I. Introduction

Each member of the Math A Panel is very passionate about the importance of learning mathematics. Each member of the Panel either has taught, is teaching, or is using mathematics in his/her professional position on a regular basis. We are all lovers of mathematics, and we want our children (which we define to be *all* of New York State's children) to be proficient in mathematics. Each member of this Panel applauds the efforts of the Board of Regents and the Commissioner of Education to provide ***all*** children with access to high level mathematics curricula. We have seen very positive outcomes of these efforts, and we want to go on record as supporting the continuation of efforts to provide high quality programs to every child.

In this report, the Panel has been guided by the nine elements of our charge. (See Appendix A.) We viewed the charge as focusing on these broad areas:

* **The Math A standards**. What do we expect our students to know in Math A? Are there improvements the Panel can recommend?

* **The Math A assessment**. How is the Math A Regents exam created and scored? Are there improvements the Panel can recommend?

* **The infrastructure**. How prepared are New York State's schools to help every child reach the Math A standards? Are there improvements the Panel can recommend?

In our discussions, additional issues were raised that do not "neatly" fit into the above categories. These are also addressed in this report.

The Panel felt a tremendous weight of responsibility. All of its members are keenly aware of the Math A graduation requirement. This report is not about Math 4 or Math 8, tests to see if children need assistance. This report is not about Math B, the advanced math test that will be taken by most students heading to college (and certainly all students contemplating a future in a field requiring a strong mathematics background). This report is about Math A, an exam that must be passed before a student can receive a high school diploma. The estimates with which we were working for our Interim Report were that approximately 2/3 of the students failed the June 2003 Math A exam. (These estimates did not include data from New York City, which were not yet available.) With the adjustment recommended by the Panel, a scale score of 47 became a passing grade of 65. This adjustment holds this year's students to the same standard as their counterparts were held in June 2002 yet, even after this substantial adjustment, the early estimates were that 1/3 of the students still failed. At our September 19 meeting, we were provided with the final statewide results. They show that, after the rescaling we recommended:

- 45% of the State's children failed at the 65 level (33% at the 55 level);
- 59% of New York City's children failed at the 65 level (47% at the 55 level).

Unless these students pass a Math A exam in the future, they will not earn a high school diploma, which will render them ineligible for a wide range of jobs in our society, many of which do not require high level mathematics skills.

The Panel is also concerned that, even after the adjustment we recommended, an estimated 16% of the 9th graders who took the exam failed at the 65 level (11% at the 55 level). Ninth graders who take this exam are typically considered as strong math students. For one out of six of these students to fail a Regents examination required for graduation points to a problem that must be solved.

The weight of responsibility this Panel feels is about finding balance. On the one hand, our schools must ensure that our high school graduates have strong math skills; on the other hand, failing rates such as the ones we have seen with Math A are, we believe, unacceptable. The Panel has placed its primary focus on the standards, the assessment, and the infrastructure. For the sake of our children, we adults need to get this right.

II. Executive Summary

The Panel supports the Commissioner and the Board of Regents in the quest to raise standards for all children, and we write this report in the hope of recommending solutions to the problems the Panel has identified, so that our State may continue on its path of providing a top quality education for every child. Based on the Panel's perusal of math graduation exams from other states, it is the Panel's opinion that New York State has the highest math standards in the nation; our recommendations are intended to keep it that way.

Before Math A, there were two ways students could meet the math graduation requirement, either by passing the Course I Regents exam, or by passing the simpler Regents Competency Test (RCT). Math A is a much more challenging requirement than Course I; it tests more content and it has more problem solving. It is a challenge to move students from Course I to Math A. The challenge of moving students from the RCT to Math A is much greater. Early on, concerns were raised about the extent of this challenge. An SED report issued in 1998 entitled "Mathematics Standards and Assessment Review Committee Report" contains the following caution:

> Until (1) the Standards are clearly stated and communicated to teachers, students, parents and other interested parties; (2) proper support systems are put in place to give ALL students a fair chance to meet the Standards; and (3) proper support systems are put in place to provide ALL teachers with opportunities to develop mathematical content knowledge and pedagogical strategies, it may be unfair and unrealistic to expect the passing of the Mathematics A exam to be a requirement for a high school diploma. (p. 4.)

This Panel has concluded that the standards are not clear, and that the necessary support systems for students and teachers are not in place. It is not within this Panel's charge to discuss graduation requirements; the Panel's work has focused on changes to Math A to make it more successful.

As noted above, even after a substantial adjustment recommended by this Panel, 45% of the students failed the June 2003 Math A exam at the 65 level. Such a failure rate on an adjusted exam points not to a single problem or a few simple problems; it points to a systemic problem. This Panel focused on identifying the various facets of this systemic problem, and on making broad recommendations to the Board of Regents, so these problems can be addressed, and so we can continue moving forward to raise all of our students to high levels of math competency.

The Panel spent hours examining pages and pages of information, graciously provided by SED staff. The Panel spent hours speaking with SED staff, who patiently put up with our questions day after day. The Panel spent hours discussing what we were seeing, and then trying to reach consensus on our recommendations. The Panel believes it has

identified a number of areas in which changes need to be made, so that we can continue moving forward on the path to higher standards for all children.

The Panel has identified 16 findings, and has developed a total of 22 recommendations, many with several parts, for a total of 41 recommendations. These are summarized below.

The Math A Standards

Finding 1: *The Math A standards lack clarity and specificity* (p. 15).

Recommendation 1A: *Educationally useful standards must be developed in mathematics for each grade, K-8, and for Math A and Math B, that consist of a clear, well-defined set of skills, the mastery of which is demonstrable* (p. 19).

Recommendation 1B: *SED should establish a mathematics standards committee to rewrite the standards into functional form, and to meet regularly in the future to analyze test results, thus ensuring continuous relevance* (p. 19).

Recommendation 1C: *SED should develop and disseminate suggested curricula for mathematics instruction for each grade K-8, and for Math A and Math B* (p. 19).

Recommendation 1D: *To benefit from the extensive research and deliberation of the current Math A Panel, some of the current Panel members should be included in both new committees recommended in this report, i.e., the standards committee, and the curriculum development committee* (p. 20).

Finding 2: *The design concept that the Math A exam should be taken by the typical student after three semesters of instruction has not been successful* (p. 20).

Recommendation 2: *The standards and curricula should be structured so that the typical student will take the Math A exam after one year of high school mathematics* (p. 20).

The Math A Exam

Finding 3: *The June 2003 Regents Math A exam was harder than past Math A exams* (p. 25).

Recommendation 3A: *Establish and maintain narrow statistical targets for difficulty of Parts I, II, III, and IV of the Math A exam forms* (p. 28).

Recommendation 3B: *Review the Math A item pool* (p. 28).

Recommendation 3C: *The difficulty of problems in the anchor item set, in the guidance documents provided to teachers, and on the actual tests should be aligned* (p. 28).

Recommendation 3D: *The weighting of the open-ended items, number of scale points possible on the open-ended item rubrics, and other aspects of the scoring of open-ended items should be reconsidered* (p. 29).

Recommendation 3E: *The Math A test should focus on a more limited, more clearly-specified set of content standards and indicators* (p. 29).

Finding 4: *The Math A tests have not been able to maintain a consistent performance standard over time* (p. 29).

Recommendation 4A: *Alternative equating designs should be considered* (p. 32).

Recommendation 4B: *Sampling procedures for estimating item performance must be improved* (p. 32).

Recommendation 4C: *Replace the anchor item set* (p. 33).

Recommendation 4D: *Revisit performance standards (cut scores)* (p. 33).

Finding 5: *The New York State Education Department cannot accurately predict performance on Math A test* (p. 33).

Recommendation 5A: *SED should implement procedures for predicting the performance of test forms and groups of students on future Math A exams* (p. 34).

Recommendation 5B: *Policies for field testing and data collection should be revised* (p. 34).

Finding 6: *Support and oversight for the Math A exam program should be improved* (p. 34).

Recommendation 6A: *SED should immediately increase in-house content and technical expertise resources by a minimum of one psychometrician and two math content specialists* (p. 35).

Recommendation 6B: *SED should clarify the responsibilities assigned to its technical advisory committee, and should request this group to provide regular*

reports, including technical analyses, reactions to proposed changes in test programs, and suggestions for improving State testing programs (p. 35).

Recommendation 6C: *SED should increase demands placed on contractors* (p. 36).

Recommendation 6D: *Internal coordination and documentation should be improved* (p. 36).

Infrastructure Issues Related to the Attainment of Math A Standards

Finding 7: *Passing rate data for the State as a whole were not available until three months after the exam; no data are collected regarding student performance on individual items, nor even regarding student performance on the four parts of the exam* (p. 37).

Recommendation 7: *SED should increase its data collection capacity to include item level data, and should accelerate its data collection timetable* (p. 37).

Finding 8: While *the most important use of student performance data is to inform instruction, statewide data mining models that would enable local schools and teachers to use these data effectively are not generally available* (p. 37).

Recommendation 8: *SED should substantially broaden its efforts to assist districts in data collection, and the use of data to inform instruction* (p. 37).

Finding 9: *The mathematical background of teachers delivering math instruction varies widely; yet, raising almost three million children to higher levels of math achievement will be impossible without highly skilled teachers* (p. 37).

Recommendation 9A: *SED and higher education need to continue and to strengthen their partnerships to ensure strong teacher education programs, both pre-service and in-service* (p. 37).

Recommendation 9B: *The certification requirements for elementary teachers and special education teachers should include a minimum of nine credits of college level mathematics (see Recommendation 9C), and three credits of teaching techniques in mathematics* (p. 37).

Recommendation 9C: *Mathematics courses required for certification, both for mathematics teachers and elementary and special education teachers, should be specific not only in terms of number of credits required to be taken, but also in*

terms of coursework required to be taken, e.g., calculus, number theory, algebraic structures, probability and statistics, etc. (p. 38).

Recommendation 9D: *The Panel believes that, for any teacher responsible for teaching mathematics at any level, the 175-hour professional development requirement should include specific mathematics requirements. The Panel's thinking is that:*

- *teachers who teach mathematics exclusively should be required to take 100 of the 175 hours in the area of mathematics;*
- *secondary teachers who are certified in, and who teach in, more than one subject area, should be required to take 50 of the 175 hours in the area of mathematics;*
- *teachers who teach mathematics as part of a broad set of teaching responsibilities, e.g., elementary teachers and special education teachers, should be required to take 30 of the 175 hours in the area of mathematics.*

Additionally, the range of possible courses that would satisfy these requirements should be clearly specified (p. 38).

Finding 10. *The public has very little awareness of Math A, and may have misunderstandings about the goals of Math A* (p. 38).

Recommendation 10: *Make greater use of SED communications capacity to engage the public in conversations about the importance of strong mathematics skills* (p. 38).

