPLARY ☐
• Thinking: little demonstration of mathematical thinking; work includes no reflection about process or product, or contains statements in terms of general features, such as likes and dislikes	• some work may include fragments of conjectures, arguments or generalizations; conclusions may be inconsistent with data; work may include very general reflective statements about student having learned or having experienced difficulties	• work demonstrates sound mathematical thinking involving conjectures, arguments, predictions, generalizations, or logical conclusions; process or product may be described in terms of mathematical ideas	• work demonstrates a high level of mathematical thinking; errors are explored until problems are resolved; thoughtful questions are expressed; may include unusual insights into the resolution of difficulties or details about ways work was or could be improved in terms of qualities of good work
• Investigation and problem solving: work may include attempts to complete unstructured problems or investigations; work demonstrates little or no understanding of problems or questions posed	• parts of structured problems or investigations are completed adequately; work demonstrates partial understanding of the problems or questions posed; goals and assumptions are left implicit; plans are sketchy and may not be carried out	• a variety of strategies are used to solve problems and/or investigate questions; work demonstrates understanding of the problems or questions posed; goals and assumptions are made explicit; plans are made and carried out, although may contain minor errors	• in addition, problems are solved in more than one way; new problems are formulated or novel strategies are devised; work demonstrates student has ways to get unstuck; situations may be investigated beyond initial purposes
• Using ideas: work demonstrates little or no purposeful use of mathematical ideas	• work demonstrates some successful use of specified mathematics to do assignments; may attempt to use ideas purposefully	• a variety of mathematical ideas are successfully put to work for a purpose	• in addition, mathematical ideas are put to work for a variety of purposes; mathematical ideas are brought to bear to make sense of situations; connections are made among mathematical ideas in various situations
• Using tools and techniques: little or no purposeful use of mathematical tools, techniques and resources	• tools, techniques and resources may be used to complete simple tasks, but used ineffectively or inappropriately in purposeful activities	• work includes appropriate, purposeful use of a variety of mathematical tools, techniques and resources	• in addition, appropriate mathematical tools and techniques are intentionally selected for particular purposes; new tools and techniques are invented, or new resources brought to bear when needed

Very Clear, Possible Benchmark ☐

Day 1 Day 2 early am late am early pm late pm Table Leader Initials _____

Portfolio Number _____ First Reader _____

Check Most Appropriate Score for MATHEMATICAL CONTENT

NOT ENOUGH EVIDENCE to score this dimension ☐

BEGINNING ☐	DEVELOPING ☐	ACCOMPLISHED ☐	EXEMPLARY ☐
• Accuracy: the work is incomplete or contains serious errors	• some calculations and procedures are completed reliably; some efforts are sketchy, incomplete, or inaccurate	• almost all calculations and procedures are performed accurately and completely	• in addition, the work shows multiple correct responses and alternative formulations, especially of the student's own invention; may contain minor computational errors
• Depth and quality: work demonstrates little or no understanding of concepts or procedures; some work may be below grade level for the student	• some of the work shows limited understanding of concepts or procedures; understanding is narrow or partially correct	• the work shows clear understanding of a variety of concepts and procedures, and one or more unifying ideas drawing from multiple strands	• in addition, the work may show integration among ideas, and connections of topics to the world or to one another, unusual insights into the nature of the mathematics, or revision of ideas and understanding over time; some work may be beyond the grade level of the student

Very Clear, Possible Benchmark ☐

Day 1 Day 2 early am late am early pm late pm Table Leader Initials _____

Portfolio Number _____ First Reader _____

Check Most Appropriate Score for COMMUNICATING MATHEMATICS

NOT ENOUGH EVIDENCE to score this dimension ☐

BEGINNING ☐	DEVELOPING ☐	ACCOMPLISHED ☐	EXEMPLARY ☐
• Explanation: work typically includes answers, with little or no attempt to describe, explain or justify results, or includes explanations that are incomplete, unclear, and/or unreasonable	• work includes attempts to describe, explain or justify mathematical questions, ideas, processes, or results in writing, through drawing, or orally; communication may be somewhat unclear, or expression may be articulate but include little mathematical substance	• work includes clear descriptions, explanations, and justifications of mathematical questions, ideas, processes or results, in writing, through drawing, or orally; explanations are reasonable, whether or not they are completely correct, or include conventional notation or technical vocabulary	• in addition, work includes multiple modes of expression for specific purposes; communication is exceptionally clear, complete, and well organized; communicates creatively to identified audiences
• Representation: work typically includes no or very few symbolic expressions, physical representations, diagrams, or graphs; those included are incomplete, unclear, and/or unreasonable	• work includes symbolic expressions, physical representations, diagrams, or graphs; may be partially correct, hard to interpret, or disorganized	• work includes a variety of representations, including words, symbols, physical materials, graphs and diagrams; representations are used appropriately and are clear and understandable, whether or not they include conventional notation or technical vocabulary	• in addition, words, symbols, physical materials, graphs and diagrams are used purposefully, as tools to solve problems or investigate relationships

Very Clear, Possible Benchmark ☐

16

Day 1 Day 2 early am late am early pm late pm Table Leader Initials _____

C. SCIENCE

Comments and Evaluation:

The Science Content Team recommends the Preparation Plus design option for the High School Assessment in science. This group feels that this option typifies what is needed in an assessment to support the instructional reform promoted in the Core Learning Goals. Many components of the Prep Plus option are desirable in enhancing instruction. Questions can be designed which are based on actual student-generated laboratory data, thus linking instruction with assessment. The analysis of such data, extracting true meaning from investigations, would allow students to experience the full realm of scientific reasoning. The laboratory experience would provide an opportunity for students to manipulate traditional science equipment and new technology.

This option also provides the ability to incorporate questions of various types that are connected to a common theme. Students can be assessed on their knowledge of concepts as well as processes as they "think through" each scenario presented, whether it be real world application, hypothetical situation or laboratory data.

The Science Content Team recommends that if the Prep Plus option proves impractical, the Combination option be used because it incorporates many of the same benefits although, without the laboratory, the impact on instructional reform will be diminished.

The Portfolio Plus option is seen as having too many unresolved issues and the Limited Combination option does not promote instructional reform nor reflect the systemic change advocated for science. The design of the test would be too narrow, making it more difficult to incorporate the higher order thinking skills and the processes of science called for in the CLGs.

Illustrative Questions:

1. Title: Portfolio and Extended Constructed Response (Science)

1.1. Item Text

This item format is appropriate for the Extended Constructed Response item in the Portfolio Plus design.

1.2. Item Text and Comments:

As of this date, ETS has no examples of portfolios used in their high-stakes assessment programs in science. A special study, the National Assessment of Educational Progress (NAEP) Science Portfolio Study was conducted with 4th and 8th grade science students in 1995-1996. The goal of this portfolio study was to "...assess what 4th and 8th grade students know and can do in science without assessing their writing ability, or their science knowledge."

Two other portfolio programs were briefly examined to learn how they handle the issue of high-stakes assessment usage:

A) Golden State Science Portfolio - California Department of Education. The science portfolio was a collection of student work put together over a year in the areas of biology, chemistry, physics, or second-year coordinated science. The portfolio was to be combined with other test items (selected response, brief and extended constructed response) as part of a total score in cases where student's total score is improved by inclusion of the portfolio score. The portfolio scores are not used by themselves to evaluate student performance.

B) Connecticut Academic Performance Test (CAPT) - The introduction to the 1994 Science Administration Released Items states: "...the science component of the CAPT is not a test of high school science content, but rather a cumulative assessment of science proficiency covering content and skills that students should have acquired in grades K-10."

1.3. This item links to the following Goal - Expectation - Indicators:
The production of the portfolio will meet a great majority of Core Learning Goal # 1. The activities can be chosen to meet any of the indicators in biology, chemistry, physics, or earth/space science.

1.3. This item links to the following Skills for Success:
The production of the portfolio will meet a great majority of all of the Skills for Success.

1.4. Item's source:
1994 Golden State Science Portfolio - California Department of Education
1994 Connecticut Academic Performance Test - Released Items and Scored Student Responses

2. Title: Extended Constructed Response (Science)
2.1. Item Text and Comments
This item format is appropriate for the Extended Constructed Response item in the Portfolio Plus, Preparation Plus, or Combination design. While this particular example is from an AP Biology Examination, taken mostly by juniors and seniors, this format can be adapted for use at any grade level.

This item consists of an introductory statement providing the student with some background information and introducing the main theme of the item. Following the data, there are a series of bulleted questions and instructions. It should be noted that while students are required to answer all parts of the item, a high score can be obtained by answering fewer parts. Also, items of this type can be developed to allow for internal choice. That is, students are given the chance to choose which of several topics they wish to discuss as part of one question. See the example, Free-Response Question 2, attached.

As indicated by the enclosed scoring guide, this item type requires human scoring. The scorers need to be subject specialists because they are sometimes called on to make decisions on the appropriateness of a response. Scoring guides would be too long and unwieldy if they included every possible correct example or explanation. Other scoring issues include the time it takes to train retraining scorers. Very high administrative costs need to be recognized.

Item types like the Extended Constructed Response allow for broad coverage of the Core Learning Goals. Items can be designed to integrate several Indicators by tying them together with a common theme or concept.

See two attached free-responses items.

