

**Global History and Geography Regents
Examination--Data and Information Related to
Standard Setting**

**A study performed for the New York State Education
Department by**

Gary Echternacht
Gary Echternacht, Inc.
4 State Park Drive
Titusville, NJ 08560
(609) 737-8187
garyecht@aol.com

Unedited Draft: March 17, 2000

GLOBAL HISTORY STANDARD SETTING STUDY

Introduction

The New York State Board of Regents has established learning standards that all students must meet to graduate from high school. One set of learning standards is for social studies. The standards pertain to:

- History of the United States and New York
- World History
- Geography
- Economics
- Civics, Citizenship, and Government

Key ideas, performance indicators, and sample tasks further describe each learning standard. They are also further broken down by educational level—elementary, intermediate, and commencement¹. To assess the extent that students have met the learning standards, the New York State Department of Education has developed a testing program. The content of the tests reflect accomplishment of the learning standards. For social studies, the State Education Department has developed a Global History and Geography Examination to reflect accomplishment of the world history and geography learning standards. Students must pass this test to graduate from high school.

Although scores for the test are placed on a numerical scale, essentially there are only three scores—fail, pass, and pass with distinction. The test items have been developed by New York state teachers using professionally established procedures and have been pilot tested and field tested on samples of students.

The purpose of the study described in this report is to obtain information that the State Education Department can use to establish scores that will classify test takers into fail, pass, and pass with distinction categories. Setting passing scores requires judgment. This study employs professionally established methods to quantify and summarize the judgement of experts related to how individuals who have met the learning standards will perform on the test.

The Global History and Geography Regents Examination

The Global History and Geography Examination is a three part examination that is administered in three one hour sessions. The test content is based on the commencement level performance indicators found in Learning Standards for Social Studies, developed and adopted by Board of Regents. The three parts of the examination are as follows:

- Part I consists of 50 multiple choice questions, with the questions from each historical era and social studies standard. The multiple choice part of the test is intended to comprise 55% of the total test score.

¹ Learning Standards for Social Studies, new York State Education Department, June 1996.

GLOBAL HISTORY STANDARD SETTING STUDY

- Part II consists of a thematic essay and is intended to contribute 15% of the total test score. The thematic essay requires students to write in depth about a major theme in the global history and geography section of the Social Studies Resource Guide with Core Curriculum. In the thematic essay, students are asked to compare and contrast events, analyze issues, or evaluate solutions to problems in a comprehensive and cohesive essay that includes a clearly articulated introduction statement and a logically drawn conclusion.
- Part III consists of document based essay questions and intended to contribute 30% of the total test score. The document based essay questions require students to identify and explore multiple perspectives on events or issues by examining, analyzing, and evaluating textual and visual primary and secondary documents. The document based part of the examination has two item types. One type asks students to answer a question concerning a specific document. There 6 of these items. The other item is an analytic essay based on a document. Each of these item types is intended to contribute 15% of the total test score.

The multiple choice section of the examination is scored by counting the number of correct answers. Parts II and III of the examination are scored by trained teachers in their districts, following guidelines designed to produce reliable scores. Each of the extended essays is scored holistically using a 0-5 scale. Each short answer response (e.g., those in Part III) are scored on a 0-2 scale. The raw scores for each item type are weighted and added to make the total test score. The weights used are created to both reflect the intended weightings of the parts and to assure that different forms of the test are equivalent in difficulty.

A complete description of the examination, including test specifications and scoring rubrics is given in a test sampler.²

Methods Employed

Data related to the performance standards for the test was obtained from a committee of experts. Judgments from committee members were quantified using standard practices employed by psychometricians who conduct standard setting studies. The committee made their judgments with respect to one test form, which had been designated as the anchor test form. This form was designated as the anchor test because the score scale and passing standards were to be developed for this form. Subsequent forms of the test would be equated to the anchor test form so that all scores would be comparable over different test forms.

Committee members were given definitions of three performance categories—failing, passing, and passing with distinction. The State Education Department has developed these category definitions and they are applied to all of the Regents tests that are being developed. In addition, committee members were given an exercise designed to help familiarize themselves with the examination and an exercise where they were asked to categorize some of their students into the performance categories as defined by the State Education Department.