Finding 11: *There is often a "disconnect" between K-12 and higher education* (p. 38).

Recommendation 11: *SED should encourage conversations at the local and regional levels of K-12 teachers of mathematics and higher education professors of mathematics, for the purpose of sharing curriculum, and exploring professional development opportunities and other possible collaborations, to bridge the gap between K-12 and higher education* (p. 38).

Finding 12: *Raising the level of mathematics achievement of all students to high levels must start when children are very young, and must go beyond the school day for school aged children* (p. 39).

Recommendation 12: *SED should encourage through grants and other means the expansion of mathematics education initiatives beyond K-12, such as the creation of partnerships between schools and libraries, and the greater use of public television and museums* (p. 39).

Additional Issues --
Scoring Rubrics, and Communication to the Field Regarding Grading

Finding 13: *The scoring rubrics do not give credit for a variety of mathematically correct approaches* (p. 40).

Recommendation 13A: *Develop more generally worded, holistic scoring rubrics which permit credit to be granted for atypical, but mathematically correct, student responses* (p. 40).

Recommendation 13B: *Rubrics should be designed so students do not lose 33% or 50% credit for a minor arithmetic error* (p. 40).

Finding 14: *There is a serious "disconnect" between the perception of the SED content specialists and the perception of field classroom teachers regarding the application of the scoring rubrics* (p. 40).

Recommendation 14: *On each set of directions for the Math A exam, a statement should be added confirming that the scoring rubrics are a guide and should be applied using professional judgment* (p. 40).

Finding 15: *There needs to be better communication of SED grading interpretations during the grading process for the Math A exams* (p. 41).

Recommendation 15A: *SED should continue on its path of setting up a website during Math A Regents exam grading to provide up-to-date clarifications to teachers grading the exam* (p. 41).

Recommendation 15B: *SED should explore ways of sending up-to-date grading clarifications to the school districts during the grading period following the administration of the exam, as a backup to the website, to ensure the greatest possible consistency of grading across the State* (p. 41).

Additional Issues --
Calculator Use on the Math A Exam

Finding 16: *Allowing the option of using a graphing calculator on the Math A exam provides some students with an advantage on the exam, thus creating an inequitable situation* (p. 41).

Recommendation 16: *The use of calculators on the Math A Regents exam should be standardized* (p. 42).

The January 2004 Exam, and All Math A Exams until A New One Is Designed

Recommendation 17: *Until the standards are rewritten, new curricula are developed, the new course is delivered, and a new Math A Regents is designed and field tested, the Math A Regents exam should be restructured so the exam includes: 30 Part I items, 5 Part II items, 2 Part III items, and 2 Part IV items* (p. 43).

Recommendation 18: *The exam should be reviewed by a group of practitioners, including math teachers, university mathematicians and mathematics educators, with representatives from this Panel, prior to the administration of the exam* (p. 43).

Recommendation 19: *Until new items are developed and properly field tested, the exam items should be scaled in accord with the procedures used for the August rescaling of the June 2003 exam* (p. 43).

Recommendation 20: *The scaling should not be finalized until after the exam has been administered and after a post equating procedure has been implemented to ensure the fairness of the test* (p. 43).

Recommendation 21: *The 55 passing option on the Math A Regents Exam for a local diploma should be continued until after the standards have been clarified, after new curriculum has been developed and disseminated, and after a new exam has been developed and administered for at least one school year (to ensure that it is performing in accord with its design)* (p. 44).

Recommendation 22: *The math RCT safety net for special education children should be continued until after the standards have been clarified, after new curriculum has been developed and disseminated, and after a new exam has been developed and administered for at least one school year (to ensure that it is performing in accord with its design)* (p. 44).

The Panel believes our recommendations, taken together, will successfully address the problems we have identified in our independent investigation. A suggested timeline for implementation has been developed and is included in the report. (p. 45).

III. The History of Math A

During the 1990s, discussion ensued about raising the standards for mathematics education in New York State. Ultimately, a decision was made to phase out Course I, Course II, Course III and to replace this three-year sequence with Math A and Math B. Conceptually, Math A was to include topics from about a year and a half of the Course I, II, III sequence; and Math B the remainder. A major shift in emphasis was toward more contextual problems and with a greater emphasis on *genuine* problem solving, i.e., mathematics within a context, where problem-solving strategies can be used. While there was to be a Math A exam and a Math B exam, there was not a curriculum developed. Rather, schools were informed of the math standards, expressed in seven "Key Ideas" which, in turn, were subdivided into 103 "Performance Indicators." Schools were told that they could reach these standards in whatever way they wished but were advised that students would be assessed on these 103 Performance Indicators. Over time, schools worked to develop courses to meet the new standards.

The first Math A exam was administered in June 1999. For several years, SED produced both the old and new exams, and schools could offer either one. The last Course I Regents was administered in January, 2002. It is no longer an option.

During this same time period, the Board of Regents made a series of policy decisions that resulted in high school graduation becoming contingent upon the passing of five Regents exams, with a math exam being one of those exams. Now, with Course I no longer available, the exam required for graduation has become Math A. Prior to this policy change, students could graduate with different types of diplomas. Some students met the requirement by passing the Math Regents Competency Test (RCT), a fairly basic test of skills, whereas others met the requirement by passing the Course I Regents exam.

From the beginning, all knew Math A was a substantial change, more for some students than for others, but a change for all. In 1998, a group of math experts expressed concerns about the difficulty level of Math A. As the exams were phased in, concerns from the field grew about the difficulty level and the wording of problems. When the June 2003 Math A exam was administered, the concerns became an outcry. Teachers saw that the test was very difficult. Early anecdotal evidence from the schools pointed to a very high failure rate. SED responded by requesting data from schools. When the data confirmed a high failure rate, the Commissioner made the decision to set aside the test for current 11th and 12th graders, and to permit them to substitute their course grade for the purpose of the graduation requirement.

Shortly thereafter, this Math A Panel was created by the Board of Regents and the Commissioner, and asked to respond to a nine-element charge. (See Appendix A.)

The Panel dedicated three full days (and held extensive conversations between meetings) to the first part of its investigation, which was whether the June 2003 exam was more difficult than previous exams and, if so, what to recommend as a rescaling to

the Commissioner. The determination was made that the exam was, in fact, more difficult. In an Interim Report, the Panel recommended rescaling the June 2003 exam based on the June 2002 results, using 9th grade students as the basis, as the 9th grade groups in both years were similar. The Panel's estimate was that this adjustment would raise scores in the middle of the distribution about ten points. The Commissioner accepted the Panel's recommendation, and directed SED staff to implement the adjustment. Within days, SED generated a new scale for the June 2003 exam; it converted an old 47 to a new 65. According to an SED press release at the time, the estimates of the impact on passing rates were as follows:

9th graders:	from 61% to 80% passing
10th graders:	from 32% to 64% passing
11th graders:	from 28% to 60% passing
12th graders:	from 28% to 55% passing

The Panel then continued with its work on the remainder of the elements of the charge. (See Appendix A.) This document is the Panel's final report to the Commissioner and Board of Regents.

IV. The Development of the Math A Exams

Each Math A Regents exam is the result of a multi-year cycle of test development, which results in four actual tests being created each year. Three of these tests are for the expected administrations, and one is held in reserve in case it is needed. (The June 2003 exam was the first exam used from a four test cycle. The January 2004 exam is scheduled to be the second of the four exams from the same test development cycle.)

The first Math A test development cycle occurred in 1997 and 1998; this cycle resulted in the setting of standards levels which are applied to this day.[1]

Each Math A exam has 35 items, 20 multiple choice and 15 open-ended questions, The test specification calls for point values as follows:

Part I: 2 points each for all 20 multiple choice items (totaling 40 points)
Part II: 2 points each for 5 of the open-ended items (totaling 10 points)
Part III: 3 points each for 5 of the open-ended items (totaling 15 points)
Part IV: 4 points each for 5 of the open-ended items (totaling 20 points)

Thus, the raw score point range is 0 to 85 points. This is scaled onto a traditional 0-100 range using the equating techniques summarized elsewhere in this report.

The test development process starts with teachers being invited to Albany to write multiple choice and open-ended items. Once the items have been written, SED staff and consultants then select items for pretesting. Math A pretesting involves the creation of between 20 and 30 forms, each of which consists of 5 or 6 multiple choice items and 3 or 4 open-ended items. Schools are sampled with the goal of pretesting each item on 250 representative students statewide.

When the pretest forms are returned, the items are graded, and the results are sent to an outside consultant for analysis. Part of this grading process is called "rangefinding." This process is an effort to assign point values to various levels of response to the open-ended items. It involves establishing rules for grading each open-ended item and it involves selecting student papers which are exemplars of each point value assignment. This is done by classroom teachers under the coordination of an outside contractor, the purpose being to create a guide for the grading process in local schools.

---

[1] There are many critically important facts about the early development of the Math A test that cannot be answered, because of staff turnover at SED and because some areas are hard to document. The Panel does not know how the initial set of problems used to set original bookmarks were developed, e.g., what were the directions to the item writers, what were the backgrounds of those writers, and who were the students whose field test efforts on these problems were used by the benchmarking committee to assess the difficulty of these problems? Likewise, the backgrounds of the members of the benchmarking committees are not known. The panelists worry that the item writers, the field test students, and the benchmarking committee members may not have been properly representative of their counterparts statewide.

In order to set the standard for passing and passing with distinction, during the first cycle of test development (1997-1998), a "bookmarking" standard-setting process occurred. This involves taking the items, after their relative difficulty has been determined, and arraying them from easiest to hardest. Then, a large group of math teachers convenes and holds several discussions for the purpose of determining where the "cut points" should be for passing (65) and passing with distinction (85). Once these standards/cut points are set, they are used for all future administrations of the test, until a new standard setting process occurs. All Math A exams have used the same cut points, through and including the June 2003 exam.

Also for the first cycle only, items that seem strong, both in terms of measuring the content and in terms of their item statistics, are selected as "anchor items."[2] These anchor items are used as the basis for equating all future exams. (For Math A, the original set of anchor items included 35 items. At some point, these were pared down to 18 items. While SED staff members cannot recall the rationale for doing this, most speculate that it was to shorten the test so it could be administered in one class period.) The set of 18 anchor items has been used for several test administrations in a row, up to and including the present. The June 2003 exam was equated based on these items, as have been the remaining three exams, including the one scheduled to be administered in January 2004.

Once the pretest results are obtained, four field test forms are prepared. These forms are intended to be pretty close to the actual exams that will be given. Each has 35 items and looks like a Math A Regents exam. The items are selected from the pretested items by SED staff and outside consultants. The selection is based on content coverage and item statistics from the pretests. (About half the pretested items survive to the field test level.)[3]

Because SED's experience is that schools are more accepting of field tests which last no more than one class period, each of the four full field test forms is divided into three field test "mini-forms," yielding a total of 12 mini-forms. A representative sample of schools is chosen and asked to administer the field test. When the forms are sent to the schools, a 13th form is sent, a form with the 18 anchor items mentioned above. (There has been a mixed practice over the years. In some years, the 18 items were interspersed with the field test items on the same form; in other years, the anchor form was a separate form and was "spiraled," which means it was given to randomly selected

[2] Please refer to the section on the Math A exam for definitions and additional discussion of technical issues.