2.2. This item links to the following Goal - Expectation - Indicators

Goal 1 - Skills and Processes

Expectation	Indicator(s)
2	all
4	4.1
5	5.1, 5.2, 5.5
6	6.1, 6.4
7	7.1

Goal 3 - Concepts of Biology

Expectation	Indicator(s)
1	all
2	all
5	5.1, 5.5, 5.6

2.3. This item links to the following Skills for Success

Thinking Skills

Expectation	Indicator(s)
1	all
2	all
3	3.2 - 3.5
4	all

Communication Skills

Expectation	Indicator(s)
1	1.1 - 1.2, 1.4,
2	2.3, 2.5

Interpersonal, Learning, and Technology skills are not relevant to this item.

2.4. Item's source

Both items were published in the 1996 Advanced Placement Biology Examination Free-Response Section.

Free-Response Question 2

Structure and function are related in the various organ systems of animals. Select <u>two</u> of the following four organ systems in vertebrates:

respiratory
digestive
excretory
nervous

For each of the two systems you choose, discuss the structure and function of <u>two</u> adaptations that aid in the transport or exchange of molecules (or ions). Be sure to relate structure to function in each example.

Question 2 Standards

Overall Commentary for Question 2

Question 2 tested students' understanding of the relationship between structural and functional adaptations related to the transport of molecules and ions in representative vertebrate systems. Students were to choose two adaptations in each of two organ systems to discuss; and for the most part had no trouble choosing the two systems. They did have trouble choosing two adaptations within each system. There was some concern in regard to whether students really understood what an adaptation is.

Respiratory System

Structural adaptation	Functional significance (HOW it aids transport)
Alveolus	Provides increased surface area Moist, related to how it aids exchange/transport One cell-thick membrane (basement membrane) Site of oxygen/carbon dioxide exchange (elaboration) Close proximity of capillary bed to surface promotes absorption
Lung	Same as for alveolus — well described
Gill	Same as for alveolus
Muscles (diaphragm or intercostal muscles)	Muscular contraction promotes bulk flow of air
Cilia	Moves mucus/dirt removed from tract
Counter-current exchange	Increases gas transport through gill filaments
Jaw, throat pouch, nostril valves (amphibians)	Provides positive pressure gradient into lungs
Cartilage rings in trachea and bronchi	Keeps airways open
Respiratory pigments (etc.)	Establishes concentration gradients
Moist skin	Facilitates transport of gases (amphibians)
Moist pleural membrane	Allows lungs to expand upon contraction of diaphragm and intercostal muscles Maintains lung inflation
Bone structure in birds	Allows continuous air flow

Free-Response Question 3

Numerous environmental variables influence plant growth. Three students each planted a seedling of the same genetic variety in the same type of container with equal amounts of soil from the same source. Their goal was to maximize their seedling's growth by manipulating environmental conditions. Their data are shown below.

Plant Seedling Mass (grams)

	Day 1	Day 30
Student A	*4*	*24*
Student B	*5*	*35*
Student C	*4*	*64*

(A) Identify three different environmental variables that could account for differences in the mass of the seedlings at day 30. Then choose one of these variables and design an experiment to test the hypothesis that your variable affects growth of these seedlings.

(B) Discuss the results you would expect if your hypothesis is correct. Then provide a physiological explanation for the effect of your variable on plant growth.

Question 3 Standards

Overall Commentary for Question 3

Question 3 is composed of two discrete parts in which Part A has two components. A perfect score of ten could not be obtained unless at least one point was earned for each part or component (i.e., one point for naming three variables, plus at least one point for developing an experiment linked to some variable mentioned, plus at least one point for results expected, and at least one point for a discussion of physiology linked with the same variable.)

A. Environmental Variables and Experiment

Variables* . . . need three for 1 point

Light (Intensity-duration-wavelength)
Water
Temperature
CO_2
Humidity
Wind
Soil Type (Adj) — (Sand-vermiculite)
Soil Chemistry (Adj) — (pH-fertilizers)
Elevation
Competition
Hormones (added)
Predation

(not an exhaustive list)

Experiment . . . 6 Maximum

(1) Control—Constants

(1) Manipulation of variable
(how manipulated)

(1) Measurement of growth
(measured as [mass-length-dry-wet] initial vs final-% change-duration)

(1) Verification (sample size-repetition)

(1) Elaboration (of any one of above)

(1) Overall exceptional experimental design

(1) Hypothesis (includes measurable predictions and clearly states experimental conditions)

B. Results and Physiological Explanation

Physiology . . . 4 Maximum

(1) Results — linked to experiment

(1) Physiological function affected
(linked to variable)

(1) Concept of physiology
(Carbon of CO_2 incorporated in carbon chains)

(1) Elaboration (of Results or Physiology)

* A score of 10 cannot be earned without the point for Variables.

3. Title: Brief Constructed Response (Science)

This item format is appropriate for the Brief Constructed Response item in the Portfolio Plus, Preparation Plus, or Combination design. This particular example is taken from published NAEP (National Assessment of Educational Progress) examination taken by 7th and 11th graders.

3.1. Item Text and Comments

In this item, students are asked to sort a collection of small-animal vertebrae into three groups and explain how the bones in the groupings are alike. To complete this task, students need to make careful observations about similarities and differences among the bones and to choose their categories according to sets of common characteristics.

Materials required for this item include a package of 17 plastic replicas of small-animal vertebrae. If desired, it can include a set of measuring calipers and a metric rule. The measuring calipers and metric rule would enable students to make accurate measurements of the bones and their characteristics. These measurements would help them to make fine distinctions between subcategories. The activity can be done in a small group (no more than 4 students) or individually. If this is a small-group activity, it is recommended that the observations and measurements be done in the group setting, but that the discussion and explanations be written on an individual basis. The cost of producing the objects to be classified is relatively low. Other everyday items such as seeds or nails can be substituted for the plastic replicas. The task can be completed in 20 to 30 minutes.

The scorers need to be subject specialists because they are sometimes called on to make decisions on the appropriateness of a response. Scoring guides would be too long and unwieldy if they included every possible correct example or explanation. Other scoring issues include the time it takes to train retraining scorers. Very high administrative costs need to be recognized.

Item types like the Brief Constructed Response allow for broad coverage of the Core Learning Goals. Item sets can be designed to integrate several Indicators by tying them together with a common theme, concept, or activity.

See attached Brief Constructed Response items embedded in the activity.

3.2. This item links to the following Goal - Expectation - Indicators

Goal 1 - Skills and Processes

Expectation	Indicator(s)
1	1
2	5, 6
3	1-3
4	1, 3, 5
5	1, 2, 3
6	4
7	4

Goal 2 - Concepts of Earth/Space Science

Expectation	Indicator(s)
6	2

Goal 3 - Concepts of Biology

Expectation	Indicator(s)
2	1
4	1, 2
5	2

3.3. This item links to the following Skills for Success

Thinking Skills

Expectation	Indicator(s)
1	1 - 3
2	1 - 4, 6
3	1 - 6
4	1 - 5

Communication Skills

Expectation	Indicator(s)
1	2
2	2 - 5
3	1, 2
4	1, 2

Learning Skills

Expectation	Indicator(s)
5	1 - 3

Interpersonal Skills

Expectation	Indicator(s)
1	1 - 3
2	1 - 6, 8

Technology Skills

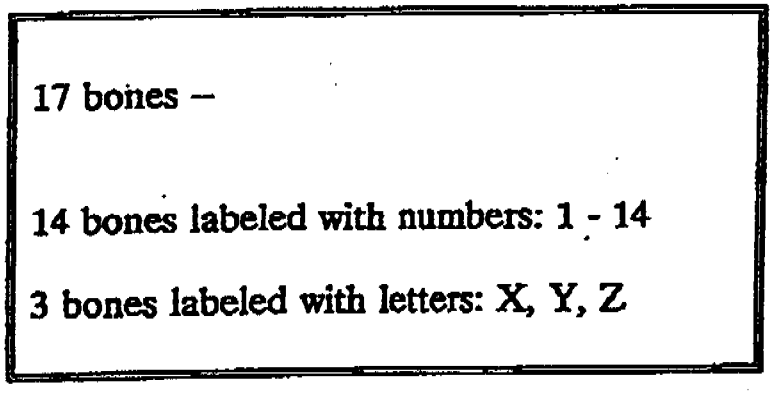
Expectation	Indicator(s)
1	2
2	1 - 3, 5
5	2

3.4. Item's source

"Learning By Doing" - National Assessment of Educational Progress

Sorting Bones Based on Similarities and Differences in Structure

For this task, you have been given a kit that contains materials that you will use to perform an investigation during the next 30 minutes. Please open your kit now and use the following diagram to check that all of the materials in the diagram are included in your kit. If any materials are missing, raise your hand and the administrator will provide you with the materials that you need.



The bones that you have just taken out of the bag are *vertebrae* from a particular species of animal – when all of the vertebrae of the animal are linked together, they form the backbone, or *vertebral column*, of the animal. The bones that you will be working with are just a sample of the entire set of vertebrae that make up the backbone of this animal. All of these bones are actually lifelike plastic models of bones.

In this activity, you will be asked to examine the bones and sort them into three groups based on similarities and differences in the bones' structure. Follow the directions step-by-step and write your answers to the questions in the space provided in your booklet.