² Global History and Geography Regents Examination, Test Sampler Draft, New York State Education Department, June 1999.

GLOBAL HISTORY STANDARD SETTING STUDY

The committee met as a group on March 10, 2000 at the State Education Department.

Judgments for the 50 question multiple choice section of the test were obtained by a modified Angoff procedure. The Angoff procedure is the most commonly employed procedure for obtaining judgments related to passing scores. Committee members were asked to make two judgments for each of the 50 items:

- The committee member provided his or her best estimate of the percentage of students who would answer the question correct and who were on the cusp between having achieved the learning standards and not having achieved the learning standards, and
- The committee member provided his or her best estimate of the percentage of students who would answer the question correct and who were on the cusp between having achieved the learning standards and having achieved them at a level that demonstrated superior mastery of the learning standards.

To aide committee members in making their judgments, P-values (the percentage of students who answered the question correct in field testing) for each item were given to the committee members. Committee members worked independently in providing their judgments. Committee members provided the above judgments on each item before proceeding to the next item.

Once committee members had completed their work on the multiple choice section, the study director determined the respective passing scores (i.e., passing and passing with distinction) and reported the results back to the committee. Committee members could then change their estimates if they so chose.

The Angoff procedure is used to quantify judgments when the test items are scored right-wrong. The long essays (thematic essay and document based essay) are scored holistically using a 0-5 scale. Zero scores are given only if the essay is off topic. The six short answer document based questions are each scored with a 0, 1, or 2.

Committee members were given copies of essays that are used as marker papers in the scoring. Marker papers define the scores in that they are the best representation of their scores. Scorers should consistently score marker papers if they are scoring properly. Committee members were given two marker papers for each score point for each of the essays.

Committee members were then asked to judge which of three kinds of students was most likely to have written the marker paper. Committee members did not know the essay's score. They were asked to judge whether the paper was most likely written by a typically failing student, a typically passing student, or a typically passing with distinction student. Committee members recorded their judgments by recording a F, P, or D for each essay they reviewed

Committee members were also asked four overall questions about test performance, whose answers might aid the state in setting appropriate performance standards on the test. Those questions were:

GLOBAL HISTORY STANDARD SETTING STUDY

- Each committee member's estimate of the percentage of students in the state who have mastered the learning standards.
- Each committee member's estimate of the percentage of students in the state who have mastered the learning standards with distinction.
- Which was the more serious error—to pass a student who has not mastered the learning standards or to fail a student who has mastered the learning standards?
- Which was the more serious error—to pass with distinction a student who has not mastered the learning standards at that level or to fail to pass a student with distinction who had achieved that level of mastery.

Committee Members

The New York State Department of Education's Office of Assessment assembled a committee of 21 people to provide judgments for the study. Committee members were, with one exception, classroom teacher. One committee member was an experienced teacher, but was employed by the teacher's union. All committee members were recognized as very knowledgeable of the learning standards for social studies and how students perform on standardized tests similar to the Global History and Geography Examination. Some had worked on an aspect of either the standards or development of the tests.

Committee members, their schools, the number of years experience each has in teaching global history, and the number of global history students they are currently teaching are given in the table below.

Committee Member	Schools and Location	Years Teaching Global History	No. of Current Students
Dexter Alleyne	Maxwell Vocational High School Brooklyn	2	34
Gloribel Arvelo-Park	Wilson High School Rochester	10	49
Adrian Bordoni	John F. Kennedy High School Bronx	7	60
Jphn DeGuardi	Ballston Spa High School Ballston Spa	10	35
Scott Dolan	Morris High School Bronx	2	110
Judith DuPre	Fairport High School Fairport	29	85
Margaret Durant	Corcoran High School Syracuse	2	20
Maria Gallo	Harry S. Truman High School Bronx	10	60

