

ASSESSING THE ASSESSMENTS - JUNE 2002

CURRICULUM AND INSTRUCTION COMMITTEE

NYSCOSS

Walter J. Sullivan, Chairperson

Introduction

This document, as in the past, is a synopsis of feedback concerning the assessments from various Superintendents as well as an analysis from the Skaneateles School's Curriculum Coordinators and staff. It provides a review of those assessments that New York's youngsters have taken during the 2001-2002 school year. It is meant to be an objective review of New York's assessments as it affects the education of youngsters in New York State. Each and every year the Curriculum and Instruction Committee of NYSCOSS attempts to provide the State Education Department and superintendents with a capsule view of this past year's assessments.

2002 Administration of ELA 4th Grade Assessment:

Here are some observations regarding the administration, content, and scoring of the Grade 4 ELA 2002.

1. The content of the multiple-choice questions, Listening, Reading, and Writing were within reasonable limits. Some of the response questions were vague and required information that seemed difficult for 4th graders to understand. However, in past years when the questions were of lesser quality, the scoring corrected it by allowing remote answers. This year the scoring seemed to be more rigid, especially in the scoring of writing. The rubric did not allow many options that were typical of 4th grade responses.
2. The scoring training done by Measurement Inc. on the video was fairly helpful in training the teachers for scoring. However, teachers had to edit and enhance the training so that they would be efficient scorers. Again, Measurement Inc. seemed to have changed their perspectives on the interpretations of responses. It appeared to be more rigid.
3. Overall the instrument this year seemed to be adequate. It continues to be excessively long in administration for this age student. It is especially difficult for many struggling students to sustain this type of focus over this period of time. If the results were diagnostic, then the time and effort might be more commensurate with the effort. Our observation is that the Grade 4 ELA is not that useful in guiding program modifications.
4. We believe that the percentage of students at level 4 dropped this year, probably due to the more rigid scoring techniques. More data analysis statewide needs to be done on this issue.
5. Pattern scoring vs. IRT – The Curriculum and Instruction Committee is still somewhat confused about the apparent changeover from the so-called “pattern scoring” to the IRT scoring where the items are weighted on the level of difficulty and are randomly selected. Our technical committee needs to continue to work with Gerry DeMauro, Director of Testing, so that we are able to communicate this change to the Superintendents at large.

Time Devoted to ELA 4 Test

In an average size class in New York State (148) approximately 64 hours was devoted to the administration of the test, 20 hours of training to score the test, 72 hours to actually scoring the test, for a total of **156** total hours involved in the ELA 4.

2001 Administration of the 5th Grade Social Studies Assessment:

The following are observations regarding the administration, content, and scoring of the Grade 5 Social Studies 2001.

1. The content of the first Social Studies exam seemed to be similar to the test sampler. The students had to spend two sessions, each test took one and a half hours long to take. The teachers and students generally felt it was a fair test.
2. The scoring of the test was done right after the administration. The scoring used a grid to give you the total score or grade. The scoring was easy and efficient.
3. Overall, the new Social Studies test was fair and appropriate for the students. However, it was too long for the amount of minimal information that it generates. It appears that students who can read and write well will also do well on this social studies test. It is also difficult to decide what should be remediated if a student fails since the social studies content for grade 5 and 6 is not American or New York History.

2002 Administration of Math 4 Assessment:

The following are observations of a) the test construction and b) impression of each section of the assessment, c) anticipated results, and d) scoring.

Session I

Test Construction: The 30 multiple-choice questions addressed each of the 7 Key Ideas of the NYS Math Standard (including fractions, place value, money, patterning, area, line segments, units of measure, time, and estimation) in seemingly direct ways. There were few distracters placed as choices.

Student/Teacher Impression: Question 18 was presented in a way that required the students to read carefully. The question only provided partial information. The rest of the information to be considered was in the choices provided.

“Ms. Star bought cans of soda packaged in groups of 6 cans. If Ms. Starr bought only complete packages, how many cans could she have bought?”

F 26

G 28

H 30

J 32

The vocabulary used in the test construction was grade level appropriate. Most teachers appeared to feel that the test was fair, with most students able to complete it within the 40 minute allotted time-period. Students with modifications were able to complete the test within the limitations of their IEP.