1. Examine the bones labeled 1 through 14. (Set aside the bones labeled X, Y, and Z for now.) Look carefully at the structural features of the bones: for example,

the overall size and shape of the bones,

the size, shape, and location of the holes in the bones,

the size, shape, and location of the long spines and short knobs on the bones.

2. Sort the 14 bones into three different groups -- Group 1, Group 2, and Group 3 -- based on similarities and differences in structure. There should be at least 3 bones in each group. All of the bones in a group should be similar to each other in several ways, and they should also be different from the bones in the other two groups.

3. What are the numbers on the bones that you placed in Group 1? _____

Describe three ways that all of the bones in Group 1 are similar to each other.

_____

_____

_____

_____

What are the numbers on the bones that you placed in Group 2? _____

Describe three ways that all of the bones in Group 2 are similar to each other.

_____

_____

_____

_____

4. Based on the similarities and differences that you observed in the 14 bones, describe how you sorted the bones so that someone else could sort the 14 bones into the same three groups that you did.

5. Now look at the vertebrae labeled X, Y, and Z. These bones are from different animals than are bones 1-14. In which of your three groups (Group 1, Group 2, or Group 3) do you think the bone labeled X should be classified? _____ Explain why you classified bone X in this group.

In which of your three groups do you think the bone labeled Y should be classified? _____ Explain why you classified bone Y in this group.

In which of your three groups do you think the bone labeled Z should be classified? _____ Explain why you classified bone Z in this group.

D. SOCIAL STUDIES

Comments and Evaluation:

Each of the four design options was reviewed by the content team and College Board/ETS staff for the implications specific to Social Studies. The resulting comments follow:

*1. **Portfolio Plus*** - There is no commonly used or agreed upon model of "portfolio" assessment for Social Studies. Members of the content team discussed what such an assessment might look like, and while there was general agreement that such an assessment would support and advance current efforts at curriculum reform, the substantial investment of resources required to make such an assessment possible led most to agree that it was an impractical option to pursue.

*2. **Preparation Plus*** - Members of the content team discussed this model at length and saw it as perhaps the model that best captures the ideal of instruction and assessment in Social Studies to which most educators aspire. As envisioned, a common research task undertaken by all students prior to the assessment would provide both a common foundation on which to build the assessment as well as a means of integrating the assessment with classroom instruction. In fact, Frederick County already has an assessment of this type in place in which performance is evaluated during an overall assessment that takes place over several days. The requirements of a high-stakes statewide assessment, however, make clear the difficulties that emerge from an assessment that depends so greatly on a *common* foundation of instruction and preparation. The content team indicated that this option would continue the instructional reforms promoted in the earlier grades by MSPAP.

*3. **Combination*** - While reluctant to give up on the possibility of a common preparatory task (Prep Plus), the content team found that the "Combination" model realistically addressed two of their main concerns: first, that the assessment be closely linked to instruction and curriculum reform; and second, that the demands on teachers and students be fair and help to achieve the Core Learning Goals. As envisioned for Social Studies, this option would essentially follow an Advanced Placement model by providing an assessment that incorporated a writing portion. To serve the HSA population, Extended Constructed Response questions as well as Brief Constructed Response questions would necessarily be written in order to allow students of varying abilities to attempt responses.

*4. **Limited Combination*** - This design option was not considered viable given the goals of the Maryland HSA regarding curriculum reform in support of the Core Learning Goals. The members of the content team recognized that Selected Response questions can, in fact, be written in such a way that they go beyond simple factual recall, and acknowledged that such items would be an important part of each of the HSA assessments in Social Studies. However, the generally negative view of the current Maryland Citizenship Competency Test, and its negative impact in the state on the teaching of Government, led the team to indicate clearly that an assessment that relied solely on Selected Response questions would not achieve the goals of the HSA and would be opposed by teachers, students, and parents.

Illustrative Questions:

MARYLAND HIGH SCHOOL ASSESSMENT ITEM EXEMPLARS

Subject: United States History

Item text
(attach scoring guide if necessary)

Document-Based Free-Response Question:

Directions: The following question requires you to construct a coherent essay that integrates your interpretation of Documents A-H and your knowledge of the period referred to in the question. High scores will be earned only by essays that both cite key pieces of evidence from the documents and draw on outside knowledge of the period. Some of the documents have been edited, and wording and punctuation have been modernized.

To what extent was late nineteenth-century and early twentieth-century United States expansionism a continuation of past United States expansionism and to what extent was it a departure?

Use the documents and your knowledge of United States history to 1914 to construct your answer.

–See Documents Attached–

This item links to the following Goal - Expectation - Indicators: 2.2 A1 B2,3,4

This item is appropriate to the following Skills for Success: 2.2 A1,2,3,4

Source: AP US History

Summary of Documents

Document A: Thomas Nast cartoon, 1885

Expresses anti-imperialist, anti-expansionist sentiment. Could be some ambiguity in interpretation of document: student may draw the conclusion that since Russians, British, and Germans are engaged in global expansionism, the United States better join in the competition.

Document B: Josiah Strong, Our Country, 1885

Expression of white man's burden or manifest destiny with a new twist; Anglo-Saxon superiority; Sense of inevitability of United States overseas expansion; Social Darwinism suggested; Ideological basis for expansion/imperialism; Some continuity with the 1840s argument but more specific – departure

Document C: Platform of Anti-Imperialist League (1889)

Self-determination violated in Philippines; No intention of making citizens of the inhabitants of possessions; Philippine War – "ruthless slaughter"

Document D: Senator Albert Beveridge (1900)

White Man's Burden (compatible with Documents B and C); Economic Determinism – markets; Manifest Destiny revisited – continuity with 1840s;

Document E: Cartoon – "American Diplomacy" (1900)

Open door Diplomacy; Suggests Secretary of State Hay's Notes; Linkage to Hawaii, Guam, Philippines; Not seeking territorial possessions; Informal Imperialism; Great Power Politics

Document A

Source: Thomas Nast. "The World's Plunderers." *Harper's Weekly*, 1885.

Courtesy of the Newberry Library, Chicago.

Document B

Source: Josiah Strong. *Our Country: Its Possible Future and Its Present Crisis.* New York: American Home Missionary Society, 1885.

It seems to me that God, with infinite wisdom and skill, is training the Anglo-Saxon race for an hour sure to come in the world's future. . . . The unoccupied arable lands of the earth are limited, and will soon be taken. . . . Then will the world enter upon a new stage of its history — *the final competition of races, for which the Anglo-Saxon is being schooled.* . . . Then this race of unequalled energy, with all the majesty of numbers and the might of wealth behind it — the representative, let us hope, of the largest liberty, the purest Christianity, the highest civilization . . . will spread itself over the earth. If I read not amiss, this powerful race will move down upon Mexico, down upon Central and South America, out upon the islands of the sea, over upon Africa and beyond. And can any one doubt that the result of this competition of races will be the "survival of the fittest"?

Document C

Source: Platform of the American Anti-Imperialist League, 1899.

. . . Much as we abhor the war of "criminal aggression" in the Philippines, greatly as we regret that the blood of the Filipinos is on American hands, we more deeply resent the betrayal of American institutions at home . . . .

Whether the ruthless slaughter of the Filipinos shall end next month or next year is but an incident in a contest that must go on until the Declaration of Independence and the Constitution of the United States are rescued from the hands of their betrayers. Those who dispute about standards of value while the foundation of the Republic is undermined will be listened to as little as those who would wrangle about the small economies of the household while the house is on fire. The training of a great people for a century, the aspiration for liberty of a vast immigration are forces that will hurl aside those who in the delirium of conquest seek to destroy the character of our institutions.

Document D

Source: Senator Albert J. Beveridge. Speech to 56th Congress, *Congressional Record.* 1900.

The Philippines are ours forever. . . . And just beyond the Philippines are China's illimitable markets. We will not retreat from either. We will not repudiate our duty in the archipelago. We will not abandon our opportunity in the Orient. We will not renounce our part in the mission of our race, trustee, under God, of the civilization of the world. And we will move forward to our work . . . with gratitude . . . and thanksgiving to Almighty God that He has marked us as His chosen people, henceforth to lead in the regeneration of the world. . . .

Our largest trade henceforth must be with Asia. The Pacific is our ocean. . . . And the Pacific is the ocean of the commerce of the future. . . . The power that rules the Pacific, therefore, is the power that rules the world. And, with the Philippines, that power is and will forever be the American Republic.

Document E

Source: "American Diplomacy", 1900.

Outside Information Likely to Be Used

Following are examples of the outside information students could incorporate in their responses to the Document-Based Question. The list is not definitive; there is much outside information that is not listed here. Quantity, however, is not necessarily and indication of a good response; students who included only a small percentage of the items listed below may do very well.