[3] During its work, the Panel learned that, because of the complexity of developing item statistics for open-ended questions, these statistics are not available when the field test forms are created. The selection of the open-ended items for the field test form is based on content coverage and an estimate of difficulty. Although items can be replaced after field testing, the Panel believes that pretest item statistics for open-ended items should be available before the items are selected for the field tests.

students within the same group as the field test forms. The latter practice is the more recent one, and it was the one used for the June, 2003 exam.)[4]

Once the field tests are returned and graded, again using rangefinding for the open-ended items, the results are sent to a consultant for item analysis, and the results are reviewed by math teachers and SED staff. This can result in items being modified or replaced by SED staff or outside consultants -- without input from field mathematics specialists. Four items on the June 2003 Math A exam were replacement items from the item pool: Items numbered 14, 15, 30 and 35.[5]

Once the four forms of the exam are finalized, the equating process is applied by the consultants[6] and the 0-85 raw score scale is transformed to the traditional 0-100 scale for each exam.[7]

---

[4] Although two different techniques have been used to administer the anchor items to the students, it appears to the Panel that both methods are acceptable and should yield similar results. The difference is noted in this report only for the sake of clarity.

[5] The Panel notes that for the June 2003 Math A exam, items were replaced, and the final form was not reviewed by field math teachers. The Panel would have recommended such a step in the development process, but has been advised that this step has already been added, beginning with the August 2003 Regents exams. The Panel applauds SED for this additional step.

[6] The Panel was provided with a very professional appearing "Equating and Scaling" report from the 2000 field test used to develop the exams administered in 2001. The Panel had requested the report used to equate the June 2003 exam. At its September 10 meeting, the Panel was provided with a draft report dated June 2003, and realized this was a report on the 2001 field test used to develop the exams administered in 2002, not a report on the 2002 field test used to develop the 2003 exams. The Panel has been advised by SED staff that the consultant did provide all of the item analyses required, as well as the scaling tables, but has not yet submitted a formal report. While this does not appear to have a material impact on any of the results, the Panel believes that these equating and scaling reports should be in the hands of SED staff several months before the exams to which they pertain are used to rate schools and students. This way, if the consultant sees a problem with any of the items, there is time for adjustment. To the Panel, it appears that the 2002 field test Equating and Scaling report is now late by over a year.

[7] Decisions regarding the final test depend upon the item statistics provided by the consultant. The item statistics the Panel received for the field test which led to the June 2003 exam (and three exams to be given in the future, including January 2004) had four different sets of item statistics, three of which were crossed out. SED staff had been told by the consultant to use the one set of statistics not crossed out and to ignore the others. The Panel attempted to ascertain if there was any importance to the three sets of statistics crossed out and SED staff arranged for a telephone conversation with Panel members, SED staff and representatives of the consulting company. The consulting company was not immediately able to explain what had occurred, nor why it had occurred. This has led to some of our recommendations regarding the technical aspects of test development.

V. Findings and Recommendations

A. The Math A Standards

The Panel's work started with various analyses of the Math A exam. These analyses, which are presented later in this report, led the Panel to its findings regarding the standards presented here. Although the analyses of the exam came first in our work, we present our findings on standards first, as the standards form the foundation for the exam.

Finding 1: *The Math A standards lack clarity and specificity*.

Classroom teachers, parents and students do not know what Math A is. The Panel already mentioned in its Interim Report the failure of the June 2003 examination to cover trigonometry. Teachers were led to expect trigonometry; the Mathematics Resource Guide with Core Curriculum states:

> *Students are still expected to master basic skills of arithmetic, geometry, algebra, trigonometry, probability, and statistics. The State Education Department will continue to assess these skills and concepts with tests that will be given in secure settings, and the results of these tests will be made public each year (p. 3.).*

Teachers, trying to prepare their students for the June 2003 exam, read these words, looked at previous exams, decided based upon this guidance that trigonometry needed to be taught, and spent weeks helping their students learn this area of mathematics. As noted in the Math A Panel Interim Report, trigonometry was not assessed by even one item on the June 2003 Math A Regents. Classroom teachers have come to believe, with good reason, that they can only guess which topics are important, and hope they make the right guess as they teach their students.

Not only are the topics unclear, but the breadth and depth of the expected understanding are unclear. The standards as they are currently written do not easily translate into classroom practice, and they are confusing to teachers, students, and parents.

As just one illustration of this, we point to one of the 103 Performance Indicators, Performance Indicator 5A. This indicator states several expectations, including the Pythagorean Theorem, but the depth of expected knowledge is not specified. Is the graduation performance standard a straightforward numeric substitution using the theorem (which would be a minimal expectation) or is it a deep conceptual understanding of the theorem and its applications (higher mastery)? Is the expectation a simple statement of the theorem, or application to a right triangle, or application twice in the same problem, using the theorem algebraically and proving a right triangle? What is the expectation? It is not clear.

To elaborate, the statements of the Key Idea and Performance Indicator which include the Pythagorean Theorem are as follows:

> **Key Idea 5: Measurement**
> *Students use measurement in both metric and English to provide a major link between the abstractions of mathematics and the real world in order to describe and compare objects and data.*
>
> **Performance Indicator 5A**:
> *Apply formulas to find measures such as length, area, volume, weight, time, and angle in real-world contexts.*
> Includes:
> - *Perimeter of polygons and circumference of circles.*
> - *Area of polygons and circles.*
> - *Volume of solids.*
> - *Pythagorean Theorem*

On the next page is a table showing how this one Performance Indicator has been tested over the years, and also showing the Assessment Example provided in the guidance document provided to teachers.

Assessment Example for Performance Indicator 5A, from the Mathematics Resource Guide with Core Curriculum	**Math A Regents Exam Questions Mapped by SED to Performance Indicator 5A**			
	June 2002 Exam	**August 2002 Exam**	**January 2003 Exam**	**June 2003 Exam**
Ms. Brown plans to carpet part of her living room. The living room floor is a square 20 feet by 20 feet. She wants to carpet a quarter-circle as shown below. Find to the nearest square foot, what part of the floor will remain uncarpeted. Show how you arrived at your answer.	31. As seen in the accompanying diagram, a person can travel from New York City to Buffalo by going north 170 miles to Albany and then west 280 miles to Buffalo. *a* If an engineer wants to design a highway to connect New York City directly to Buffalo, at what angle, x, would she need to build the highway? Find the angle to the *nearest degree*. *b* To the *nearest mile*, how many miles would be saved by traveling directly from New York City to Buffalo rather than by traveling first to Albany and then to Buffalo?	31. In the accompanying diagram, x represents the length of a ladder that is leaning against a wall of a building, and y represents the distance from the foot of the ladder to the base of the wall. The ladder makes a 60° angle with the ground and reaches a point on the wall 17 feet above the ground. Find the number of feet in x and y.	30. A rectangular garden is going to be planted in a person's rectangular backyard, as shown in the accompanying diagram. Some dimensions of the backyard and the width of the garden are given. Find the area of the garden to the *nearest square foot.*	30. To get from his high school to his home, Jamal travels 5.0 miles east and then 4.0 miles north. When Sheila goes to her home from the same high school, she travels 8.0 miles east and 2.0 miles south. What is the measure of the shortest distance, to the nearest tenth of a mile, between Jamal's home and Sheila's home? [The use of the accompanying grid is optional.]
		35. Determine the distance between point A(–1,–3) and point B(5,5). Write an equation of the perpendicular bisector of AB. [The use of the accompanying grid is optional.]		34.A straw is placed into a rectangular box that is 3 inches by 4 inches by 8 inches, as shown in the accompanying diagram. If the straw fits exactly into the box diagonally from the bottom left front corner to the top right back corner, how long is the straw, to the nearest tenth of an inch?
				2. The accompanying diagram shows a square with side y inside a square with side x. Which expression represents the area of the shaded region? (1) x^2 (3) $y^2 - x^2$ (2) y^2 (4) $x^2 - y^2$

These are all very different problems. Which one is the standard?

The lack of clarity can also be seen in the overlap of Key Ideas and Performance Indicators. For example, Question 30 of the June 2003 Math A Regents (in the table on the previous page), which was mapped by SED to Key Idea 5, Performance Indicator 5A, could also be mapped to Key Idea 5, Performance Indicator 5G. Note, however, the Assessment Example given to teachers to help them understand the standard:

Performance Indicator 5G	**Assessment Example 5G from the Mathematics Resource Guide with Core Curriculum:**
Relate absolute value, distance between two points, and the slope of a line to the coordinate plane. Includes: • *Absolute value and length of a line segment.* • *Midpoint of a segment.* • *Equation of a line: point-slope and slope intercept form.* • *Comparison of parallel and perpendicular lines*.	What is the distance between points A (7,3) and B (5,-1)?

Clearly, Assessment Example 5G is much more direct -- and much easier -- than Question 30 on the June 2003 Regents exam. The Panel has found such "disconnects" repeatedly between the types of items provided to teachers as examples, and the types of items appearing on the actual exams. What is the standard?

Yet another example of this disconnect can be found in the contrast between Question 29 on the June 2003 exam, and the example given in the SED teacher guidance document, both shown below. Notice how much more complex the test item is, when compared with the example provided to teachers as guidance.

<table>
<tr><td colspan="2">Key Idea 6 Uncertainty:
Students use ideas of uncertainty to illustrate that mathematics involves more than exactness when dealing with everyday situations.
Performance Indicator 6C:
Use the concept of random variable in computing probabilities.
Includes:
• Mutually exclusive and independent events.
• Counting principle.
• Sample space.
• Probability distribution.
• Probability of the complement of an event.</td></tr>
<tr><td>Assessment Example 6C from the Mathematics Resource Guide with Core Curriculum</td><td>June 2003 Math A Question 29 mapped to Performance Indicator 6C:</td></tr>
<tr><td>The graph below shows the hair colors of all the students in a class. What is the probability that a student chosen at random from this class has black hair?</td><td>29. A certain state is considering changing the arrangement of letters and numbers on its license plates. The two options the state is considering are:
Option 1: three letters followed by a four-digit number with repetition of both letters and digits allowed
Option 2: four letters followed by a three-digit number without repetition of either letters or digits
[Zero may be chosen as the first digit of the number in either option.]
Which option will enable the state to issue more license plates? How many more different license plates will that option yield?</td></tr>
</table>

The Assessment Example and the test item represent very different expectations; which one is the standard?