Session II

Test Construction: There were 9 open-ended questions. *Two* questions required students to provide a *written explanation*. Three questions required the students to show their work to receive full credit.

Question #37 was a graph that was to be constructed from data in a table. The directions for this item included a list of things the students were “Be sure to” include. These points were specifically included in the scoring for **full** credit. In the past, full credit was awarded even if some of the “bullets” were missing.

Question #36 instructed the students to use their counters. There were also directions and a space provided for the students to show their work. This presented some confusion for those students who did use the counters to solve the problem, but then did not understand that they were to reproduce (draw) how they used the manipulatives to arrive at the solution. Full credit could only be awarded for correct answers accompanied by work shown.

Student/Teacher Impression: Overall, teachers reported that they felt the test was a fair test. Each of the concepts had been covered. Student questions related to clarification of what questions were asking rather than students not understanding the required mathematics for the problem.

Session III

Test Construction: There were 9 open-ended questions. *Three* questions required students to provide a *written explanation* of how they arrived at their answer. *Five* questions required the students to show their work for full credit.

Student/Teacher Impression: Fair assessment of K-4 knowledge.

Student Performance: Many students demonstrated difficulty on one item, question # 47. The directions required the students Part A to “...*make 1 triangle by tracing around 2 of your pattern blocks. The sides of your triangle must be exactly as long as the sides of the green pattern bloc.*” In Part B it asked “*How many green pattern blocks would it take to cover a triangle that has sides exactly as long as the green pastern block.*” This prompted students to trace additional pattern blocks on Part A to figure out their answer. Inclusion of this additional tracing on Part A, did not allow the student to receive full credit – as they were required to trace around only 2 pattern blocks.

Time Devoted to Math 4 Test Scoring Process

Approximately 25 hours of training to score the test, 50 for scoring and a total of approximately **117** total hours were devoted to the scoring process.

2002 Administration of Science 4 Assessments:

Items on Science 4 were very similar to the items and format of the 2001 Administration of Science 4 (See Assessing the Assessments: 2001)

GRADE 8 ASSESSMENTS

Overview

The best thing that could happen for New York State's eighth graders is to keep the ELA 8 and Math 8 and eliminate everything else at that level.

The NYS 8th Grade Technology Assessment:

We applaud the Commissioner and the Board of Regents for providing an option to school districts in terms of the administration of the State Technology Test. This option provided some flexibility and therefore, relief from the burden of over testing 8th grade students.

The NYS 8th Grade Math Assessment:

This assessment was similar to the math assessments since 1999. The administration and scoring went smoothly with one exception. The week of June 17th, nearly five weeks after giving the assessment and then scoring it (districts had also sent the answer sheets to BOCES), NYSED sent a memo addressing the scoring of one extended response item. It was a question that could result in two answers, depending upon how the question was interpreted. The memo said that either answer was acceptable. Some districts were fortunate that they had teachers who went for scoring training and already knew about the ambiguous question. They had allowed for both answers. Some districts were probably unaware of the issue with the question and had to recall their answer sheets from BOCES for re-scoring. Although the process went fairly smoothly, the scores need to be received in a more timely fashion. **ONCE AGAIN, THIS HINDERS SCHEDULING FOR 9TH GRADERS.** At the district level, students are recommended for Academic Intervention Services in grade nine based on their achievement in grade eight math. The fall release of the NYS math score reports requires that we change student schedules after the beginning of school. This time frame should be rectified so that we may receive them sooner this year.

Time Devoted to 8th Grade Mathematics Test

There were 21 hours devoted to the administration of the Math 8 test, 19 hours training to score, 63 hours for the actual scoring, for a total of **103** hours devoted to the Math 8 test.

2002 Administration of 8th Grade Science Assessment:

Performance Test: The setting up and administration of this test went more smoothly this year. No one needed additional training that could not be done informally during the set up. There was an error in the student booklet on the y-axis of a graph where the scale was incorrect. One teacher discovered this error early in May (as you know the performance test can be given by a district any time in May). The teacher sent an email to the NYSED but at no time was the district officially notified by the State of the error. It probably did not affect most students, but it would have been helpful if the State acknowledged the error and gave instructions on correcting it.