Chronological Listing of Territorial Acquisitions by the United States

1783 – Territory from Appalachia to Mississippi
1795 – Yazoo Strip/Pinckney Treaty
1803 – Louisiana Purchase
1810 – West Florida
1818 – 49th Parallel/joint occupation of Oregon
1819 – Adams-Onis Treaty brought East and West Florida into United States
1842 – Webster-Ashburton Treaty; set boundary of Maine
1845 – Texas
1846 – Oregon
1848 – Mexican Cession
1853 – Gadsden Purchase
1867 – Alaska and Midway
1889 – Samoa
1898 – Philippines, Puerto Rico, Guam, Hawaii
1900 – Guantanamo
1903 – Canal Zone
1917 – Virgin Islands

Some Relevant Factual Information

Agrarian Republic/Thomas Jefferson
John Quincy Adams
New Manifest Destiny, 1890s
Manifest Destiny, 1840s
Informal empire
William Seward
Linkage of early nineteenth-century expansion with domestic policies (e.g., slavery and railroads); consider Mexican War, Gadsden Purchase
Large policy (coastal defense, coaling stations, big Navy, canal)
Industrialization
Frederick Jackson Turner
Secretary of State Hay
"Open Door"
Boxer Rebellion
"White Man's Burden" (Kipling)
Roosevelt Corollary
Hay-Bunau-Varilla Treaty
Gunboat Diplomacy
"Big Stick"
Dollar Diplomacy
William McKinley
James G. Blaine

DBQ Scoring Guide (15-point scale)

13-15
1. Strong thesis clearly developed; well organized; analytical, and focused on the question
2. Develops the issues of continuity *and* departure; need not be balanced
3. Sophisticated use of substantial number of documents
4. Substantial relevant outside information (abstract and/or concrete); chronological coherent
5. May have insignificant errors

10-12
1. Consistent, well-developed thesis; clearly organized and written
2. Addresses the issues of continuity and departure; need not be balanced
3. Effective use of several documents
4. Significant relevant outside information (abstract and/or concrete)
5. May have minor errors

7-9
1. Partially developed valid thesis; acceptable organization and writing
2. May discuss only one side of question
3. Use of some documents
4. Some relevant outside information
5. May contain errors, usually not major

4-6
1. Limited, confused and/or poorly developed thesis; weak organization and writing
2. Shows limited understanding of the question
3. Misinterprets, briefly cites, or quotes documents
4. Little outside information, or information which is inaccurate or irrelevant
5. May contain major errors

1-3
1. No thesis; disorganize, poorly written
2. Exhibits inadequate or inaccurate understanding of the question
3. Poor, confused, or no use of documents
4. Inappropriate or no outside information
5. Numerous errors, both major and minor

MARYLAND HIGH SCHOOL ASSESSMENT ITEM EXEMPLARS

Subject: United States History

Item text
(attach scoring guide if necessary)

My party's in power in the city, and it's going to undertake a lot of public improvements. Well, I'm tipped off, say, that they're going to lay out a new park at a certain place. I see my opportunity and I take it. I go to that place and I buy up all the land I can in the neighborhood. Then the board of this or that makes its plan public, and there is a rush to get my land which nobody cared particular for before. Ain't it perfectly honest to charge a good price and make a profit on my investment and foresight? Of course it is. Well, that's honest graft.

This statement was most likely made by

(A) a populist

(B) a machine politician

(C) a tenement owner

(D) an urban merchant

Key: B

This item links to the following Goal - Expectation - Indicators: 1.1 A3 B4

This item is appropriate to the following Skills for Success: 2.2 A3

Source: SAT II US History

MARYLAND HIGH SCHOOL ASSESSMENT ITEM EXEMPLARS

Subject: World History

Item text
(attach scoring guide if necessary)

One of the immediate outcomes of the United States occupation of Japan following the Second World War was the

(A) institution of Japan's first parliamentary form of government

(B) demilitarization of Japan

(C) electoral success of the Communist party

(D) beginning of Japan's industrialization

(E) opening of diplomatic relations between China and Japan

Key: B

This item links to the following Goal - Expectation - Indicators: 2,2 A 1 D

Source: SAT II World History

MARYLAND HIGH SCHOOL ASSESSMENT ITEM EXEMPLARS

Subject: United States History

Item text
(attach scoring guide if necessary)

Questions 1 and 2 refer to the cartoon below.

Bruce Shanks in the Buffalo Ne

1. Circle the decade in which you believe this cartoon was drawn.

1920's 1940's 1960's 1980's

Citing specific historical evidence, explain why you chose the decade you did.

_____

_____

_____

2. What is the main message of this cartoon?

_____

_____

_____

This item links to the following Goal - Expectation - Indicators: 1.2 A4 B4

This item is appropriate to the following Skills for Success: 2.2 A3

Source: NAEP

Scoring Guide(s)

Item 1

An **Appropriate** response explains why the cartoon was drawn in the 1960's and provides supporting detail, such as the occurrence of demonstrations and riots after the law was enacted. Or, the response identifies the 1980's and gives a reasonable explanation, e.g. in the 1980's the spirit of 1960's civil rights legislation remain unfulfilled.

A **Partial** response chooses either the 1960's or the 1980's as the decade, but explains why the cartoon was drawn in the decade chosen in vague terms.

Item 2

An **Appropriate** response correctly identifies the main message of the cartoon as being that passing civil rights laws does not mean that civil rights issues are resolved, and that social, economic, and political (and not just legal) changes were also necessary.

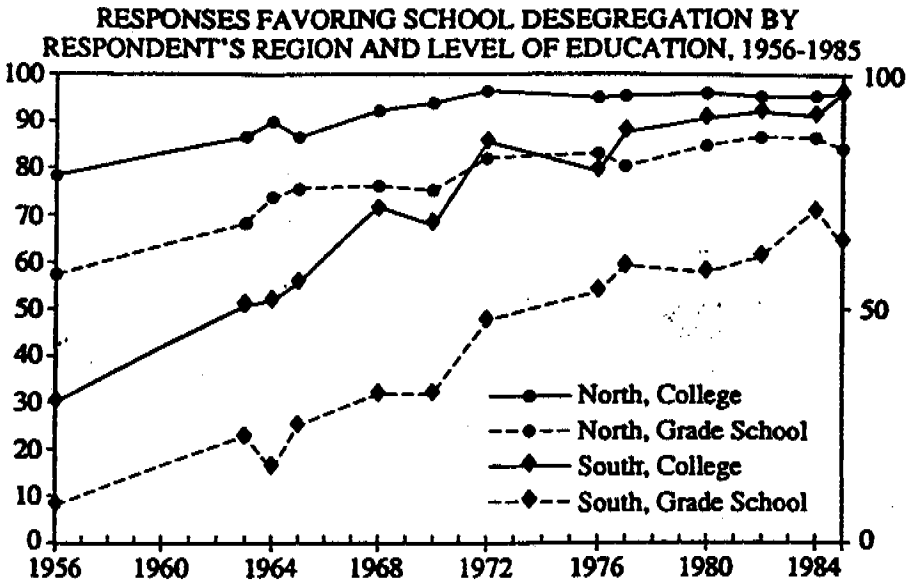
A **Partial** response shows understanding that the cartoon is about problems associated with the civil rights bill but is not able to explain the cartoon in a specific and precise fashion.

MARYLAND HIGH SCHOOL ASSESSMENT ITEM EXEMPLARS

Subject: Government

Item text
(attach scoring guide if necessary)

Short Free-Response Question:



The graph above presents public opinion on the racial desegregation of schools. Using the graph and your knowledge of United States politics, perform the following tasks.

a. **Identify the trends that were evident in Americans' attitudes toward school desegregation.**

b. **Explain why these trends occurred.**

This item links to the following Goal - Expectation - Indicators: 1.2 A2 B4

Source: AP US Government

Scoring Guide (5-point scale)

5 Answer must clearly identify at least TWO of the following trends in public opinion toward the racial desegregation of schools:

(a) Public opinion reflected a mostly increasing acceptance of school desegregation in BOTH the North and South from 1956 to 1984.

(b) Respondents from the South showed the most marked increase from 1956 to 1984 in accepting school desegregation.

(c) College graduates, North and South, tend to be the most supportive.

(d) Grade school educated respondents, North and South, generally became more accepting of desegregation between 1956 and 1984.

(e) Northerners tended to favor school desegregation more than Southerners from 1950 to 1984.

Explanation for trends must include at least TWO of the following: (A clear explanation of a trend will include a specific example that illustrates the point.)

(a) The *Brown v. Board of Education* decision in 1954 seemed to be a catalyst for growing acceptance of school desegregation.

(b) Increased media coverage of civil rights may have gradually changed those resistant to desegregation.

(c) Civil rights laws in 1964 and 1965 broke down segregation in general, and may have eventually moderated some of the antipathy White Americans held toward Black Americans.

(d) The civil rights movement itself convinced many that segregation was indefensible in today's modern world.

(e) Any other logical explanation, e.g., television presentations of Black Americans in a different light, presidential and political leadership, changing role models, or liberalizing effects of college education, etc.

4 Answer clearly identifies two trends, but provides only one explanation, OR identifies one trend and give two solid explanation. OR, it clearly identifies two trends but provides two explanations that are vague or incomplete.

3 Answer provides once clear explanation of the trends of the graph and one explanation. OR it explains two trends, but has no explanation.

2 Answer provides EITHER one trend, OR one explanation.

1 Answer attempts to deal with question, but does not provide either an explanation for the graph, trends, or for the causal factors.

0 No effort is made to answer the question.

– Blank space, frivolous response, or completely off-task.