In some ways, this situation can be likened to setting a standard that all students should run fast. We all have a sense of what this means, but in a high stakes environment, clarity and specificity are essential. Is "run fast" defined as a nine minute mile or a seven minute mile? The Panel is convinced that the lack of clarity and specificity of the standards must be addressed, as the standards are the foundation for all other aspects of this work.

As the Panel reviewed student performance data from the Math A testing program, it became very clear that we cannot limit our thinking to the high school math program. SED data show a very high correlation between not passing the 8th grade math assessment and not passing the Math A Regents examination. Almost one-half of all 8th graders statewide scored a level 1 or level 2 on the Math 8 exam; how can this apparently significant deficit be made up in a matter of months at the high school? Because mathematics is such a sequential subject, any effort to modify Math A must include efforts directed at the lower grades, K-8. Additionally, because Math A leads to Math B, any efforts regarding Math A must be extended to the upper grades. The effort to streamline and clarify the standards must extend to the other grades.

The Panel's recommendations concerning this finding are:

Recommendation 1A: *Educationally useful standards must be developed in mathematics for each grade, K-8, and for Math A and Math B, that consist of a clear, well-defined set of skills, the mastery of which is demonstrable.*

Recommendation 1B: *SED should establish a mathematics standards committee to rewrite the standards into functional form, and to meet regularly in the future to analyze test results, thus ensuring continuous relevance.*

This committee should include a large cross section of adults including mathematics teachers, university mathematicians, professors of mathematics education, special education teachers, parents, and adults who work with mathematics in real work applications, both in the professions (for example, engineers and accountants) and in the trades (for example, carpenters and electricians). The Panel envisions that this group would meet at least once a year to review the exams against the standards, in order to provide continuity over time.

Recommendation 1C: *SED should develop and disseminate suggested curricula for mathematics instruction for each grade K-8, and for Math A and Math B.*

The Panel wishes to make it clear that it does not recommend this as a mandated or required curriculum, but rather as additional guidance to the field. No curriculum, no matter how strong, can take the place of a gifted classroom teacher. The Panel wishes not to discourage in any way individual creativity either at the classroom level or the

district level, but, rather, to provide struggling teachers and schools with a suggested starting point upon which they may build. The Panel also wishes to state that we see this as a need because of the highly sequential structure of mathematics; this idea does not necessarily transfer to other subject areas. The Panel envisions a curriculum development committee of mathematics teachers, and representatives of the mathematics standards committee.

Recommendation 1D: *To benefit from the extensive research and deliberation of the current Math A Panel, some of the current Panel members should be included in both new committees recommended in this report, i.e., the standards committee, and the curriculum development committee.*

Finding 2: *The design concept that the Math A exam should be taken by the typical student after three semesters of instruction has not been successful.*

The Panel understands the thinking behind the original design concept that the Math A exam should be given to the typical student after a year and a half of coursework. However, the "disconnect" between this model and the academic year has been problematic. First, we are all aware of the research demonstrating the "drop" which students experience during the summer, especially weaker students; this impacts most those students who are struggling with Math A, and it leaves the teacher of the third semester before the Math A exam being responsible for closing that gap. Another issue is the rhythm of the school year, a force which cannot be ignored. Students often experience one teacher in the first and second semesters of instruction leading to the Math A exam, and another teacher for the third semester. It would seem logical that teachers should be scheduled so that they remain with one group of students for all three semesters. However, with course sign-ups, singleton courses, etc., this becomes a very difficult goal to meet. Additionally, teachers who are leaving or retiring are encouraged to do so in June, so as not to disrupt student instruction; and new teachers are hired as of September 1, for the same reason. Yet, with Math A typically being a year and a half course, these very efforts to limit disruption of the continuity of instruction, actually *cause* disruption.

Additionally, those students who complete Math A in a year and a half, and who choose not to enroll in Math B, need to take an additional year and a half of mathematics. Schools are left with the problem of "inventing" half year options to help these students continue their education. The Panel believes these students would be better served by taking two full year courses after passing the Math A exam.

It is also noted below in the section of this report on the Math A Exam that the current configuration is creating a content validity issue for the exams.

Recommendation 2: *The standards and curricula should be structured so that the typical student will take the Math A exam after one year of high school mathematics.*

The Panel believes that SED, working with the curriculum committee mentioned above, should redesign Math A into a *one-year course*, by realigning topics in K-8, by streamlining topics, and by providing a specific scope and sequence. (The Panel wishes to make it clear that it does not see one year as a mandated or required length of the course. Local districts should have the option of providing alternative time frames for course completion to tailor the course to the needs of the student population.)

B. Findings and Recommendations Concerning the Math A Exam

Introduction

Math A was designed to raise the standards of mathematical knowledge and problem-solving ability of New York high school graduates. The Panel supports the efforts to provide access to high level programs for all children, and the efforts to raise math skills across the State. However, the Panel's work has led it to the conclusion that these new standards were not well-defined by clearly-specified objectives, an adequately structured curriculum, specific courses, or sufficient professional development. Rather, it was required of teachers, students, and others to make strong inferences about Math A based largely upon its operationalization in the form of the Math A examinations.

The introduction of the Math A test with its higher standards presented an array of challenges to the New York State Education Department (SED) staff. These challenges would be daunting under ordinary circumstances, but the difficulties were, we believe, compounded by staffing inadequacies, and technical constraints imposed by New York's Truth in Testing law.

The Commissioner's Panel investigating the June 2003 results was charged with responding to nine elements of the Commissioner's charge. One subcommittee of the Panel focused more squarely on technical issues. For example, we looked at whether Math A exams in general (including the June 2003 exam) have been designed and implemented in compliance with appropriate professional test standards. We found no material problems in this area. (See Section 1, below.)

We then proceeded to investigate technical concerns specific to the June 2003 Math A test. We investigated issues related to item writing, test development, equating, technical analysis, and oversight of contracted services related to the Math A examinations. We evaluated the infrastructure that supports the conceptualization, development, administration, and reporting of test results. It was in these areas that the Panel found serious inadequacies.

In the following sections, we first briefly address compliance with relevant professional standards. We then address the technical issue of comparability of examinee groups. Finally, we turn to problems we identified with the Math A assessment and the infrastructure supporting it. For each problem, we provide a summary of the evidence that led us to conclude a problem existed, followed by one or more recommendations for addressing the problem.

1. Compliance with Appropriate Professional Standards

Element number 1 of our charge from the Commissioner of Education was to answer the following question:

> Did the June 2003 Regents Math A exam measure achievement of the New York State mathematics standard three as defined through the core curriculum--consistent with generally accepted standards for assessment? (Refer to the so-called "Joint Standards.")

Many relevant professional standards exist, including the *Code of Fair Testing Practices in Education*, the *Code of Professional Responsibilities in Educational Measurement,* and others. The reference in the Commissioner's charge to the "Joint Standards" is a reference to the single most authoritative source of guidelines for appropriate practice in educational testing. That document, which bears the formal title, *Standards for Educational and Psychological Testing,* is the result of the joint efforts of the three leading organizations representing expertise in educational measurement. Those organizations are the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The *Standards* are comprehensive in that they contain chapters on every relevant topic that could be addressed when evaluating the quality of a test (e.g., validity, reliability, bias, and so on). The *Standards* themselves are updated on a recurring basis. The latest edition of the *Standards* was published in 1999. (One author of this report served on a review committee during the development of the 1999 standards and currently serves on the Joint Committee on Testing Practices of NCME.)

A detailed comparison of the degree to which the Math A examinations are developed, administered, and reported in accordance with the *Standards* was beyond the scope of our time frame and resources.[8] Instead, we reviewed the Math A assessment program to identify potential, serious violations of the *Standards* on the most important aspects they were intended to inform. For example, the *Standards* require that testing programs provide evidence of the validity of examinations. One way to satisfy this requirement would be to demonstrate that each item on a test was linked to an established content standard. On this count, we were presented with evidence that each item on the Math A test was written by teachers with backgrounds in mathematics teaching, curriculum, and by others with editorial skills. Each item in the June 2003 Math A test was reviewed, among other reasons, to ensure alignment with the Math A content standards.

We were also presented with evidence that the June 2003 Math A test comported with appropriate professional standards in other areas. For example, it met requirements for gathering and reporting reliability evidence; statistical and judgmental procedures were in place to screen items for potential differential functioning in various groups (i.e., to ensure items that are equally fair to various sex or ethnic groups); contemporary, accepted procedures were followed for establishing the performance standards (i.e., the "cut scores") defining the levels of performance on the test; careful sampling plans for field testing were provided, and so on.

[8] There are 264 individual standards listed in the document. Of those, not all would apply to K-12 testing programs in education, such as the Math A program. Nonetheless, a detailed evaluation of even the relevant standards was not possible.

On the other hand, time permitting, the Panel would have wished to pursue compliance with each, specific, relevant *Standard* in greater depth. For example, we would have liked to more thoroughly investigate the specific qualifications of item writers. We would have liked to obtain information on the specific characteristics of those who set the performance standards (i.e., the cut scores) for the Math A examination in the first place. We would have liked to delve deeper into the Math A item pool to investigate to what extent item classifications are clear and unambiguous, the extent to which the pool has sufficient content-valid items to support the development of truly equivalent test forms, and so on.[9]

Admittedly, in the Panel's prioritization of elements in the charge, this element did not receive our focused attention until other more time-critical elements had been addressed. It is our understanding that SED has documentation on many aspects of alignment with the *Standards for Educational and Psychological Testing,* although we have not yet completed a review of that documentation.

It is our opinion that detailed scrutiny against all of the relevant standards would likely reveal areas for which improvements could be made. This speculation notwithstanding, it is our opinion that the June 2003 examination was developed, administered, and scored essentially in compliance with the applicable specific guidelines of the *Standards* as well as within the spirit of that document.

2. Groups of Students Taking the Test

The Panel was asked to answer the following question, which is the third of the elements presented to us by the Commissioner:

> Were groups of students taking the June 2003 Math A exam statistically similar to or different from those taking previous Regents Math A exams?

This charge was difficult for the Panel to address and our conclusions are not founded on definitive data. The Math A testing program is precluded from collecting comprehensive, timely data on test takers. And, as we will see in the next section of this report, the statistical procedures which might aid in providing key policy and evaluative information also suffered from inaccuracies which result in part from State requirements that constrain appropriate test design.

Overall, our evidence on the question of group similarity is somewhat mixed. On the one hand, as we noted in our Interim Report presented to the Commissioner on August 25, 2003, there are anecdotal reports to support the conclusion that "there are some differences between the June, 2002 population and the June, 2003 population in that, this year, more students who are struggling in math took the Math A exam because the

[9] It is our understanding that SED has documentation that demonstrates the level of compliance with each of the relevant standards, produced previously for another purpose. Further, we understand that SED is gathering information regarding current compliance for review by the Panel. This documentation, however, was not obtained in time for a complete analysis and evaluation by the Panel.