GLOBAL HISTORY STANDARD SETTING STUDY

Committee Member	Schools and Location	Years Teaching Global History	No. of Current Students
Miguel Garcia	De Witt Clinton High School Bronx	6	95
Audrey Goropeuschek	Valley Stream North High School Franklin Square	7	160
Preya Krishna-Kennedy	Bethlehem High School Delmar	4	55
Nancy Maguire	Cornwall Central High School Cornwall	25	79
Victoria Milne	Liverpool High School Liverpool	14	130
John Orzel	Whitney Point Central High School Whitney Point	34	25
Joseph Palya	Charles E. Gorton High School Yonkers	32	135
William Russo	Leonardo da Vinci High School Buffalo	8	75
Elizabeth Sheffer	New York State United Teachers Albany	19	0
Kenneth Siepes	Shaker High School Colonie	32	53
Brendalon Staton	Hempstead High School Hempstead	4	87
Margo Ulmer	Naples High School Naples	23	92
Andrew Zawacki	Loudonville Christian School Loudenville	5	31

Committee members were chosen so that they would represent a wide range of schools and different types of students. Each committee member was asked to complete a short background questionnaire that included questions about their sex, ethnic background, the setting for their school and the percentage of students in their classes who were American-Indian, African-American, and Hispanic. Results of the questionnaire tabulations are given in the table below.

GLOBAL HISTORY STANDARD SETTING STUDY

Characteristic	Percent of committee
Sex	
Male	48%
Female	52%
Ethnic Background of Committee Member	
African-American	10%
Hispanic	24%
White	62%
Other	5%
School Setting	
Urban	45%
Suburban	40%
Rural	15%
Percent of students in committee member's classes who are American Indian, African-American, or Hispanic	
More than 50%	50%
10%-50%	5%
Less than 10%	45%

Findings Related to the Multiple choice Section

Committee members record their judgments in terms of a percentage. In the Angoff procedure, the percentages are summed. This yields a passing score for the individual committee member. In this study, committee members were making two judgments for each item. The resulting committee averages and standard deviations for the two points are given in the table below.

Standard	Committee Average	Standard Deviation
Passing	20.3	3.7
Passing with distinction	38.5	3.2

There are 50 items in the multiple choice section of the test. The averages and standard deviations in the table are in terms of raw score points.

GLOBAL HISTORY STANDARD SETTING STUDY

Typically, one takes these points and others in the neighborhood of these points and finds the percentage of students who will likely fail to meet the passing standard or who will pass with distinction. That was not possible in this study because field testing was not performed on the test on a whole. Thus, there are no score distributions for the multiple choice section when given as a whole.

An estimate of the score distribution for the whole multiple choice section was obtained by adding the P-values for the items and using that sum as the average for the score distribution. The standard deviation for the score distribution was derived by making assumptions about the KR-20 reliability of the section and deriving the standard deviation. An additional assumption was made that students will perform better on operational testing than on field testing. It was assumed that students would perform on average about 3 raw score points better on operational testing than on field testing. This factor was then added into the score distribution.

By making the further assumption of a normally distributed score distribution, an exact form of score distribution was obtained. The purpose of this distribution is to provide some guidance to the state with respect to the nature and quality of the committee judgments. Passing scores may or may not be set at the committee average, but they generally are set in the neighborhood of those averages.

The estimated percentage of students who would fail the test, assuming only the multiple choice section were given, is presented for various passing points in the table below.

Passing Score	Estimated % Failing
24.5	17%
22.5	12%
20.5	8%
18.5	5%
16.5	4%

In the table, the committee average is highlighted in bold.

Typically, passing scores are set at a score level that is in between two scores to avoid confusion about whether a specific passing score is passing or failing. Recall also, that the percentages in the table are obtained from the estimated total multiple choice score distribution. Key assumptions were made about the reliability and amount of improvement students would make from field testing to operational testing in deriving that score distribution. Caution is urged in over interpreting the accuracy of that distribution.

GLOBAL HISTORY STANDARD SETTING STUDY

The estimated percentage of students who would pass with distinction is given in the table below. In this table that percentage is reported for a number of possible

Passing With Distinction Score	Estimated % Passing With Distinction
41.5	11%
40.5	14%
39.5	17%
38.5	20%
37.5	23%
36.5	27%
35.5	31%

The committee average is again highlighted in bold. Similar caution is urged in interpreting the percentages reported.