Written Test: The written test was straightforward and seemed easy for students. Grading was an issue just like last year. In an average size middle school, seven teachers spent one day scoring the written test and half a day bubbling the answer sheets. Similar to last year some of the extended responses that the students answered were actually multiple-choice questions. So, the students

circled their answers in the test booklet and the teachers had to bubble in their responses on the answer sheets. This is ridiculous! Last year we raised the issue of the answer sheet. We really need two answer sheets for the science assessment – one for the performance test and one for the written test. The answer sheets should be specific to the test, just like the math assessment. Generic answer sheets may be cheaper to print, but they take longer for the teachers to complete. The costs are shifted to the districts. We hope next year we can have two answer sheets for these tests and that they are easier to use.

Districts which score over 90% of students at level 3 and level 4 should be **exempt** from the test for the next five years or, at the very least, districts could just give the performance test which tests science process skills and be exempt from the content test.

Time Devoted to the 8th Grade Science Assessment Test

There was approximately 7.5 hours for set-up, 40 hours for the administration of the test, 14 hours for training to score, 101 hours for the actual scoring, for a total of **162.5** hours devoted to the Science 8 Assessment.

2002 Administration of the Social Studies in 8th Grade Assessment:

Some of this information is the same as previous years, but, of course, many of the issues continue from one year to the next and in some cases, the outstanding issues remain unresolved.

The test itself is a fair assessment of student progress. The biggest issue with this exam is the enormous burden it places on teachers and administrators to rate the exam. Once again, an average-size school district had five teachers devote 4-12 hours each rating the exam. In addition to that time, the teachers had to be trained for the sorting. Finally, there are three different forms for the partial scores each student receives in the exam. All of these score reports have to be filled out, reconciled, and then finally transferred onto the final answer sheet. This paperwork alone takes hours and hours.

The exam itself should be given in ONE three-hour sitting. Currently, Part I is given in a 90-minute block. Any student who finishes early must wait until the 90 minutes have passed before he or she can start Part II. This is another significant waste of time.

Time Devoted to Social Studies 8th Grade Assessment

There was approximately 6 hours for the set up, 18 hours for the administration of the test, 30 hours of training to score the test, 60 hours for the actual scoring of the test for a total of **114** hours devoted toward the Social Studies 8 Test.

The 2002 Administration of 8th Grade English Language Arts:

Although the test itself is fairly well designed, it is still far too long. 8th grade students should not be expected to sit for two lengthy sessions on two consecutive days (followed by two more sessions for math!). There is most certainly a reliable and valid way to assess students in Language Arts using a single three-hour session. Furthermore, since the test preparers insist on using social studies material for the readings (this year's passage was on food and culture from the

Italian Renaissance) why not combine this assessment with the social studies test (identified last year as well). After all, that test assesses reading and writing abilities as well.

Again, the scoring of this exam is a significant burden. Even in an average-size school, seven teachers participated in the scoring. Each had to miss a day in the classroom for both training and rating. This rating becomes yet another burden on the middle school. All of this preparation, training and scoring results in a final score for the student of 1-2-3 or 4. These scores do not tell us a single thing about the students' ability in Language Arts that we didn't know before the exam.

Time Devoted to ELA 8 Test

There was 10 hours devoted toward setting up the ELA 8 test, 20 hours for the administration of the test, 21 hours for training to score, 49 hours for the actual scoring for a total of **100** hours devoted toward the ELA 8 Assessment.

Below you will find a summary of the chart that indicates for an average size school in New York State the amount of time involved in the testing. The grand total devoted toward New York State Assessments in grade 8 was approximately 558.5 hours for an average size school in New York State.

Staff-hours spent in the spring of 2001 on the NYS Assessments in Grade 8

	Set Up	Administration	Training to Score	Score	Total
*Technology 8	N/a	4	18	8	30
Science 8	7.5	40	14	101	162.5
Social Studies 8	6	18	30	60	114
Math 8	N/a	21	19	63	103
ELA 8	10	20	21	49	100

Grand Total

509.5 hours

*If districts choose to give State Assessment

In Grade 8, assuming staff works a 7-hour day, approximately 80 staff days were devoted to these tests.

General Considerations

- The seriousness and maturity level of a typical eighth grader, particularly if the test is construed as not counting both by students and staff creates a situation where it affects the quality of their work.
- The multiplicity of the testing removes the classroom teacher from the classroom, robbing more time from the curriculum.
- It also creates a situation with middle school youngsters who are active already become more stressed out.