Course I exam is no longer an option." On the other hand, one segment of the population we examined -- 9th grade students -- has remained reasonably similar. As stated in our first report, this group "has included, and continues to include, only those students who are strong in math and who the teachers feel can challenge this exam at that early stage of their high school career." It was our comparison of this relatively more stable and homogeneous group's performance from a sample of 400 school districts that suggested some adjustment of the scaling for the June 2003 Math A examination was in order.

3. Problems with Math A Assessment and Infrastructure

The Panel was asked by the Commissioner to address other questions related to the technical aspects of the Math A assessment program. A subcommittee of the Panel focused intensively on the following questions:

> Element 2: Were there anomalies in the test preparation process that could account for real or perceived changes in the level of difficulty in the June 2003 Regents Math A exam in comparison with prior Math A exams? This includes but is not limited to item writing, pretesting and field testing (including adequacy of the samples), production scheduling, scaling, equating, final test assembly, and review of the completed exam.
>
> Element 4: Is the 2003 Regents Math A exam of the same level of difficulty as prior Regents Math A exams? (That is, in addition to the equating included in question 2, consider the content, cognitive demand, and perceived difficulty of the exam.)

Our observations, data, and technical and logical analyses cause us to conclude that the Math A test has gotten more difficult over time. It is clear that certain psychometric procedures were not working properly; that relevant field test populations and performance were unstable and poorly understood; and that SED lacked appropriate and sufficient infrastructure to forecast, prevent, or respond to these problems.

Finding 3: *The June 2003 Regents Math A exam was harder than past Math A exams.*

Mathematics teachers on the Panel were unanimous in their assessment that the June 2003 Math A test, particularly Parts III and IV, were harder for their students than the previous Math A tests they reviewed. In our Interim Report, we provided evidence that supported the ultimate decision to rescale the June 2003 Math A test.

Specific to the June 2003 exam is the finding that a statistical indicator of the difficulty of test items (called "b-parameters") was higher in Parts III and IV of the June 2003 June test than in the June 2002 test. This statistical observation is confirmed by content experts on the Panel who judged that the June 2003 items (particularly those in Parts III and IV) were substantially more cognitively and linguistically complex.

We also compared the statistical and/or judgmental difficulty of three groups of items: (i) the items appearing on the June 2003 Math A exam; (ii) sample items, intended to be representative of Math A item difficulty, presented in the Mathematics Resource Guide and the 1998 Math A Test Sampler; and (iii) the set of 18 anchor items, created in 1998, and used in every subsequent field test as the basis for calibrating Math A items and equating Math A tests.[10] Content experts and non-content experts on the Panel concluded that the differences in these groups of items were striking, with the sample items and anchor items being dramatically less linguistically and conceptually complex than the comparison items in the June 2003 exam. Because we did not go back to also examine pretest data for these items, we cannot say whether the increasing difficulty of test items is more due to a change attributable to the way items are created (i.e., to changes in item writing practices) or to the way tests are created (i.e., to changes in test construction practices). There are at least three hypotheses for why the items appearing on the June 2003 Math A exam were, in real terms, harder than items appearing on previous tests. First, the increased difficulty may be due to a systematic evolution of items in the Math A item pool. It is possible that more straightforward items were selected for use on earlier examinations and that those items that remained in the pool for inclusion on the June 2003 exam were those implicitly judged to be less than optimal. A second hypothesis is that item writers for the Math A exam (likely unknowingly) evolved in their item writing practices -- writing more straightforward, easier items in the beginning and crafting more complex items as they exhausted their initial ideas for items, or as they became gained more experience or a changing perception of the level of complexity intended to be tapped by the Math A assessments. Finally, it is possible that a preference for items of increasing complexity (again, likely unknowingly) affected the decisions of those assembling the June Math A test in their choice of items.

These hypotheses are, of course, hard to test. However, we believe that the increase in item difficulty can be traced, at least in part, to some chronological constraints of the Math A test development process. We note that items in the June 2003 test were created at least as far back as the fall of 2000 and assembled into test forms in fall 2001. At this time, Math A instruction was just starting and item writers would likely

---

[10] At this point, a few definitions may be helpful. *Anchor* items are test questions for which the difficulty level of the question is considered to be known or fixed, based on the performance on those items by a reference group. In this case, the difficulty of the anchor items was established by the performance on those items of the first group to take the new Math A test in 1998.

Once the difficulty levels of the anchor items are known, the anchor items are administered along with new/field test items in subsequent years. In each subsequent year, a comparison of student performance on "known" anchor items provides a basis for determining the difficulty level of new/field test items. This process is referred to as *calibrating* the new items.

Finally, a statistical method called *equating* is used to determine a level of overall performance on a subsequent test comprising new items such that the standard of performance required to pass is the same for the group taking the current form of Math A test as it was for groups that took previous forms of the Math A test.

have been aware of the fact that the first few Math A test administrations had very high pass rates. This situation could have created an expectation that more challenging items would be appropriate in the future when Math A instruction was more established.

Finally, we observed two inherent design problems that affect the difficulty of the test. The first involves an aspect of the Math A test itself and centers on the weights assigned to and rubric scale values possible on the *open-ended* (also called constructed-response items).[11] As we have noted, the difficulty of the open-ended items in Parts III and IV of the June 2003 exam was substantially greater than problems on previous forms. However, as these items become harder, the effect of the rubrics and scale values used to assign partial credit increases. The current test specifications mandate a small number of scale points possible for the open-ended items. If the total points possible on an open-ended item can only be obtained if a student's response is error-free, then even the most minor arithmetic error will lead to loss of 33% credit on a 3-point free response item and a loss of 50% credit on a 2-point free response item. This design problem does not explain the lower performance on the June 2003 Math A exam, but it can lead to unexpected fluctuations in mean performance from year to year.

The second design problem concerns the lack of close alignment of the instruction with the content assessed on any given Math A exam. This characteristic, which we believe is a design flaw in the assessment system, operates in the following manner. There are many indicators – 103, each with varying levels of difficulty -- that form the content standards to be taught for Math A. Many indicators encompass a variety of distinct problem types and can be tested in a variety of ways at widely variable levels of complexity. The Math A test, however, is constructed to consist of only 35 items. Our review indicates that it is not uncommon for some of these items to require mastery of the same indicator. It is the consensus of the mathematics educators on the Panel that it is impossible for teachers to cover all possible combinations of indicators, problem types, and levels of complexity in the preparation of students. Thus, a student's probability of success on the Math A exam is related, in part, to the relative emphasis his or her teachers place on each of the indicators. For example, if a teacher emphasizes trigonometry, but if no trigonometry indicators are represented on a particular Math A test (as was the case on the June 2003 test) the student's skills will be underestimated. Conversely, if a teacher emphasizes mastery of the Pythagorean theorem, and if that knowledge is represented on a particular Math A test (as it was in several items on the June 2003 test), the student's competence may be overestimated.

In situations such as presented by the Math A assessment system where there are a large number of indicators, it would be reasonable for teachers to look to sample items provided by the State for guidance. However, as previously noted, the sample items

[11] An *open-ended* item is one for which a student must generate his or her own response, such as an essay or showing the work to arrive at the solution for a problem. This item format differs from a *select-response* item (such as the multiple-choice format) where the student selects from a fixed set of provided choices. *Rubric* refers to the scoring key used to evaluate open-ended items. For both item formats, a completely incorrect response would ordinarily receive zero points. However, whereas multiple-choice items have a fixed point value for a correct response, open-ended items are usually evaluated such that a better response earns more points than a weaker response.

provided as exemplars seriously misrepresent the overall level of complexity and difficulty of items on the June 2003 exam. Teachers who used this resource as a basis for aligning instruction and adjusting their teaching what they perceived to be the level of challenge of the Math A exam would, through no fault of their own, have erred.

The problem created by content underrepresentation in the current Math A assessment system cannot be overstated. It is an obvious validity concern. Beyond that, the problem prevents the system from capitalizing on a known phenomenon in assessment: instructional alignment. Many states have implemented higher standards and required mastery of more rigorous content. As might be expected, when new standards are introduced, overall performance is often at lower-than-desirable levels. However, when the new content standards are clearly specified, when instruction can be focused on the content standards, when tests can be created that are more fully representative of and aligned to the content standards, fairly large increases in average student performance are routinely observed.

Recommendation 3A: *Establish and maintain narrow statistical targets for difficulty of Parts I, II, III, and IV of the Math A exam forms.*

The relative difficulty of the four parts of the Math A exam must be stabilized so that the parts are more homogenous, equivalent, and stable as possible. The means and range of item difficulties should be consistent across parts and across forms. Having these targets in place will not only result in the reality of statistical stability, but will also promote the perception of fairness that items on each section of the test are of approximately the same level of challenge.

Recommendation 3B: *Review the Math A item pool.*

There exists a pool of field tested items that are available for use on future Math A test forms. However, the extent to which these items vary in linguistic and conceptual complexity and indicator coverage is not known. Obviously, if it is decided that the current Math A content standards and indicators are to be revised, each item in the pool would need to be reviewed to determine whether it is well aligned to the new content specifications. However, even if no changes are made to the content standards and indicators comprising Math A, the entire pool of old items must be reviewed initially and periodically to determine if item writing practices are inducing drift in complexity or misalignment.

Recommendation 3C: *The difficulty of problems in the anchor item set, in the guidance documents provided to teachers, and on the actual tests should be aligned.*

We noted the serious mismatch in difficulty and complexity among three sets of items: the set of items used as anchors, the set of items provided to teachers as samples of the content and complexity of Math A tests, and the set of operational items appearing on the June 2003 Math A exam. We discuss later in this report the problem introduced by a mismatch between anchor and scored items; this problem is of a more technical

nature. However, differences in difficulty among *any* of the three sets of items are of obvious concern, particularly the extent to which misrepresentation of difficulty and scope of coverage in the sample item set can lead to misalignment of classroom instruction.

Recommendation 3D: *The weighting of the open-ended items, number of scale points possible on the open-ended item rubrics, and other aspects of the scoring of open-ended items should be reconsidered.*

Recommendation 3E: *The Math A test should focus on a more limited, more clearly-specified set of content standards and indicators.*

While the Panel strongly supports the higher standards envisioned by the Regents for Math A, we believe that the current configuration of Math A content standards and assessments jeopardizes the attainment of those higher standards. Observations and recommendations relative to the Math A curriculum presented elsewhere in this report support this conclusion. Recommendation 1B presented earlier in this report recommends that a new process be put in place to review and revise the current standards.