Findings Related to the Essay Sections

Essay passing scores are derived for each individual essay and committee member from the pattern of classifications made by the committee member. The goal is to deduce passing scores that will best reproduce the committee member's pattern of classifications. Decision rules are used to deduce the passing points. The decision rules used to determine a passing point for an individual on an individual essay were as follows:

1. Find the lowest essay score that is classified as passing. That score is the base number.
2. If the other essay at the same score is failing, add 0.5 to the base.
3. If there is an essay with a higher score that the committee member has classified as failing, add 0.5 to the base.

The decision rules for the passing with distinction point for an individual on an individual essay were similar and as follows:

1. Find the highest essay score that is classified as passing. Add one to that score, which becomes the base.
2. If the other essay at the same score is passing with distinction, subtract 0.5 from the base.
3. If there is an essay with a lower score that the committee member has classified as passing with distinction, subtract 0.5 from the base.

GLOBAL HISTORY STANDARD SETTING STUDY

In the document based short answer section of the test, there are six questions, each scored either as 0, 1, or 2. Scores on this part of the test may range from 0 to 12. The above decision rules were applied to each essay and individual. The six derived passing points were added for each individual. The average committee passing points and standard deviations are given in the table below.

Standard	Committee Average	Standard Deviation
Passing	3.8	1.4
Passing with distinction	10.9	.9

Entries in the table are in terms of raw score points for the section (12 possible raw score points).

As is the case with the multiple choice section, there are no total score distributions available from the field testing. Adding the score averages over the six essays, however, results in an expected average score of about 7.5. The distributions of scores³ over the six essays from the field test is given in the table below:

Essay Number	Essay score			Average
	0	1	2	
1	.14	.63	.23	1.1
2	.06	.61	.30	1.2
3	.12	.55	.27	1.1
4	.15	.11	.65	1.4
5	.03	.05	.83	1.7
6	.15	.55	.20	1.0
Total				7.5

Based on the table above, it seems reasonable to assume that a passing score of 3.5 (approximately the committee average) would result in somewhere between 5%-15% failing this section of the test. In choosing the committee's passing with distinction point of 10.9 would result in somewhere between 5%-20% of students passing with distinction.

The two longer essays are scored holistically using a 0-5 scale. Committee averages and standard deviations for the two passing points and the two essays are given in the table below. The passing points are expressed in terms of raw score points on the 0-5 scale.

³ The rows do not add to 1 because those students who did not answer an essay are not included in the table.

GLOBAL HISTORY STANDARD SETTING STUDY

Essay and Standard	Committee Average	Standard Deviation
Document based essay		
Passing	1.8	.5
Passing with distinction	4.1	.7
Thematic essay		
Passing	2.4	.7
Passing with distinction	4.6	.4

Unlike the other test sections, there are score distributions from the field testing for each of the longer essays. Those score distributions are given in the table below.

Essay Score	Document Based Essay	Thematic Essay
0	.05	.05
1	.16	.11
2	.25	.35
3	.25	.27
4	.10	.13
5	.03	.06
Average	2.0	2.5

If the respective passing points for the longer essays (1.8 and 2.4) are applied to the two essays, it appears that about 20%-30% would fail the document based essay and about 50% would fail the thematic essay. If the respective passing with distinction points for the longer essays were applied (4.1 and 4.6), it appears that about 3% would pass the document based essay with distinction and about 6% would pass the thematic essay with distinction.

Other judgments obtained

Committee members were asked to provide their best judgment of the percentage on students in the state who would not achieve at least a proficient level of performance on the learning standards as well as the percentage of students in the state who would achieve competence in the learning standards at a distinction level. The committee averages and standard deviations are presented in the table below.

Standard	Committee Average	Standard Deviation
% not proficient	33%	11.2%

GLOBAL HISTORY STANDARD SETTING STUDY

% proficient with distinction	17%	10.7%
-------------------------------	-----	-------

The data in the table above relates to the passing scores for the test in that the committee on average was indicating that in their judgment almost one third of the students in the state were not achieving at the minimum level suggested by the learning standards. This assessment was made without test scores and is independent of the test scores. Similarly, the committee on average that one-sixth or about 17% of students were achieving at the distinguished level.