III. IMPLICATIONS OF END-OF-COURSE ASSESSMENTS

A. Turnaround Time for Scores

The report discussed the most critical issues concerning turnaround time for assessments. Toward the end of the public engagement process, information on turnaround time was presented to principals, superintendents, and the High School Task Force. Some participants in these groups generally felt that the projected turnaround may render the proposed graduation use for the HSA unfeasible. That is, they believe that two conditions must be met that appear operationally incompatible. These are:

1. Assessments MUST be administered as close to the end of the course as possible to have educational value; this is especially critical for schools on semester-block schedules, and
2. Results must be available to students well before the end of the course to provide schools an opportunity to (a) counsel students, (b) propose and plan for needed remediation, and (c) retest students within close proximity of completion of the course. This is also necessary to provide students with opportunities to (a) explore options, (b) enroll in remedial programs or undertake other learning activities, and (c) retest within close proximity of the end of the course.

Basically, many principals and superintendents believed that these two conditions MUST be met out of basic fairness to students. Anything less was unacceptable and unfair.

Unfortunately, today, both conditions cannot be met. An aggressive plan to develop and implement a computer-based testing program could overcome these conflicting needs; however, the costs and delivery problems would far exceed those of any design options included in this report. Eventually, as technology expands, costs decline, and access increases, the HSAs could be migrated to computer-based testing. We recommend that MSDE consider this as one goal of the test development activities -- to ensure that development proceeds in a way that would facilitate the eventual migration to computer. However, a comprehensive operational testing program such as the HSA, with the projected volume of test takers, could not be conducted in the overwhelming majority of the state's schools for the foreseeable future. An additional restriction is that computer-based testing (CBT) cannot currently accommodate the types of performance assessments widely called for by educators and proposed for the HSAs. This leaves us with the above conflict which is more fully framed earlier in the report.

One alternative that could minimize the conflict would be to implement two separate administrations for each assessment. That is, each assessment would have two components: (a) performance assessments administered 7 weeks before scores are required, and (b) Selected Response items (multiple-choice, grid-ins) administered about 3 weeks later. Scores from both components would be combined and released as a final composite score. However, this would possibly double the scheduling burden for schools, increase the proportion of students not completing the test (since students would need to be present on two testing dates for each assessment, rather than just one date), and probably increase logistical, operational, scoring, and database costs and burdens for the state and schools.

A second alternative would be to relax assumptions concerning the "end-of-course" nature of the assessments. That is, if assessments basically measured those CLGs generally covered in the previous course in the content area, as well as the initial indicators that might be covered in the current course, assessments administered in November and March would not be prohibitive. For example, assume that the three English exams are to correspond with 9th, 10th, and 11th grade English courses at most schools. Students enrolled in 10th grade English would be given an assessment in March which covers the CLGs from 9th grade English (English assessment I) and that portion of the CLGs from 10th grade English commonly covered early in the year across all districts (English assessment II). This approach might also be used in math and social studies, although there are certainly more complex instructional issues involved in these subjects. Because science courses do not appear to have one or two common sequences for students, additional variations would be required. For example, students completing the Biology course in June might take the assessment the following October or November. Additional difficulties would emerge for seniors if they take any courses associated with the HSAs in their final year.

B. Score Reporting and Scheduling Test Administration in the Schools

School administrators in particular asked what a typical school testing schedule might look like. Table A2 below illustrates a potential schedule for testing students in April/May. This would need to be duplicated in November/December for schools on a semester-block schedule. A word of caution: this is only a proposed schedule that might be applicable across the state. The primary need is to schedule assessments to minimize the probability that the same student would be required to take two different tests at the same time on the same date.

Table A2 Proposed Statewide Testing Schedule

Time Slot	Monday	Tuesday	Wednesday	Thursday	Friday	Make-Up Day *
1 AM	ENG. 1	ALGEBRA	BIOLOGY	EARTH/ SPACE SC	WORLD HISTORY	all tests offered
2 AM	ENG. 3	GEOM.				"
3 PM	ENG. 2	CHEM.	US HIST.	PHYSICS	GOVERN.	"
4 PM						"

** All tests would be offered for students absent on the day of the test. Alternative forms or mixed formats would be employed to minimize security concerns.*

In the above example, both the AM and PM slots would be 2-2½ hour blocks for testing determined by the district or school. On Monday, students taking the English 1 assessment might go to rooms A, B, C, and D; students taking the English 3 assessment might go to different rooms. Any students required to complete both these exams during the same testing period would need to take one of them on the make-up day. The same would be true for Algebra and Geometry on Tuesday. However, we expect these combinations would result in the fewest number of students requiring both tests offered at the same periods on the same day.

IV. FEASIBILITY OF ACCOMMODATING SPECIAL CIRCUMSTANCES

Throughout the public engagement activities numerous representatives from various key constituency groups expressed concerns about individual students and specific groups of students, and potential the negative consequences they believed would result from the current plans for the HSA. Many of the specific issues addressed in this section concern operational aspects of the program that can largely determine the "buy-in" or "acceptability" of the HSA for many of these key constituency groups. These operational issues can only be addressed once MSBE considers the delicate and often controversial policy issues that are involved in the programs operations. Exceptions, accommodations, and special circumstances for individual students and groups of students must be considered by MSBE before many of these operational issues can be fully addressed.

This section attempts to contrast the desires expressed by many groups for exemptions, accommodations, and special circumstances with the implications and operational realities for the HSA. One strategy for reaching closure on some of these issues would be for MSBE to better define the range of acceptable options and then further engage key stakeholders in arriving at a final resolutions.

Two operational issues which impact the entire program have already been discussed in the body of the report:

A. Student Absences And Make Ups
B. Alternative Evidence Of Competency

These issues must be resolved by MSBE for Phase II work to proceed on schedule. The remaining issues of accommodating special circumstances are of equal importance to the first two issues, but they do not require an immediate resolution by MSBE in order to proceed with Phase II work.

C. Process to Begin a Revision of the *Requirements and Guidelines ... for Assessment Programs*

Currently, exemptions, accommodations and special circumstances for Maryland's state assessment programs are documented in the *Requirements and Guidelines for Exemptions, Excuses and Accommodations for Maryland Statewide Assessment Programs* (Maryland State Department of Education, 1995). This document lists the acceptable exemptions, excuses and special accommodations for three types of students:

- Limited English Proficient Students
- Students with Disabilities
- Transfer Students

Separate but consistent requirements and guidelines have been developed for the major state testing programs (Functional Testing Program, MSPAP, and CTBS/4). The *Requirements &*

Guidelines.... document will require revision to incorporate the HSA. Several types of revisions will be required. First, in addition to the three types of students mentioned above, any state approved exemptions for students in accelerated courses and examinations (see the Report, V, C) such as Advanced Placement or International Baccalaureate may also need to be incorporated. Second, approved accommodations for the HSA must be added. Third, documentation must be developed which details how students performance on the HSA will be used to determine any high stakes uses for individual students (e.g., graduation, eligibility for retesting) and schools/districts (e.g., how school scores will be computed, formulas for inclusion of data from excused students). Perhaps the general guidelines for the last category of issues will require a new document, but exemptions to these guidelines may still be cited in the revised edition of the *Requirements and Guidelines* (MSDE, 1995).

Maryland must begin to consider a process to undertake a revision of these requirements and guidelines and to develop additional documentation for the issues discussed above and how best to involve key stakeholders in the process to ensure their concerns are adequately addressed. While it is premature to embark on a revision of the document until more operational issues concerning the HSA program are decided, MSBE must be aware that once all these issues have been determined it may be unfeasible and financially prohibitive to incorporate additional modifications and accommodations. A coordinated and iterative process that has the general test and program development process working collaboratively with a task force that is both proposing potential accommodations, exemptions, and exceptions, and reviewing plans for the full program, may be most appropriate for the HSA.

Specific concerns and issues which emerged with special groups of students are addressed in the remainder of this section.

D. Students with Disabilities

The presumption from the beginning of the HSA design effort has been that students with disabilities or special education needs should receive the same accommodations that are provided for MSPAP. In many cases these accommodations are specified in the student's IEP (Individual Education Plan) and are similar to the accommodations made for instruction. These may range from allowing extra time to take the assessments, to providing readers or audio tapes, to providing an amanuensis, or providing a large-print edition. Advocates for these students would like to see additional accommodations incorporated into the HSA. The two such additional accommodations most frequently proposed were: (1) breaking the assessments into smaller sections, or modules, that can be administered directly after the topics have been taught, and (2) using computers as tools for administering and responding to the assessments.

It should be noted that much of the burden of providing the current accommodations falls on the school the student attends. The assumption is that MSDE would approve most accommodations which are currently permitted under existing testing programs. However, there are some areas where these or additional accommodations may cause additional problems which must be successfully resolved. For example, security concerns will increase as school flexibility increases for scheduling test administration. This is a substantial concern for a high stakes graduation test

and becomes more difficult to manage when a substantial proportion of a test is devoted to one or two constructed response tasks that can easily be recalled and provided to students yet to take the test. Current MSDE policies overly permission in this area.

The request to break the test into modules or tasks that would be completed immediately after instruction also presents a non-standard condition that would impair efforts to compare scores of with students required to complete the entire test at the end of the course. Combining student performance on test items and performance tasks completed after instruction into a single test score is simply not comparable to scores produced from an end-of-course test. In addition, the "modularize" assessment design proposed also raises the same security concerns mentioned above[3].