The configuration of Math A coursework is also relevant to the problem of content underrepresentation on Math A tests that results from the number of indicators that comprise the current framework. It is important to note that Math A was originally conceived of as a challenging three-semester course. For many students, however, it is taught as a four-semester course. Such a structure may have been thought to be necessary given the larger number of indicators to be addressed. However, it is not possible to adequately assess a large and representative enough sample of indicators in a three-hour, 35-item examination. There are, of course, two possible remedies. Doubling the number of items on the examination would more fully represent the content. However, we judged that a single mathematics examination requiring six hours of assessment to be unacceptable from many perspectives: public support, cost, student fatigue, and others. Thus, as recommended elsewhere in this report, we believe that consideration should be given to reducing the number of content standards and indicators, and structuring Math A as a two-semester course.

Finding 4: *The Math A tests have not been able to maintain a consistent performance standard over time.*

Equating is the process by which a standard of performance (i.e., the level of performance indicated by a cut score) is maintained over time. There are a variety of designs possible for implementing equating. The equating design used for the Math A test consists of including blocks of anchor items along with field test forms so that items in a field test form can be calibrated and the passing standard (i.e., cut score) used for subsequent operational test forms can be adjusted to ensure comparability with previous years' tests. The current equating design is perhaps the best procedure possible given the constraints imposed by New York law on item release and

constraints imposed by the current practice of voluntary participation in pretesting and field testing. However, the equating design used is also highly susceptible to the introduction of fairly large and consequential errors.

A first consequential weakness in the equating design is that the anchor items are administered along with the field test items under conditions that do not have sufficient controls in place to assure confidence in the resulting statistical properties of *either* set of items. For example, the test is given under what are termed "non-motivated" conditions. There are no consequences for students and no diagnostic information provided to teachers as a result of their students' participation in a Math A field test. It is well known that students do not put forth their best effort under non-motivated conditions. We noted that up to nearly 20% of students simply opted not to answer some of the multiple-choice items administered during field testing; the proportions of students not putting forth their best -- or even typical -- effort was routinely even worse on more complex, open-ended items requiring a constructed response. Consequently, statistical estimations regarding how these field test items will perform when they really "count" are extremely tenuous.

Compounding this problem is the fact that the field test samples can be of woefully inadequate size and of unknown representativeness. Ideally, a large and representative sample of students from across the State of New York should participate under motivated conditions so that only technically-sound, fair items appear on subsequent Math A exams. Ordinarily, *minimum* sample sizes of 1500 students -- carefully chosen to proportionally represent important demographic characteristics in the State -- would be desirable to obtain stable, useful information regarding each item's difficulty, potential bias, and other characteristics. However, because participation in field testing is both voluntary and of essentially no consequence or benefit to test takers, we observed that sample sizes as low as 250 were used. To obtain even this many respondents, it was sometimes the case that the samples were not as representative of the State as would be optimal. To the extent that item parameters (i.e., the technical characteristics of items) are misestimated because of small, unrepresentative samples, the equating of the Math A tests (that is, the ability to ensure that the passing standard is equivalent from year-to-year) is jeopardized.

A second consequential weakness in the equating design is that the block of anchor items used to equate current forms of the Math A test consists of the identical block of items first used to anchor the score scale in 1998. It is the judgment of the content experts on the Panel that these anchor items most closely resemble items that would be appropriate for assessing the old Course I. They appear to be uniformly less linguistically and conceptually complex than, for example, the (non-anchor) items that comprised the June 2003 test. While item writers producing items for the 1998 Math A test may have attempted to generate items aligned with what they conceived of as the new Math A standards, it is clear that the conceptualization, understanding, and implementation of Math A as it has evolved are dramatically different. Thus, there has been an increasing disconnect between the knowledge and skills measured by the

anchor items and the knowledge and skills measured by operational items on current forms.

Previously in this report, we described the practical consequences of this disconnect. However, there are also serious technical consequences of this disconnect that likely resulted in an inaccurate equating of the June 2003 Math A test to previous forms and which, if uncorrected, have the potential to affect future test forms.

Our investigation revealed serious malfunctioning of anchor items which, we suspect, is likely attributable to the fact that the anchor items tend to reflect one "version" of Math A, while the other items in Math A forms reflect another "version." The statistical process of equating requires homogeneity of content tested -- that is, a single "version" of Math A should be evident in the anchor items and recent field test and operational items. This requirement is sometimes called *unidimensionality*. To the extent that instruction in classrooms resembles one version or another and, as a consequence, if students perform better on one type of item than another, the requirement of unidimensionality is violated and the statistical process of equating becomes inaccurate.

In the case of the 2003 Math A examination, the inaccuracy likely occurred because of the following:

- students performed approximately the same, or slightly worse, on the anchor items compared to their performance on the operational items (perhaps due in part to changing instructional practices, content coverage, etc.);
- this differential made the group appear relatively weaker than previous groups, and made the operational items appear relatively easier; and
- these relative differences resulted in the statistical procedure of equating producing a higher raw cut score for the June 2003 exam.

There is convincing evidence that this hypothesis explains the resulting lower passing rate initially observed on the June 2003 exam. The Panel requested certain analyses to shed light on this hypothesis. SED personnel and its contractor, Measured Progress, provided us with information in graph form that showed the relationship between the difficulty values (adjusted and unadjusted b-values) of the items in the 2002 operational form and its anchor items. These relationships appeared to be uniform and a statistical test of the slopes of the two regression lines would likely be non-significant (indicating similarity between what is measured by the anchor and operational items). However, the same information provided for the June 2003 test reveals a marked difference in these relationships, suggesting that the anchor and non-anchor items were, in fact, measuring somewhat different constructs for the 2003 administration. SED personnel have indicated that they will perform thorough statistical analyses on these relationships. We fully expect that the hypothesis stated above will be upheld by such analyses.

Finally, there is a secondary statistical problem with the anchor item set. This problem does not necessarily explain the anomalous results witnessed on the June 2003 Math A

exam, but it has the potential to cause instability in test equating for *any* administration. This problem involves the stability in performance of anchor items over time. In the most common and stable equating procedures, anchor items are administered under motivated conditions to the population of students as embedded items in an operational form. Thus, the difficulty levels (i.e., b-parameters) of the anchor items are usually highly stable. In such situations, changes in the b-parameters of an anchor item on the order of .30 in difficulty would routinely cause the item to be *excluded* from use as an anchor item in equating.[12] Recall, however, that for the Math A tests, the same anchor items are administered each year in order to calibrate new/field test items. Recall also the previously mentioned problem of small, potentially non-representative field test samples. Our investigation revealed large swings in anchor item b-values, with the magnitude of instability in the range of 1.0 logits (absolute value). Under such circumstances, any equating procedure would be highly unstable.

Recommendation 4A: *Alternative equating designs should be considered.*

Under current law, all operational test items must be released for public scrutiny. As a consequence of this constraint, New York is precluded from taking advantage of the most common and preferred equating design -- one that is known as an *internal anchor* design. Under an internal anchor design, a small proportion (approximately 20% of the total number of test items) of scored anchor items, representative of the test specifications as a whole, is embedded into the operational test. The internal anchor design is preferable because it ensures that anchor item information is based on the largest number of representative, optimally-motivated students as possible, and that each item counts toward a student's score and each item provides information used to make the eventual pass/fail decision.

Because an internal anchor design is not permissible under current New York law, an *external anchor* design may be an appropriate alternative. Using an external anchor design, a small, representative set of anchor items is still administered, but performance on the anchor items is excluded when calculating students' scores. Such a design is, however, sometimes subjected to the criticism that it is inefficient and educationally less desirable not to include information on anchor item performance when estimating students' overall competence.

Recommendation 4B: *Sampling procedures for estimating item performance must be improved.*

Sampling procedures must be revised to ensure that larger, representative, and more optimally-motivated samples of students participate in pretesting and field testing of Math A items. Changes in regulations to require mandatory participation of sampled units would be one possibility; education and persuasion would be another. Enhanced

[12] It may be helpful to realize that these difficulty levels (b-parameters) ordinarily range from -3.0 to +3.0 on a scale called the "logit scale," with an average or middle difficulty level of zero and a standard deviation of approximately 1.0. Thus, a change in b-value of .30 logits represents a change of nearly a third of a standard deviation in the item's difficulty.

auditing and monitoring of sample demographic characteristics and motivation by SED is also recommended to ensure stability of representativeness and comparability with operational data.

Recommendation 4C: *Replace the anchor item set.*

As noted above, the anchor item set is no longer representative of the remaining operational content covered by Math A exams. The equating item set must be reconfigured using items that represent a "mini-version" of the operational Math A test forms beginning with the January 2004 administration and for all subsequent forms. That is, the anchor item set must be representative of the breadth and depth of content coverage, complexity, and difficulty of currently operational Math A items.

Recommendation 4D: *Revisit performance standards (cut scores).*

Because the cut scores for the two levels of performance on the Math A exam were established under a different conceptualization of Math A, it seems imperative that the appropriateness of the current cut scores be reviewed and, if appropriate, revised. It is possible (and technically defensible) to maintain current standards via various methods (assuming that it is desired to maintain the standards eventually accepted for the June 2003 exam). Options would range from simply maintaining the same standard as was eventually applied for June 2003 to "affirming" that the standard is correct using a content expert/judgmental review. However, serious consideration should be given to revisiting the cut scores if the content standards and indicators remain the same. If the content standards and indicators change, a new standard setting study would be essential.

Finding 5: *The New York State Education Department cannot accurately predict performance on Math A tests.*

In educational testing programs, it is often desirable to be able to predict performance (such as overall pass rates, changes in subgroup performance, etc.) on tests. These predictions are usually only accomplished for a short-term outcome -- such as for the next administration of a test. However, even short-term predictions are useful for informing policy makers, gauging the resources that may be necessary to provide remediation or advanced coursework, and other uses. In many educational testing programs, the technical aspects of testing have been rigorously designed, refined, and controlled. And -- though we would wish otherwise -- in many educational settings, dramatic, pervasive progress or regress in learning over a short term is rarely observed. These two characteristics combine to result in changes in overall performance that can be reasonably accurately predicted.

For the Math A tests, largely as a consequence of our observations related to Finding 4 (see above), SED has been unable to predict performance characteristics of operational Math A test forms. It would be helpful, for example, if field test data can be used to

better estimate individual item performance, to signal warnings about future operational test performance, to suggest areas of strength and weakness in curriculum, to yield estimates of operational form pass rates, and so on.

Recommendation 5A: *SED should implement procedures for predicting the performance of test forms and groups of students on future Math A exams.*

It is likely that some research will first be required to identify statistical approaches that would be best suited for the context of Math A exams and would yield the most accurate predictions. However, there is an advantage that SED has data on past test form characteristics and group performance already in hand. Retrospective studies can be conducted to identify promising approaches.

Recommendation 5B: *Policies for field testing and data collection should be revised.*

In a previous recommendation (4B) we identified suggested changes in field testing data collection to address another issue. The adoption of this recommendation would also aid in addressing the issue of predictability identified here.