GLOBAL HISTORY STANDARD SETTING STUDY

Any time there is a classification of people made based on a test score, errors in classification occur. Errors in classification cannot be eliminated, but they can be managed. There are two errors that occur:

- Passing a student who should fail (i.e., the student's true level of achievement is below that required for passing), and
- Failing a student who should pass (i.e., the student's true level of achievement is above that required for passing).

With respect to the passing standard, committee members overwhelmingly (86% vs. 14%) indicated that it was a more serious error to fail a student who should pass. This type of error can be minimized by lowering the passing score. As the passing score is lowered, this error is decreased while the other type of error increases.

With respect to the passing with distinction standard, committee members were almost equally divided (43% vs. 57%) about which type of error was the more serious error. This suggests that raising or lowering the passing with distinction point to adjust for the two errors is not warranted.

Consistency of committee Judgments

It is not uncommon to find that committee members report varying judgments about student performance on assessments. This variation or inconsistency is due to two factors:

- Unfamiliarity with the judgment tasks they are asked perform, and
- Sincere differences standards committee members have internalized.

For classroom teachers, it is rare that they are asked to make quantifiable judgments of how hypothetical students will perform on a test as they are required to do in making Angoff judgments and in sorting essays. Also, standards themselves are influenced by the teacher's own experience. Teachers who have for years worked only with high achieving students will have different standards of performance than will teachers who for years have only worked with lower achieving students.

This inconsistency in resulting judgments is neither good nor bad. If the committee members are all very consistent in the judgments they provide, it does, however, suggest that statistics like the committee average should be adopted as the passing standard. When there is inconsistency, it suggests that the ultimate standard setting authority exercise flexibility and consider several points of view with respect to the final standard.

GLOBAL HISTORY STANDARD SETTING STUDY

To assess the consistency with which committee members provided their judgments on the multiple choice section of the test, and analysis of variance was performed on each set of judgments (those for passing and those for passing with distinction) to estimate the variance components and a coefficient of consistency. This coefficient of consistency is similar to a reliability coefficient for a test. This coefficient of consistency is the proportion of total variation that is accounted for by the variation in people and items.

The coefficients of consistency for the passing and passing with distinction data were .55 and .42 respectively. These coefficients are not extremely high, especially for the passing with distinction standard. The coefficients indicate that committee members varied considerably in the Angoff estimates that they made and suggests that the state act cautiously in interpreting the resulting judgments.

With respect to the essay judgments, the number of F, P, and D judgments were counted for each essay. From those counts, the percentage of committee members who made the most common judgment was calculated. The size of this number indicates the consistency with which the essay was judged. If every committee member agreed, then the percentage would be 100%. If there were maximum disagreement, then an equal number of committee members would have judged and essay as F, P, and D and the resulting percentage would be 33%.

Percentages obtained for the 36 short answer document based essays appear in the table below.

% of committee in majority	Frequency
90% +	6
80%-90%	5
70%-80%	10
60%-70%	6
50%-60%	8
< 50%	1

The data in the table suggests that committee members were reasonably consistent in their judgments, though the consistency was not overwhelming.

Data for the longer essays is given in the table below. The table indicates for each essay, the score of the essay, the number of committee members who judged the essay as failing (F), passing (P), or passing with distinction (D), and the percentage making up the majority of judgments.

The table suggests inconsistency at about the same level as that for the shorter essays, particularly for the document based essay.