Computerization of the assessment tasks and responses may also present unique problems. For example, many of the performance tasks considered for some design options would use detailed graphics, photos, or other visual stimuli that would not lend them selves to display on a computer currently. Similarly, the types of responses that will be elicited with some performance items may require drawing, graphing, and other skills which would be more complex or prohibitive when responding on a computer. Of course, the multiple choice sections, short answer questions and essays could both be administered with a computer and students could provide their response with a computer using current technologies if schools have the equipment and MSDE has the resources to implement this accommodation for a portion of each test.

Finally, special needs students must be included in all pretesting to ensure accommodations and exemptions are appropriate and manageable. Score comparability studies can also be undertaken and such pretesting will also help to familiarize the community with the tests, student expectations for performance, and accommodations. However, the primary value of such pretesting is to ensure that students with varies types of disabilities can, in fact, complete the various types of tasks and items that will be included in the assessments and can effectively use the accommodations that are provided. These are only a sample of the operational issues that must be confronted before approving existing or additional accommodations. Research shows that scores from tests completed under many non-standardized conditions (e.g., extended time) are not comparable to scores for students testing under standard conditions. To the extent that more exemptions and accommodations are introduced for more and more students, issues of fairness and comparability must be addressed. Tests administered under non-standardized conditions are not generally comparable to those administered under standardized conditions. Relying solely on test scores for any important decision (e.g., graduation, school-level rewards/punishments) raises serious professional, ethical, and legal issues for Maryland given this concern and others raised throughout the report.

E. Limited English Proficient Students

Maryland has a significant number of students for whom English is a second or third language. Some of these students receive special instruction so that they can progress in various subjects

[3] However, these security concerns may be minimized as additional test forms are developed that can be used for creating alternative forms for such special accommodations, make-ups or retesting.

while learning English. Currently, the only accommodation available for LEP students is extended time on the MSPAP. In most instances, LEP students can be exempted from one or more administrations if they have appropriate documentation.

It would appear that exemptions would still be required for many LEP students who do not have the requisite skills in English. During the public engagement activities there were comments supporting offering mathematics, science, and social studies assessments in languages other than English; there also were comments arguing that a Maryland diploma should mean that students function sufficiently well in English to be able to pass high school assessments in that language.

Offering an assessment in a language other than English entails much more than translating the words. The translations themselves would require substantial resources for the number of tests and test forms currently proposed for the HSA. However, even more problematic are the underlying assumptions and conceptual models which will vary significantly from those used by English speakers. It would take a very considerable effort to ascertain that assessments in other languages were equally valid measures of the Core Learning Goals and that the reported scores were comparable to those derived from the English-language assessments.

F. Students Enrolled in Evening School, Alternative, or Private Schools.

Some Maryland districts have indicated that they have significant numbers of students completing their high school graduation requirements through night school or other alternative instructional programs. The presumption is that such students will also be subject to the HSA requirements. Yet the implications would need to be considered and a phase-in of HSA for these populations would certainly be required initially. It would appear that private schools could voluntarily participate in the assessment program if there was significant interest and approval from the state. Of course, there would need to be assurances that the same standardized conditions for administration, scoring, and quality control of assessments are incorporated in any migration of the assessments to non-public schools if there are to be comparisons of student scores across schools and programs. The additional costs that would be incurred for expanding the program beyond public schools would also present some policy issues for MSBE to consider.

G. Transfer Students

There have been repeated comments from a variety of constituencies about the transient nature of segments of Maryland's population and the need for a policy on implementing the HSA requirement with students who transfer into Maryland during high school. Several alternatives have been proposed during public engagement activities:

- Transfer students should be held to the same requirements, i.e., passing 10 HSAs, if they enter Maryland schools prior to the last semester of the senior year. The precedent of the community service requirement is cited in support of this position.
- Transfer students should be expected to take and pass the end-of-course HSA for any courses they take in Maryland schools, but the requirement should be waived for courses successfully completed before coming to Maryland.

- Transfer students should be expected to pass some minimum number (less than 10) of HSA assessments regardless of whether they take the courses in a Maryland high school.

These are complex policy issues that must be resolved with input from local school districts on the feasibility and implications of the various alternatives. Some of these options would appear to hold students responsible for content which they have not been exposed to and would raise legal risks for the state and district. It would seem unfair to hold students responsible for any assessment which corresponds to CLGs which that they have not had an opportunity to learn.

To handle within-state transfers, as well as to maintain the data generated from the HSA, it has been suggested that there should be a central data base of student scores, maintained by MSDE, that would provide the "official" record of student performance on HSA, regardless of whether they had moved from one high school or district to another. Others have supported such a data base as a means of verifying school records or recreating missing records. Further, the data base would be critical to any longitudinal research involving HSA. MSDE should consider the resources required to develop and maintain a relational database that can both serve the HSA program and other state and local needs for student and school-level data. Additional staff or external contractors would be required to maintain this service.

H. Testing Middle School Students

The HSA requirements are currently designated to be implemented with the Class of 2004. Parents and teachers have repeatedly pointed out that some significant portion of that class will already have taken certain requisite courses (e.g., Algebra, Earth/Space Science) as 8th graders in 1999-2000, the academic year before implementation. A number of alternatives have been suggested for dealing with this disjunction:

- Waive the HSA requirement for those courses taken by the Class of 2004 while in middle school.
- Require students who take such courses in 1999-2000 to take the corresponding HSA the following year as 9th graders.
- Accelerate the development of the affected assessments so that they can be administered in the Spring of 2000.

Many stakeholders also raised additional concerns that many 8th grade students will both be expected to participate in some high school assessments (this is most likely to occur with Algebra and some science courses) as well as MSPAP. There were concerns both about student motivation and the loss of instructional time.

V. FLEXIBILITY IN ACCOMMODATING VARIATIONS AMONG LOCAL DISTRICTS

If assessments do not reflect the curriculum offered in a school then they lack curricular validity for the educators who must implement them and the students who must complete them. Early in the conceptualization of the HSA program, it was proposed to modularize certain of the

assessments in order to permit local districts greater flexibility in selecting the configurations which most reflect its curriculum. The idea of modules was particularly salient to mathematics and science because a number of districts were known to have implemented integrated courses which were not congruent with the proposed set of assessments. After a great deal of study and discussion, this report recommends that the idea of *modules not be pursued* but rather, that the major variations in course offerings can be better accommodated through the addition of two additional assessments each in the areas of mathematics and science. MSDE staff have gathered data which indicate that these four additional assessments will accommodate the vast majority of students enrolled in integrated mathematics or science courses.

Modules would require substantial additional costs for development and implementation, would introduce additional burdens to the assessment and management systems, and raise a number of significant psychometric concerns in comparing student performances and ensuring a fair and level playing field for all students. Assessments composed of different combinations of modules would result in differing levels of difficulty among the assessments--resulting in some schools administering a more difficult assessment that in other schools--an inherent inequality. Each combination of modules would require additional work comparable to that for a distinct assessment:

- setting the standard or proficiency levels for passing
- scaling the reporting scores
- equating the reporting scores
- programming score reports
- preparing interpretive information

In addition, designing the assessments to be split into modules would put severe limitations of the specifications for each assessment, for example, on the content of an extended response question or in giving heavier treatment to a particular topic solely to fill out the content of a module.

Not only would it be nearly impossible to scale or equate a number of different configurations of the same assessment based on different modules, it would be impossible to equate subsequent forms of each assessment to ensure that a "3" on one form has the same meaning as a "3" on other forms of the assessment. In some instances, it may be impossible to scale or equate all configurations of these modules to previous forms of the test because insufficient numbers of students complete a module or due to other psychometric constraints.

Modules would complicate the database and systems required to keep a cumulative record of a students' progress on the HSAs. Administrative burdens would increase as districts must determine well in advance how many students at each school would take each of the possible combinations of science or mathematics modules. As simple a task as ordering adequate quantities of testing booklets becomes a complex task for each district. If districts ordered the wrong quantities of test booklets for specific modules, students would be denied taking an assessment at the end of the appropriate course .

Finally, parents and the general public might be confused when students at different schools complete assessments composed of different combinations of modules but are still held to the

same graduation standard across these assessments. Public perception would be particularly problematic when different proportions of students pass different combinations of modules. For example, what happens if 75% of students pass the biology and chemistry assessments when offered at the end of the respective courses, but a different population of students taking a Bio/Chemistry modular test achieve only a 50% pass rate. While the objective would be to equate all forms and combinations of these assessments, the public might attribute performances differences to the test content and difficulty rather than to differences among students taking the alternate combination of modules.

Because of the many difficulties associated with the concept of modules, it is likely that the expenses connected with separately defined assessments which correspond to commonly occurring integrated courses in mathematics and science will be no greater than those associated with the development. implementation, and administration of assessments composed of modules. Separately defined assessments will also avoid a number of the constraints inherent in the modular model. In addition, they can be designed to be more congruent with the alternative courses than if they were composed of modules which had to be appropriate to both traditional curricula and to integrated curricula., e.g., an assessment in Environmental Science could situate a biological principle in the context commonly taught in such courses, rather than in the context or application used in teaching the same principle in a Biology classroom.