In addition to revision of field testing procedures, policy changes should also be considered that would permit SED to gather more timely and complete data on statewide operational test administrations. In order to analyze test information, policy makers and educators need timely and accurate data, and regulations should be put in place to assure rapid and efficient data reporting.

Finding 6: *Support and oversight for the Math A exam program should be improved.*

It goes without saying that all states are struggling to meet many demands. The requirements imposed by the No Child Left Behind Act, public pressures for greater accountability in education, shrinking budgets in a time of economic uncertainty, and other forces have placed stresses on all corners of the educational system. Important testing programs, such as the Regents examinations, have not been able to hide from these pressures, nor would they necessarily be immune to fiscal belt-tightening at a time when sacrifices must be borne by all.

On the other hand, high-quality, high-stakes testing cannot be done “on the cheap.” It is the strong impression of the Panel that the Math A assessment program has lacked the support it must have in order to produce with consistency tests that gauge the success of New York high school students on content they must master in order to gain a diploma. We sense that there is actually strong psychological support for the work that has been accomplished to date. Our conversations with the Commissioner and his associates revealed to us genuine respect for the talent, enthusiasm, and commitment evident in those who have worked on the Math A program. However, other forms of support are essential for ensuring the production of tests upon which important consequences hinge.

The occasion of having to examine the development of Math A exams also availed the Panel of several opportunities to observe internal and external processes. We believe that there is room for improvement in these processes. For example, in some data gathering, it was unclear where the primary responsibility rested for an activity; this made it difficult for the Panel to know where to direct requests for information and likely made it difficult for SED personnel to rapidly respond to such requests. In other cases, primary responsibility for an activity was diffused; in such cases, it did not appear that any single person had "the big picture" required for effective oversight.

Recommendation 6A: *SED should immediately increase in-house content and technical expertise resources by a minimum of one psychometrician and two math content specialists.*

The Panel observed that SED leaned heavily on external content experts to guide the development and support activities for Math A (and other) examinations. The number of content support positions at SED has been reduced over time. Further, we observed that, with the exception of the director, there is essentially no internal expertise in psychometrics -- that is, in *testing.* This strikes us as fundamentally inconsistent with the mission and activities of the testing unit. The director of the office does have a high level of expertise and experience in psychometrics, but the administrative duties of the director, i.e., oversight, support, and management of personnel and activities for 70 tests annually, greatly dilute this resource. While we are aware that there is a person assigned as a statistical support person, expertise in statistics differs from expertise in psychometrics (much like one cannot interchange a dentist for a physician). Additional support is essential to this effort.

The Panel did not engage in a full analysis to identify the precise level of staffing that would be appropriate for the assessment activities of SED; such an effort would likely be useful to ensure economical use of resources. Nonetheless, an elaborate study is not required to discern that current support levels are insufficient. It is our estimate that personnel should be increased by at least one person with psychometric expertise and at least two people with content expertise and experience in mathematics education.

Recommendation 6B: *SED should clarify the responsibilities assigned to its technical advisory committee, and should request this group to provide regular reports, including technical analyses, reactions to proposed changes in test programs, and suggestions for improving State testing programs.*

As an aid in the oversight, trouble-shooting, review of proposed changes, and initiation of new ideas for improvement in their testing programs, and other functions, many state assessment programs rely on technical advisory committees to supplement their internal resources and expertise. Such committees often consist of four to eight diverse external experts drawn from academics with expertise in psychometrics, alternate assessment for LEP or special needs students, directors of assessment from other states, or similar backgrounds. Such committees usually meet from two to four times

per year to review critical aspects of a state's testing programs, suggest ways to respond to technical challenges, identify and recommend ways to avoid potential problems, assist in developing plans of action, review or offer suggestions for proposed changes, develop alternative strategies for accomplishing key goals of the assessment program, and other activities as directed by the leadership of the state assessment program. New York State does have such a committee. This Panel is not clear as to how often the group has met, what its responsibilities have been, nor what its recommendations have been. The Panel believes clarifying these responsibilities and requiring regular reports in the future, would be helpful in terms of addressing technical issues.

Recommendation 6C: *SED should increase demands placed on contractors.*

In the course of requesting and gathering information for its investigations, the Panel had a few occasions on which to observe the activities or results of activities performed by external contractors. Our observation is that contractor performance is too variable. For example, one contractor conducted special analyses overnight when an urgent request was made. That was good. In another case, when the panel needed information on a Math A exam from last year, we were informed that a contractor had not yet provided the routine, annual documentation well after a year beyond when the test had been given. We believe SED should take a firmer approach to hold all contractors responsible for timely, accurate reports, documentation of all procedures, and responsiveness for data requests and analyses.

Recommendation 6D: *Internal coordination and documentation should be improved.*

As we noted previously, the Panel sometimes observed that roles and responsibilities related to production of the Math A exam may be too discrete. We believe that SED should consider reorganization plans that would enable coordination of each testing program and locate "the big picture" for a project within a single individual. Further, SED should develop its own, internal "historical annals." Such documentation would consist of organized, centrally-located documentation in which all relevant technical and other related information about a test is maintained. Beyond assisting in time of need, such as was the case for the current Math A investigation, such documentation would also assist SED in times of personnel changes, for training purposes, for effecting smooth transitions and sharing of information between contractors, and other benefits.

C. Findings and Recommendations Concerning Statewide Infrastructure Issues Related to the Attainment of Math A Standards

Finding 7: *Passing rate data for the State as a whole were not available until three months after the exam; no data are collected regarding student performance on individual items, nor even regarding student performance on the four parts of the exam.*

The Panel was surprised at the lack of data concerning the Math A test. As this is being written in September, total failure rates have just become available. SED cannot analyze the functioning of its assessments if it does have item level data. The Panel recognizes that SED is moving toward a more comprehensive program of data collection, and believes this needs to be an important priority.

Recommendation 7: *SED should increase its data collection capacity to include item level data, and should accelerate its data collection timetable.*

Data should be collected earlier, and data should be collected at the item level to determine whether the assessments are functioning in accord with their design.

Finding 8: While *the most important use of student performance data is to inform instruction, statewide data mining models that would enable local schools and teachers to use these data effectively are not generally available.*

To practitioners, data is only useful if it is available and can inform instruction. Although the Panel is aware that there are some efforts to assist districts with the effective use of data, these efforts need to be broadly expanded.

Recommendation 8: *SED should substantially broaden its efforts to assist districts in data collection, and the use of data to inform instruction.*

Finding 9: *The mathematical background of teachers delivering math instruction varies widely; yet, raising almost three million children to higher levels of math achievement will be impossible without highly skilled teachers.*

Recommendation 9A: *SED and higher education need to continue and to strengthen their partnerships to ensure strong teacher education programs, both pre-service and in-service.*

Recommendation 9B: *The certification requirements for elementary teachers and special education teachers should include a minimum of nine credits of college level mathematics (see Recommendation 9C), and three credits of teaching techniques in mathematics*.

Recommendation 9C: *Mathematics courses required for certification, both for mathematics teachers and elementary and special education teachers, should be specific not only in terms of number of credits required to be taken, but also in terms of coursework required to be taken, e.g., calculus, number theory, algebraic structures, probability and statistics, etc.*

Recommendation 9D: *The Panel believes that, for any teacher responsible for teaching mathematics at any level, the 175-hour professional development requirement should include specific mathematics requirements. The Panel's thinking is that:*

- *teachers who teach mathematics exclusively should be required to take 100 of the 175 hours in the area of mathematics;*
- *secondary teachers who are certified in, and who teach in, more than one subject area, should be required to take 50 of the 175 hours in the area of mathematics;*
- *teachers who teach mathematics as part of a broad set of teaching responsibilities, e.g., elementary teachers and special education teachers, should be required to take 30 of the 175 hours in the area of mathematics.*

Additionally, the range of possible courses that would satisfy these requirements should be clearly specified.

Finding 10. *The public has very little awareness of Math A, and may have misunderstandings about the goals of Math A.*

Several Panel members recall the emphasis SED placed on the importance of increased literacy, and sees that as a model that can be applied to mathematics.

Recommendation 10: *Make greater use of SED communications capacity to engage the public in conversations about the importance of strong mathematics skills.*

Finding 11: *There is often a "disconnect" between K-12 and higher education.*

A few years ago, SED encouraged local conversations between leaders of K-12 schools and higher education, i.e., regional meetings involving college presidents and superintendents of schools. As the Panel has reflected upon the enormity of the task of raising every child to Math A levels, it would seem advantageous for the gap between K-12 and higher education, in mathematics, to be bridged. We envision meetings of local high school math teachers and college math professors to review their programs and curriculum, and to explore collaborations.

Recommendation 11: *SED should encourage conversations at the local and regional level of K-12 teachers of mathematics and higher education professors of mathematics, for the purpose of sharing curriculum, and exploring professional development opportunities and other possible collaborations, to bridge the gap between K-12 and higher education.*

Finding 12: *Raising the level of mathematics achievement of all students to high levels must start when children are very young, and must go beyond the school day for school aged children.*

We know from brain research that learning is connected with neurological development, and such development occurs at an early age. In order for children to be proficient in mathematics at the high school level, they need exposure to good mathematics at a young age. Typically, in schools, children are exposed to mathematics for approximately one period, or 45 minutes per day. The Panel believes we need to move beyond the capacity of public schools, perhaps establishing partnerships with local public libraries to implement programs to children at a very young age, and also on an afternoon, weekend, and summer basis during a child's school career. We would envision that these programs would be designed to help young children become as excited about ideas of mathematics as they are about reading a new book. We believe that, for mathematical skills and concepts to be learned, they have to be viewed as important as reading skills, and this means bringing other partners to the table.

The Board of Regents is in a unique position, given its broad oversight of educational functions in New York State. The Panel believes there are opportunities to expand this effort well beyond the doors of the state's K-12 schools. Libraries have programs for pre-school; public television is viewed by even the youngest children; some museums have science programs and exhibits, which can be expanded to include more mathematics. To enable all children to reach high standards in mathematics will require societal and cultural changes which will only occur if all of the forces are aligned in the same direction. The schools alone will not be able to do this work.

Recommendation 12. *SED should encourage through grants and other means the expansion of mathematics education initiatives beyond K-12, such as the creation of partnerships between schools and libraries, and the greater use of public television and museums.*

D. Findings and Recommendations Concerning Additional Issues

1. Scoring Rubrics, and Communication to the Field Regarding Grading

Finding 13: *The scoring rubrics do not give credit for a variety of mathematically correct approaches.*

While an important goal of Math A is to encourage multiple approaches to solving a problem, some Math A exam questions force a student to solve a problem by one particular approach, e.g., item 35 on the June 2003 Math A exam. The June 2003 scoring rubrics also had instances in which only one approach received full credit, or where more favorable treatment in partial credit was given to one approach over another. The Panel recognizes that it is very hard to develop a comprehensive grading rubric that anticipates the credit that should be earned by unexpected approaches, whether students get the right answer or are on the right track. A more holistic approach to scoring rubrics may be needed, one perhaps more similar to the rubrics used in the International Baccalaureate Math program. The Panel believes there must be room for teachers to apply professional judgment in the grading of student work.