GLOBAL HISTORY STANDARD SETTING STUDY

Essay	Essay Score	F	P	D	% Majority
Document 1	1	19	2	0	90%
Document 2	1	9	12	0	57%
Document 3	2	3	17	1	81%
Document 4	2	6	13	2	62%
Document 5	3	0	9	12	57%
Document 6	3	0	20	1	95%
Document 7	4	0	11	10	52%
Document 8	4	0	11	10	52%
Document 9	5	0	10	11	52%
Document 10	5	0	4	17	81%
Thematic 1	1	20	1	0	95%
Thematic 2	1	16	5	0	76%
Thematic 3	2	11	10	0	52%
Thematic 4	2	8	13	0	62%
Thematic 5	3	3	18	0	86%
Thematic 6	3	1	19	1	90%
Thematic 7	4	0	14	7	67%
Thematic 8	4	1	15	5	71%
Thematic 9	5	0	3	18	86%
Thematic 10	5	0	3	18	86%

Discussion and Recommendations

The purpose of this study is to obtain data and information that the state may use in setting passing points for the Global History and Geography Examination. The data should be used to guide those decisions.

The committee that provided the data was diverse and well represented the diversity of students, teachers, and school districts. With that diversity, it is not surprising that committee judgments varied and were somewhat inconsistent.

The overall committee averages (in raw scores) and the estimated percentage of students who would fail the passing standard or exceed the passing with distinction standard is given in the table below.

GLOBAL HISTORY STANDARD SETTING STUDY

Test Section	Passing		Passing with Distinction	
	Committee Average	% Failing	Committee Average	% Achieving
Multiple Choice	20.3	8%	38.5	20%
Document Based--Short Answer	3.8	5%-15%	10.9	5%-20%
Document Based--Essay	1.8	20%-30%	4.1	3%
Thematic Essay	2.4	50%	4.6	6%

In independent judgments, the committee estimated that statewide about 33% are failing to meet the learning standards and about 17% are achieving those standards with distinction. When making judgments about the seriousness of errors of classification, the overwhelming majority of committee members believed that failing to pass a student who should pass is more serious than passing a student who should fail. Committee members were evenly divided on the severity of errors when it came to passing with distinction.

Without total test score distributions, it is impossible to accurately estimate the impact of establishing specific passing points for the total test score. In fact, as this report is written, there is as yet no algorithm for obtaining a total test score. Before implementing any passing score points, the study author urges the state to obtain at least estimated total score distributions to determine the impact of whatever passing points are proposed.

The final passing points will be a composite of section passing points. As a composite, students will be able to compensate for poor performance on one section of the test with better performance on another section of a test. Generally, more people pass the test than pass any one section of the test when the total score is a composite.

With respect to the passing with distinction point, it appears that the committee averages provide reasonable section passing scores. A composite of those passing points employing the same weights as the total score composite should provide a passing point that is consistent with the committee expectations.

The simple passing point is of greater concern. The committee overwhelmingly indicated that failing to pass a student who should pass is a serious error. This argues for establishing a passing score that is somewhat lower than might be set had the committee been more indifferent about the errors in classification. The problem area is the thematic essay, where about half of students score 0, 1, or 2 all of which are below the passing point of 2.4.

Examination of the scoring rubric for the thematic essay tends to confirm the appropriateness of the committee average. The rubric for a score of 3 (lowest passing) is

- May make only vague reference to time and place.
- Addresses most aspects of the task or addresses all aspects in a limited way.

GLOBAL HISTORY STANDARD SETTING STUDY

- May contain minimal factual errors.
- Analyzes issues and events but not in any depth.
- Writes a satisfactorily developed essay, demonstrating a general plan of organization.
- Includes an introduction and/or conclusion.

The rubric for a score of 2 (highest failing score) is:

- Attempts to address the theme, but uses vague and/or inaccurate information
- Fails to address all aspects of the task
- Weak or no reference to time and place
- Develops faulty analysis of belief systems
- May contain some inaccurate information
- Narrative goes off on tangents; essay lacks focus
- Has vague or missing introduction and/or conclusion

In the judgment of the study author, the committee's average passing score does not seem unreasonable given the scoring rubric, the compensatory nature of the total score composite, and the assessment that about one third of students are not meeting the learning standards.

In the absence of total score distributions, the study author recommends that the state take the committee averages, place those in the appropriate composite weighting for total score, and use those for the initial operational year. The state might also consider subtracting some number of points from the composite passing score to account for the committee's belief in the seriousness of the error or failing a student who should pass. Once total score distributions are known, the state should review the appropriateness of the passing points, and make adjustments as needed.