The proposed makeup of an Environmental Science assessment from a biology module and a chemistry module raises several areas of concern. One is that the assessment would lack several key content areas. Second is the philosophical concern with teaching a topic in one context and testing the same topic in another context.

In 1996, the Advanced Placement program of the College Board conducted a survey of college and university biology departments to determine what content areas are covered in introductory environmental science and ecology courses. They also conducted a survey of several of the major college texts on environmental sciences. The results indicated that current environmental science courses are a composite of material taken from the disciplines of biology, chemistry, physics and geology. While the Environmental Science module will not cover the material at the same amount of depth that the AP Environmental Science course does, it is important to have the module cover material that is considered important to the discipline. The currently proposed Environmental Science module arrangement lacks coverage in the area of physics (notably energy transfer - Biology: Expectation I Indicator 3; Earth/Space Science: Expectation III Indicator 1, 2, 3) and the geological sciences (notably matter cycling - Biology: Expectation I Indicator 3; Earth/Space Science; Expectation V Indicators 1, 2; dynamics of the solid earth - Earth/Space Science: Expectation III Indicator 1, 2, 3; Expectation IV Indicators 1, 3, 4, 5; Expectation V Indicators 1, 2; and the atmosphere - Earth/Space Science: Expectation III Indicator 1, 2, 3; Expectation V Indicators 1, 2).

In the proposed model, concepts would be transferred from a Biology or Chemistry context to an Environmental Science context. For example, the concept of pH would be transferred from the Chemistry Goals to the Environmental Science Goals. It will be very difficult for students to make the transfer of the concept of pH from the Chemistry goals on their own. The concept of

pH should be taught in an Environmental Science context not in a Chemistry one. There are also challenges to consider if the candidates and public are told that the test items are coming from the Chemistry Core Learning Goals, but the test items only refer to pH in an Environmental Science context.

Many additional suggestions for introducing some level of local choice and flexibility into the HSA program have been discussed in the full report.

VI. SUPPORTING STUDENTS WHO DO NOT DEMONSTRATE COMPETENCY

Here, the report addresses some additional issues which must be considered by MSBE in the next few months. MSBE must consider ways to implement the HSA so students are not adversely effected, but rather benefit from the higher standards and expectations associated with the assessments. Students should have adequate opportunities to complete the assessments or demonstrate generally comparable levels of competence on the CLGs through other methods. In addition, students who are not initially successful may need some form of remediation or additional instructional opportunities to overcome any weaknesses of deficiencies in skills related to the CLGs.

In addition, MSDE must establish procedures to document that students in each district have had one or more opportunities to learn (OTL) the CLGs that are incorporated in the various assessments. One method for collecting such data is administering surveys to teachers and students simultaneously with large-scale assessments (Herman, Klein, Heath & Wakai, 1995). Delaware has developed case studies relating to OTL that go beyond linking OTL only to student achievement and have uncovered significant OTL between-county differences (Winters, 1995). Because of the high stakes uses intended for the HSA, MSDE should consider multiple methods to establish OTL before the program becomes operational.

A. Alternative Evidence of Competence[4]

The High School Assessment Task Force proposed the development and use of alternative mechanisms for students to demonstrate competence in the Core Learning Goals without having to repeatedly retake tests. They recommended that the state provide assistance to local districts for developing the alternatives, training staff, and ensuring that the options and the HSA assess comparable levels of proficiency on the Core Learning Goals.

If students who do not pass the state tests are to have alternatives for demonstrating competence in the same Core Learning Goals, the implication is that these "alternatives" might differ from the HSA in terms of format, task assessed, or period of performance. Such differences, even when

[4] Much of the discussion concerning the application of modules (see pages 21-22 in the report and Section V of Appendix A) is relevant to the discussion of different options for demonstrating competence. Whether different modules or entirely different assessment options are employed, the same psychometric difficulties arise in attempting to compare student performance across districts. Also, many of the same logistical, administrative, scoring, and financial issues are similar with both elements of an assessment program.

content covered is generally similar, will result in assessments that are not statistically equivalent. Thus, it would be inappropriate to assert that a student obtaining a proficiency score of 3 on the HSA has demonstrated the equivalent proficiency of students obtaining similar scores on alternatives offered by LEAs. If each (or most) district develops its own alternatives, it would be impossible to establish psychometric links to derive any common meaning or understanding about comparable performances. That is, it is nearly impossible to use alternative assessments for norm-referenced purposes.

There are measurement models that may permit different assessments to be linked in ways that would make scores comparable. Unfortunately, this would involve using a block of common items in each of the alternative assessments. This would generally preclude portfolios, extended projects, or most locally developed instruments for assessments. The only possibility that appears feasible would be to employ one or two state-created alternative assessments that are linked to the HSA by including common items. It is not clear, however, that this would be appreciably different from requiring students to repeatedly retake the HSA.

To certify that a student has demonstrated competence on the Core Learning Goals at as demanding a level as that represented by a passing proficiency score on the HSA, a rigorous process could be put in place. Content experts would be asked to examine student work collected under different conditions and to judge whether the student had demonstrated competence on the CLGs measured by the particular HSA. Student results generated in this way could not be intermixed with HSA results nor could they be compared with HSA scores. One could say, however, that based on a review of alternative evidence, the student was judged to have demonstrated competence on the relevant set of CLGs. If a model involving composite HSA scores is chosen (i.e., one in which the required score may be achieved by compensating for low scores in some subjects with higher score in other), there is no apparent way to include alternative scores in the composite.

If the state is primarily interested in criterion-referenced uses, and not normed-referenced comparisons, alternative assessments would appear more feasible. This is an important distinction for addressing the consequences and options for students who fail an assessment (or do not demonstrate mastery on the CLGs).

If the primary purpose of assessments is not to make inter-student comparisons but rather to determine if a student has demonstrated competency on the Core Learning Goals, assessment options that differ marginally in coverage of content (the extent that measure the same exact goals and expectations), format (require the same type of performance), duration (end of course tests versus extended projects, oral reports, portfolios), and statistical specifications are less problematic. Performance on alternative options could be used to make decisions about whether a student has demonstrated competency on the Core Learning Goals (let's say they have achieved a performance of a 3 on a 5-point scale), but not whether the performance is statistically comparable across the various assessments. The assessments would be on scales that are not strictly comparable, yet a 3 on any option would be set to represent competency in the Core Learning Goals as judged by content and measurement experts. Under these assumptions the state

could develop one or more alternative options that would be employed for students who fail the tests, or assist LEAs in developing their own options.

As noted above, as alternative options differ (in format, content coverage, and other essential elements) so will they differ in the meaning of their scores. One cannot derive statistically comparable meaning from performances that differ among individuals in breath, scope, difficulty and content. The more alternative options that are employed the greater the variation among them and the less comparable will be any interpretation of student performance. *If the primary concern purpose is to "certify" that a student has demonstrated competency in Core Learning Goals, this is less of a problem. If the primary concern is to ensure that students with a "3" on assessments across all districts and schools represent the same level of competence on the same indicators and CLGs, then this will be much more difficult and problematic.*

It appears most feasible to limit the number of alternative assessments to possibly just one or two alternatives per test during the initial years of development because of costs, logistical issues, and burden on the state and schools. Whether the alternatives are developed by the state or a collaborative effort between the state and one LEA, MSBE should consider the benefits of having implementation coordinated by the state to permit students in all districts access to the alternative. If districts differ in their local assessment options, then the public and policymakers will question the degree that performance differences result from "easier tests" in some of the districts, and the credibility of alternative options across the state will be questioned. A completely local model of assessment will be difficult to defend because assessments and students' performance will not be comparable, yet the same high stakes will be employed across all LEAs.

If alternative assessment options for students who initially fail the HS tests are to be considered, we recommend a separate governance committee be formed by MSBE to consider the various issues and concerns discussed in this report. These issues would include: what criteria will be used for their development? Who will approve their use? How will district and school comparability on the HS assessments be computed? What range of alternative options can be considered? Who will pay for the development of the alternative options? What psychometric evidence will be available to evaluate the options?

Does MSBE want to defer any further exploration of alternative means of assessing competence until after the design of the HSA program has been completed?

B. Opportunities for Students to Retake Tests They Do Not Pass

The design work so far has assumed that students should have the opportunity to retake a particular HSA at least once at a subsequent administration, presumably following a remedial effort by the student and school and review of the course material. Is one opportunity for retesting sufficient, or should students be expected to retest several times? The High School Assessment Task Force articulated some of the concerns that arise when students are required to continue to retake tests until they obtain a sufficient score:

- Remediation efforts are almost exclusively focused on the test, because additional instructional and learning opportunities are limited.
- Students doing remedial work may have fewer opportunities to take subsequent courses and may be denied access to a high-quality education.
- Because these are end-of-course tests, if a student has to take the tests several months after the course has been completed, he or she may be at a greater disadvantage the second time around unless there has been some form of instructional intervention.

Does MSBE want the HSA design to provide for a student who initially fails an HSA to re-take that particular HSA more than once? Will the provision be mandatory or voluntary

C. Remediation Programs

Few people assume that students who initially fail the HSA can achieve a passing level of proficiency without intervention. Consequently, the education system in Maryland must decide on the types of remediation that will be available to students who initially fail the exams.