Recommendation 13A: *Develop more generally worded, holistic scoring rubrics which permit credit to be granted for atypical, but mathematically correct, student responses.*

Recommendation 13B: *Rubrics should be designed so students do not lose 33% or 50% credit for a minor arithmetic error.*

Finding 14: *There is a serious "disconnect" between the perception of the SED content specialists and the perception of field classroom teachers regarding the application of the scoring rubrics.*

At one point, during intense discussion about a particular rubric, SED staffers stated that the rubrics are general guides for grading; the room became very quiet, and one classroom teacher stated that the field understanding is that the rubrics are to be applied with little latitude. Several Panel members joined in that view. They seem led to that thinking by the language in the scoring guide which states several times that the "specific criteria" are to be applied. The Panel welcomes the concept of flexibility and recommends this flexibility be clearly communicated to the field.

Recommendation 14: *On each set of directions for the Math A exam, a statement should be added confirming that the scoring rubrics are a guide and should be applied using professional judgment.*

Finding 15: *There needs to be better communication of SED grading interpretations during the grading process for the Math A exams.*

As the Panel discussed the grading of Math A exams, it became clear that SED staff are very available to answer questions from the field, and that they also encourage the flexibility noted above. The problem, though, is that SED does not have the resources to reach out to every district, and many teachers will not think of calling SED except in the case of a very serious matter. Thus, it is entirely possible that teachers who call SED will apply the grading rules differently from those who do not. When the Panel held this discussion, SED staff advised that they are exploring a website that would be activated during Regents exam grading that could provide up-to-the-minute responses to grading questions. Although this does not guarantee that the information will get to everyone, it is a definite step in the right direction, and the Panel applauds this initiative.

Recommendation 15A *SED should continue on its path of setting up a website during Math A Regents exam grading to provide up-to-date clarifications to teachers grading the exam.*

While the website is a step in the right direction, websites are "pull" technology, i.e., the user must pull the page up to get the information; and it is possible that there are still schools without web access. The Panel recommends that thought be given to "push" technology, whereby the information would be pushed out to every district. Right now, errata sheets are faxed to districts when there is an actual error on the scoring sheet; perhaps thought should be given to a fax every few hours after the exam (during the school day) up to 48 work hours after the exam, to send out grading clarifications.

Recommendation 15B: *SED should explore ways of sending up-to-date grading clarifications to the school districts during the grading period following the administration of the exam, as a backup to the website, to ensure the greatest possible consistency of grading across the State.*

2. Calculator Use on the Math A Exam

Finding 16: *Allowing the option of using a graphing calculator on the Math A exam provides some students with an advantage on the exam, thus creating an inequitable situation.*

Students who are able to afford graphing calculators, or who live in school districts that are able to provide them with a graphing calculator, have a distinct advantage over other students if they are permitted to use the graphing calculator on a Regents exam. While the Panel agrees that students should be taught how to use graphing calculators, permitting the optional use on the exam provides an advantage to some students. The Panel believes testing conditions should be the same for all students.

Recommendation 16: *The use of calculators on the Math A Regents exam should be standardized*

The Panel recommends that, until the State can be sure that every child has access to a graphing calculator on the Math A exam, the use of these calculators should not be permitted on the exam.

E. Recommendations Concerning the January 2004 Exam, and All Math A Exams until A New One Is Designed.

The Panel has noted earlier its concern that the January 2004 Math A exam was created at the same time as, and under the same pretest and field test conditions as, the June 2003 exam. The Panel is concerned that there is much we do not know about why the June 2003 exam behaved the way it did, particularly the items in Parts III and IV. As the Panel sees it, the recommendations contained above in this report represent a plan for redesigning the mathematics standards and assessments, and this plan will result in a completely revised Math A exam at the end of the process. The Panel believes it has a responsibility to make recommendations regarding the Math A exams in the interim, and presents these recommendations here.

Recommendation 17. *Until the standards are rewritten, new curricula are developed, the new course is delivered, and a new Math A Regents is designed and field tested, the Math A Regents exam should be restructured so the exam includes: 30 Part I items, 5 Part II items, 2 Part III items, and 2 Part IV items.*

The largest problems the Panel saw with the June 2003 exam were with the items in Parts III and IV. By reducing the number of items in those parts and increasing the number of items in Part I (which did not demonstrate the same problematic performance), the Panel believes that this somewhat modified exam can be an effective measure of student performance until a new exam is developed based on the rewritten standards, with the provisos below. Additionally, this recommended configuration reduces to some extent the concern about curriculum coverage, as it calls for 39 items rather than the current 35, thus increasing the content coverage of the exam.

Recommendation 18: *The exam should be reviewed by a group of practitioners, including math teachers, university mathematicians and mathematics educators, with representatives from this Panel, prior to the administration of the exam.*

The Panel understands SED already instituted such a quality control step for all August 2003 Regents exams, and intends to do so for future Math A exams.

Recommendation 19: *Until new items are developed and properly field tested, the exam items should be scaled in accord with the procedures used for the August rescaling of the June 2003 exam.*

Recommendation 20. *The scaling should not be finalized until after the exam has been administered and after a post equating procedure has been implemented to ensure the fairness of the test.*

The Panel understands SED intends to do so for future Math A exams.

If the above measures are put into place, the Math A exam should function somewhat similar to the June 2003 exam, after it was rescaled. The Panel believes that, until the

standards, the curriculum, the assessment, and the infrastructure are in place, students should be held to the same standard as last June's students, which leads us to these final recommendations.

Recommendation 21: *The 55 passing option on the Math A Regents Exam for a local diploma should be continued until after the standards have been clarified, after new curriculum has been developed and disseminated, and after a new exam has been developed and administered for at least one school year (to ensure that it is performing in accord with its design).*

Recommendation 22: *The math RCT safety net for special education children should be continued until after the standards have been clarified, after new curriculum has been developed and disseminated, and after a new exam has been developed and administered for at least one school year (to ensure that it is performing in accord with its design).*

F. Suggested Timeline

As the Panel reviewed its thinking with SED representatives, it was suggested that the Panel draft a timeline that might serve to guide the process. The timeline we suggest is below.

<table>
<tr><th></th><th>Test Development</th><th>Standards</th><th>Curriculum</th></tr>
<tr><td>Oct 2003</td><td>Immediately:
Create three exams to be administered:
Jan 04, June 04, Aug 04,
with format:
30 Part I (2 points)
5 Part II (2 points)
2 Part III (3 points)
2 Part IV (4 points)
using current item pool, aligned with current core curriculum item sampler, and scaled used in August 2003 rescaling of June 2003 exam, with each complete exam to be reviewed before administration.</td><td>Immediately:
Form Mathematics Standards Committee.</td><td rowspan="13">Immediately:
Examples of high quality K – 8, Math A, and Math B curricula selected and disseminated to the field.</td></tr>
<tr><td>Nov 2003
Dec 2003</td><td rowspan="2">Nov 03 - Feb 04:
Create new item pool (using Checklist of Writing Items) for tests to be administered Jan 05, Jun 05, Aug 05, Jan 06 under same conditions as above.</td><td rowspan="5"></td></tr>
<tr><td>Jan 2004
Feb 2004</td><td rowspan="4">Jan 04 - Dec 04
The Mathematics Standards Committee retools the standards.</td></tr>
<tr><td>Mar 2004
Apr 2004</td><td></td></tr>
<tr><td>May 2004</td><td>May 04: Field test.</td></tr>
<tr><td>Jun 2004
Jul 2004
Aug 2004
Sep 2004
Oct 2004
Nov 2004
Dec 2004</td><td rowspan="2"></td></tr>
<tr><td>Jan 2005</td><td rowspan="4"></td><td rowspan="4">Jan 05 - Jun 05
Curriculum committee writes or chooses exemplar curriculum, aligned to retooled standards.</td></tr>
<tr><td>Feb 2005
Mar 2005
Apr 2005</td><td>Feb 05 - Mar 05
New Items written aligned to retooled standards (and old items reviewed to salvage any that are aligned).
Work should be guided by Checklist of Writing Items.</td></tr>
<tr><td>May 2005</td><td>May 2005: Field test.
Purpose: Create three tests for actual administration:
Jun 06, Aug 06, and Jan 07.</td></tr>
<tr><td>Sep 2005</td><td>Between Sept and May, pretest, field test, set new performance standards (bookmark).</td><td>New one year Math A course taught across the State.</td></tr>
<tr><td>June 2006</td><td>First Administration of new Math A exam.</td><td></td><td></td></tr>
</table>

VI. Summary and Conclusion

The June 2003 Math A exam results clearly point to a need for substantial change. After the rescaling recommended by this Panel in its Interim Report, 45% of the State's children failed at 65; 59% of New York City's children failed at 65.

This report stated earlier:

> Many states have implemented higher standards and required mastery of more rigorous content. As might be expected, when new standards are introduced, overall performance is often at lower-than-desirable levels. However, when the new content standards are clearly specified, when instruction can be focused on the content standards, when tests can be created that are more fully representative of and aligned to the content standards, fairly large increases in average student performance are routinely observed (p. 28).

This Panel believes that, if the recommended streamlining and clarification of the standards occur, and if the Math A course is streamlined to a year long course (after the K-8 standards are aligned), and if there is greater curriculum guidance to teachers and districts struggling with this effort, and if the other recommendations in this report are accepted, many more students will reach high levels of mathematical knowledge and skills. There will still be some students who, despite enormous efforts by them and their teachers, will not reach this level. This Panel believes discussion must continue to find ways of helping these students find success; this is outside the charge of this Panel.

As we have worked, we have reflected that, in our modern society, people often look for "sound bite" answers to even the most complex problems. We see the Math A situation as very complex, and we hope our recommendations reflect that sense of complexity. We also hope our thoughts prove helpful as we move forward.

Raising the level of mathematical skill and knowledge of millions of children is a daunting challenge, but it is a challenge this Panel agrees must be faced. While taking on this challenge, we all need to appreciate the enormity of the effort, and we must be cognizant of the wide variety of children who enter our school doors every day. As our children are not homogenous, our solutions for them cannot be homogenous. Everything we do must be sensitive to their varying individual needs.

In closing, we once again express our appreciation to the staff at SED. Even as we were working, SED continued exploring possibilities. The Panel is aware of the Statewide Math Initiative recently formulated and the Panel believes this is exactly the type of creative thinking that will move this effort forward.

The members of this Panel have been honored to have been asked to help find a solution, and we offer our help in any way that might be needed in the future, so that we adults can get this right -- for the children we all serve.