- As was mentioned on a number of occasions it is the exit exams which really count, not the intermediate exams that may or may not have any correlation to the graduation requirements.

We are in favor of the ELA test being given in January or February and the Math test given in May. We would also propose that the Science performance exam be given at the end of May or earlier in June.

REGENTS

Process For Recall of Regents Exams and Scores

First, we would like to address the process of having Regents exams “called” by the State for scoring review. Exams such as Living Environment and Global Studies which were called last June 2001 for review were not returned until February 2002. A memo from Gerald DeMauro (dated February 13, 2002) said that the score report forms would be shipped 3/5/02-3/14/02. These score reports were returned, one year after the exams were “called”. A quicker turnaround, however, would have been appreciated.

2002 Administration of the New High School Regents Exams in Science:

The Living Environment and Earth Science tests were very fair, given the generous scaled score chart. As we discussed, there were serious issues with the Physics test and, to a lesser degree the Chemistry test, in terms of the scaled score chart.

2002 Administration of the English Regents:

Although the content of this year’s test is much improved, it continues to be far too long. Why is English the only Regents exam that has to be given in two three-hour sittings? Certainly a three-hour exam could be designed which would effectively and reliably assess students reading and writing. Perhaps part of a students’ Regents English score could be derived from writing completed during the school year. After all, that practice is already well established in the Foreign Language exams, as well as the science exams, where students are given credit for work done before the exam.

The content for this June’s exam was interesting and engaging for students. Even the Part I B section, reading and writing for information, seemed to hold the students’ interest. It was about the correlation between teenagers’ sleep patterns and school achievement.

Once again, the issue for this exam is the significant amount of time and human resources required to rate the exam. In an average size district, it took eight teachers five full days to train for and rate these exams. There are nearly 1300 essay readings.

The grid for this June’s exam was really quite fair. In fact, this was the highest amount ever of cells which result in a score of 90 or higher.

2002 Administration of the Global History and Geography Regents:

The exam and its results are very similar to those of the last two years. Teachers did report that the exam was somewhat more difficult. Because this exam also includes extensive student writing, scoring becomes a huge task, requiring several teachers and many hours of work.

One significant issue with the test was the documents on the DBQ. This year's DBQ was about human rights. One of the documents showed a black and white photo of a policeman looking at the passes of a group of minority people. The caption read, "Show me your pass." There was no indication about what country was being depicted or what time period. Some students interpreted the photo to be India while others thought it was South Africa. The documents should be clearly labeled. Another document has no attribution at all. It was a long text about the topic with no indication who the writer was. It could have been from some 7th graders' web page for all we know. We would never allow students to present such undocumented evidence.

2002 Administration of the US History and Government Regents:

This exam is also quite similar to last year's. Again, teachers reported that it was a bit more difficult. Just as with the Global History exam, the rating of the essays becomes an incredibly huge task. In fact, some schools had to hire a substitute teacher to help deal with the paper load.

One discrepancy that was discovered on the scoring grid for this year's exam was that there are 77 cells which yield a score of 85 or higher. That is fairly close to last June's grid which had 70. But the grid for January, 2002 has 176 cells 85 or higher! That's astounding! Ostensibly, the state will claim, that the increased number of cells makes up for a more difficult test. It is almost unimaginable, though, that the test was so difficult that it required about 150% more cells at that level. Clearly, students who took this test in January had a great advantage.

2002 Administration of the Physics Regents:

2002 PHYSICS REGENTS EXAMINATION: A PROPOSAL

I. Background

The New York State Council of School Superintendents, Curriculum and Instruction Committee received numerous phone calls and e-mails within one day concerning the physics regent's exams relative to the scoring chart on the exam itself. The results in physics this year were dramatically lower than in all previous years. In many schools the superintendents indicated that the percent of failure was significantly higher than last year. One example is a high performing district, which had 92% of its students pass last year and this year, only 40%. Last year it had 25% of its students at mastery, this year it had 6%. This type of pattern throughout New York State based on the communication and e-mails from the field is startling and unfortunate for the physics students particularly for those seniors who, in some cases, are relying on this course to graduate. The Council of School Superintendents will be willing to share the e-mails and material disaggregating the data concerning this very unfortunate outcome.