Many school administrators believe that an extensive summer remediation program will be required in each district for the 12 exams. This model requires extensive resources for teachers in all 12 areas, transportation for students in the summer, and other expenses that are not currently budgeted. Schools employing a semester-block schedule are concerned that summer remediation will not be adequate because students move to the next level of courses (and state tests) during the second half of the year.

While policies regarding remediation do not have to be resolved prior to Phase II of the design process, MSBE, MSDE, and the LEAs together must begin to consider the extent and nature of remediation that is envisioned for HSA.

When must MSBE begin examining the feasibility of implementing remediation programs and other learning interventions to support students who do not demonstrate competence on the HSA?

VII. OTHER UNRESOLVED ISSUES

Many additional issues emerged during the public engagement activities that MSDE and MSBE should consider and address. Among these issues, two are most critical and should be examined in the next few months. Both of these relate more to the general directions and objectives of Maryland's School Performance Program than to the specific design of the HSA. The HSA is one component of the Program. During public engagement many educators and other participants emphasized the need to both (a) address the preparation of students in K-8, and (b) develop the necessary remediation programs and other learning interventions needed for students who do not initially succeed on the assessments.

A few additional issues are also mentioned below because there was substantial interest and concern was expressed about them, and because they can determine the success of the HSA implementation and buy-in of HSA by the key stakeholder groups.

A. K-8 Instruction - To prepare students for the high-stakes, individual accountability dimension of the HSA, students in the earlier grades must be adequately prepared. There are various educational reform efforts underway at the K-8 level that are not part of this report and with which project staff from the College Board and ETS are only marginally familiar. MSPAP is one of these efforts, yet because individual student scores are not generated, many participants believe there is not an adequate "early warning system" to inform teachers of students at risk. A system is required to enable educators to inform parents and students of students' strengths and weaknesses in the skills and key learning processes that underlie the CLGs. If students do not enter the 7th through 12th grades with adequate preparation, it is naive to expect them to meet the high standards desired by the state. Education involves many components; assessment is just one of them. Instructional and curricular reform, staff development, parent and community involvement and support, and technology, resources, and access are just some of the additional components that participants identified as currently lacking but essential for moving students toward higher standards. Participants generated many such questions that are better addressed and considered by MSBE and MSDE.

B. Remediation - In any individual assessment system, some students will not meet the performance standards. MSBE must not only consider options for students to retake the assessments and/or demonstrate their competencies on the CLGs through alternative methods, but also consider what types of remediation will be in place by 1999-2000. Local school and district administrators emphasized that they will look to the state for guidance, and, to some extent, for additional support in this area. One educator asked (the following is paraphrased):

"Fully 1/3 or more of the school's students may initially fail one or more tests (given the stated high standards and results from MSPAP). Remediation must take place over the summer unless we are going to make students retake courses that they successfully completed and spend 6 years in high school.... Well, we will need test scores in April to plan for such a major summer remediation, but that doesn't work educationally. Where will I get the teachers for this program, the money to pay for them, and the buses required to bring the students here...? And we're better off than many other districts."

The questions posed by this educator must be explored in the next year and considered well before assessments with high stakes are administered. The systems, staff, and resources required for remediation cannot be estimated given the current number of open-ended issues surrounding the HSA, but these are real issues that will ultimately impact the success of the HSA.

C. Technological Applications - The use of technology is growing rapidly and is an increasingly important influence in both education and assessment. Educational institutions are increasingly ready to use technology, and the expectations at all levels are changing. The signs are many for this: distance learning is increasing; the increased emphasis on diversity suggests different ways of characterizing one's ability to succeed; on-line applications for college admissions and electronic

transcripts speed up the admissions process; and the increased availability of instructional software designed for the school has resulted from increased access to computers in many classrooms.

While computer-based testing has the potential to provide many benefits -- for example, increased frequency of test administrations, use of new and innovative item types, immediate score reporting -- there are many barriers that must first be addressed before computer-based delivery can be accomplished and affordable for a large-scale program such as the HSA.

For example:

- Development costs for computer-based tests are still extremely high. National programs that have maintained both paper-based and computer-based programs may charge roughly two or three times as much for the computer-based test. The higher fees required for computer-based tests reflect the additional costs incurred with such programs (e.g., large item pools required for security purposes, additional equating and comparabilities studies needed). Cost per candidate can be several times higher with a comparable computer-based test today.
- School-based delivery is still a major obstacle because many schools still lack the hardware required to administer large-scale testing programs. In the case of the HSA, a Maryland high school would need to accommodate testing for virtually all students on one or more tests within a fairly brief period of time (say 4-5 weeks) if end-of course assessments are to be given. Scheduling 200-500 students to each complete a 2 or 3 hour test (in many cases students would need to complete 2 or more tests) in such a narrow window of time would present significant logistical burdens to schools and require substantial investments by schools. If computers are located in classrooms, it may be difficult to schedule students for testing during instructional time, and educators have not been supportive of testing students at the end of the day when fatigue and extra curricular activities enter into the equation.
- Security of items requires a secure server. Tests cannot reside of the hard drive of a stand alone PC or even a local server. Tests must reside on a secure server where there is maximum security and limited physical access to ensure the integrity of the test items and data. Schools would also require additional staffing to schedule the testing and maintain the security. Staff with expertise in technology would be required to maintain the equipment and load additional enhancements to the programs and tests.
- Maintaining a paper-based test (for schools without technology resources) and a computer-based testing program simultaneously can be prohibitively expensive and raise á number of difficult psychometric issues concerning the comparability of results. Substantial psychometric support and additional research would be needed at MSDE on an ongoing basis to support such a model.
- Many of the performance tasks envisioned in the subject areas would not conform to current models for computer-based testing. Scoring of essays would be no more efficient unless remote scoring networks can be established or until automated scoring models advance further. However, even with such advancements, only certain types of performance tasks would lend themselves to computerization -- many tasks used in paper-based performance assessments would not be efficient in a computer-based modality. In addition, many of the performance tasks would require similar amounts of time to perform whether completed on a

computer or in a paper-based program and would not necessarily result in shortened testing time.

- Today major discrepancies and inequities exist both across schools and in students' homes regarding access to and familiarity with computers. Most students would need additional familiarization with completing assessments on a computer before they could take a high-stakes test under such conditions. Tutorials and other instruction provided in advance and at the front end of today's computer-based testing requires both additional student time for assessment and adds substantial costs to the delivery.

Many proponents of computer-based testing respond with a statement like the following..."*in (you fill in the blank here) years none of these issues will be relevant, all schools will be wired, all schools will have computers, and all students will be adequately familiar with computers that they will expect to take tests with them.*" Well, this may be true; however this is not a certainty. In five or ten years all Maryland schools may be wired to the World Wide Web with ten times as many computers as today; however, we cannot assume that major discrepancies between the quality and access to technology will no longer exist. In addition, as noted above, computerization requires much more than hardware and wiring. The location of the computers, the security of the server(s), and the availability of staff to monitor and schedule student testing are all issues that must be successfully resolved. Finally, while the general public has an expectation that computerization drives costs down, the experience in assessment has not borne this out. Today, computerized testing requires substantially more cost than paper-based testing.

ETS in particular, but the College Board as well, has substantial experience in developing and operating computer-based testing programs. ETS is, in fact, the world's leader in developing and operating such assessment systems, with programs such as the GRE, GMAT, TOEFEL, PRAXIS and many other tests currently delivered or transitioning to computer-delivery in the near future. The College Board's Accuplacer was the earliest computer-based placement test and is currently used by many of Maryland's colleges and universities for course placement. A computer-based SAT has recently been developed and was taken for the first time this month by students applying for gifted and talented programs at Johns Hopkins and other such national centers.

Someday many of these obstacles, added costs, and other operational barriers will be removed. Because we do not know when, how, and in what ways (e.g., technology, thus delivery, may be very different than currently envisioned with hardwired computers), we do not believe it is feasible to develop the proposed HSA end-of-course assessments for computer-based delivery in the near future. However, MSDE should ensure that the ultimate design does not evolve in ways that would prohibit or impede the eventual transition of the assessments to computer-based delivery or prevent the state from taking advantage of other technological developments that are certain to evolve. As technology advances, these issues will be successfully resolved in national assessment programs, and the HSA, as well as other state assessments, can leverage these advances and "lessons learned" by organizations such as the College Board and ETS, in migration to technological platforms.

Other Issues (also see discussion in Appendix B)

D. Functional Tests - Many educators insisted that the Functional Tests should be eliminated well before the HSA is operational because the burden and the differences between these two assessment systems would impact the acceptance of the newer assessments.

D. Staff Development - This is mentioned above, but there are substantial concerns that teachers will need additional staff development to provide quality instruction in the CLGs. Curriculum and content specialists, and other educators noted that the CLGs provide new, dynamic models for instruction and that some teachers will need help in adapting to these challenges. In addition, staff development is likely to be required for the HSA. While the Portfolio Plus and Preparation Plus design options appear to require more staff development than the other options, some training needs are associated with all design options. If teachers are to be used to score the performance assessments associated with the first three design options, substantial amounts of released time will be required for many teachers to receive training and to score the assessments at several times during the year. The costs for staff development will likely be much higher if the portfolio plus design is selected since educators will require substantial training in assembling and scoring samples of student work to enable student comparisons.