The main issue with the June, 2002 physics regents exam is the scale score conversion chart. Based on our experience with these conversions from many past administrations of Math A regents and now three administrations of the Living Environment and Earth Science exams, the June, 2002

regents examination conversion chart was in direct contradiction to the others. We do understand the scale score-setting process, matched against standards. We also know that it is not appropriate to look at percentage of points earned by a student. In the test we know that the State is interested in the difficulty of individual questions on the items and that a logarithmic process is used to generate the conversion chart. However, being aware of all of these issues, a reasonable person would have to conclude that the June, 2002 physics chart was generated the way it was because the June, 2002 regents physics exam was “easier than both the living environment or earth science examination”. This is clearly not the case. The exams did not differ significantly in terms of degree of difficulty.

We believe that the poor performance statewide has little or nothing to do with teachers not teaching to the standards, or “seniors suffering from senioritis” (juniors in many districts did just as badly). It has to do with the conversion chart.

We understand how the scores are generated on the new regents exam in physics. Unfortunately, parents, students and others may focus on percentages especially if a grade on the physics regent’s exam is the difference between a regents or a local diploma or between graduating and not graduating. The Council will leave it up to the SED to explain to parents how a student who gets 65% of the possible points on a regents physics exam correct, still fails the tests.

II. Possible Solution to the Problem

At this juncture, we would encourage the SED, in an attempt to remedy this issue, **to generate a new chart**. There is precedence for this action. In June, 1999 a corrected edition of the Math A conversion chart was distributed. We have attached a possible chart that could be utilized for this June’s physics examination. We have also attached a comparison of this year’s Learning Environment exam and the physics exam scale score charts in graphic form. You can see that a 39 out of 85 on the Learning Environment copy of the exam converts to a score of 65, but on the physics exam the student needs 58 out of 85 (see Living Environment Chart).

III. Unintended Consequences

In addition, as a result of this conversion problem, there are a number of issues that the Curriculum and Instruction Committee has identified:

- a. Who are the individuals who write the questions and set the scale score in physics? (Previously asked of other Regent’s exams)
- b. The State Education Department “regent physics sampler” did not fairly represent the 2002 regents.
- c. As recommended, many staff members in New York State spent two weeks teaching the components of the standard model – not one question was asked that required students to use the standard model chart.
- d. Similarly, two weeks of instruction was spent on reading comprehension- type questions. While there were word problems, there were no reading comprehension questions.
- e. In total, four weeks of instruction were not represented through any questions at all.
- f. Types of questions generally considered “Part II types” (multi/step calculations) were asked in Part I starting as early as question 5.

- g. The long-term effects of this exam can be very serious to many students since physics is essentially an “elective” program and future students may “vote with their feet” and simply not take physics. This certainly would be an unintended consequence that would pose a serious problem for both schools and students.

The thoughts above represent a very quick turnaround to this physics exam problem. We are confident the SED will make the proper adjustments in order not to penalize the present students who were, in a sense, victims of a conversion chart. Thank you very much for your consideration and future action.

PHYSICAL SETTING/CHEMISTRY
Regents Exam, June 2002

There are a few problems with the Regents Physical Setting/Chemistry exam. Question 39 has 3 possible answers (out of 4). Question 67 has units listed for electro negativity, (9/ml.) when there are no units. This is confusing for students.

Achievement of Mastery (>84%) will be very low throughout in New York State. There is a negative score adjustment in the chart between scores 100 and 65. For example, if a student earned 93% of the possible points, he/she will get a 90. If a student earned 85% of possible points, he/she receives a scaled score of 80. There are two ways to earn a 64; but no way to earn a

99
97
95
92
89
or 86.

There were no reading comprehension questions – just like the issue with physics. There was one in the test sampler.

Conclusion

In conclusion, some of the new assessments, particularly at the senior high school, have improved dramatically. NYSCOSS is very appreciative of some gains in the quality of a number of the tests. However, as we have been saying for four years, there is too much testing, particularly at the middle grades and the tests are not of sufficient quality to warrant the tremendous removal of both teachers and students from instruction and to justify such a complicated testing process.

The issue continues to be that teachers are “jumping through many hoops” just to determine a youngster’s number. In the end it has the effect of lowering the expectations of youngsters as it relates to academics.