

## Appendix A – Description and Illustrations of Test Similarity

Caveon’s test similarity statistic uses a nominal response Item Response Theory (IRT) model to assess the probability that an individual student will select the chosen responses. Using these probabilities, two test responses (i.e. from the actual set of answers to the test items) may be compared in order to test whether the two tests were taken independently. Extreme similarity between the test responses provides statistical evidence that the tests were not taken independently. This evidence forms the basis of Caveon’s test similarity analysis.

All possible pairs of tests within a school are compared against each other. The number of pair-wise comparisons varies greatly and can be extremely large. The test similarity statistic is the count of the number identical correct and identical incorrect responses. Under the independence assumption this multivariate statistic follows a mixed trinomial distribution. Since a large number of comparisons are made, the probability level of the statistic is transformed using the distribution of the maximum order statistic.

This information is illustrated in the following table taken from a pair of test takers for the TAKS Grade 5 Math test. The data is drawn from a school having twenty-five students in two classes. The similarity between the test responses spans both classes. 15 tests were found to be similar and 10 were not.

The item responses have been highlighted and formatted according to the following rules: If the answer is correct it is not highlighted and the answer is represented using a letter; Incorrect answers are highlighted in decreasing frequency order (within this set of tests) using gold, yellow and gray. The test score is highlighted in blue if the student met the TAKS standard. The test identifier (not derived from actual identifiers) is highlighted if the test is a member of a test similarity cluster. A cluster is a grouping of tests where each test is highly similar to at least one other test in the cluster.

The data illustrate two similar test clusters. One cluster is a pair of tests. The other is 13 tests. The most likely explanation for a pair of similar tests is answer copying. The other cluster of 13 similar tests spans two classrooms and the students appear to be organized. The most likely explanations are students who are text messaging the answers to each other or access to crib sheets with the actual test content. Other explanations seem less likely, but possible.

The actual responses in this data have been randomly reassigned in order to protect the integrity of the test’s answer key. The data have been split into two tables, Table A-1 and Table A-2, for printing purposes. The best visual comparison will occur by placing the two Tables side-by-side.

**Table A-1: Part One of Similar Tests Illustration**

Test ID	14	22	25	26	28	29	37	42	49	11	18
Score	2186	1769	2271	2016	1700	1790	2100	1905	2345	2116	2016
Matches	0	0	0	0	0	0	0	0	0	1	1
1	E	E	E	E	3	3	E	E	E	E	E
2	C	C	C	C	C	C	C	C	C	C	2
3	D	5	D	D	2	D	D	D	D	D	D
4	D	D	D	D	2	D	D	2	D	D	D
5	D	D	D	5	3	3	5	D	D	D	3
6	C	C	C	C	C	1	C	1	C	C	C
7	A	A	A	A	5	5	A	A	A	A	A
8	E	3	E	E	1	1	E	3	E	E	E
9	A	3	A	A	3	2	A	A	A	A	A
10	B	B	B	B	5	5	B	B	B	5	5
11	5	3	D	D	2	2	2	3	D	D	D
12	E	1	E	3	3	3	E	3	E	E	E
13	C	C	C	C	2	4	C	C	C	C	C
14	A	A	A	A	2	3	A	3	A	A	A
15	B	4	B	5	B	5	B	3	B	5	5
16	D	3	D	3	D	2	3	2	D	3	3
17	1	5	D	5	5	3	1	3	D	5	5
18	E	2	E	E	E	E	E	E	E	E	E
19	A	4	A	A	3	A	A	A	A	A	A
20	E	1	E	E	2	1	2	1	2	E	E
21	A	A	A	A	3	3	A	3	A	A	3
22	1	1	1	1	1	3	1	1	E	1	1
23	A	2	A	A	2	5	A	A	A	A	A
24	C	4	C	4	1	C	C	1	1	1	1
25	E	1	E	E	2	2	E	E	E	E	3
26	E	3	E	3	3	1	4	4	E	E	E
27	E	2	4	E	E	2	E	E	E	2	2
28	2	A	5	2	5	2	2	2	2	3	3
29	A	4	*	A	2	A	A	A	A	2	2
30	B	4	B	B	4	B	1	4	1	B	B
31	2	1	E	E	1	E	E	2	E	E	E
32	D	1	D	1	D	3	D	3	D	1	1
33	2	C	1	4	4	C	C	1	C	C	C
34	C	2	C	C	C	C	C	C	C	C	C
35	C	2	C	2	2	1	C	1	C	C	C
36	B	3	B	3	1	4	3	1	B	4	1
37	2	5	2	2	C	C	2	2	2	5	5
38	C	1	C	5	5	1	C	C	C	C	C
39	3	1	E	E	2	1	E	E	E	1	3
40	3	*	D	D	1	3	3	D	D	D	D
41	C	1	C	C	C	C	C	C	C	C	C
42	4	E	E	1	4	1	4	4	E	E	E
43	A	4	3	2	3	3	A	3	A	A	A
44	E	2	E	4	4	E	3	E	E	E	2

**Table A-2: Part Two of Similar Tests Illustration**

Test ID	39	19	40	12	32	15	43	30	13	44	38	41	33	47
Score	2271	2212	2271	2161	2212	2306	2240	2240	2116	2186	2186	2161	2271	2186
Matches	12	11	11	10	10	10	9	8	7	6	6	4	3	1
1	E	E	E	E	E	E	E	E	E	E	E	E	E	E
2	C	C	C	C	C	C	C	C	C	C	C	C	C	C
3	D	D	D	D	D	D	D	D	D	D	D	D	D	D
4	D	D	D	D	D	D	D	D	D	D	D	D	D	D
5	D	D	D	D	D	D	D	D	5	D	3	D	D	D
6	C	C	C	C	C	C	C	C	C	C	C	C	C	C
7	A	A	A	A	A	A	A	A	A	A	A	A	3	A
8	E	E	E	E	E	E	E	E	E	E	E	E	E	E
9	A	A	A	A	A	A	A	A	A	A	A	A	A	A
10	B	B	B	B	B	B	B	B	B	B	B	B	B	B
11	D	D	D	D	D	D	D	D	2	D	D	D	D	D
12	E	E	E	E	E	E	E	E	E	E	E	E	E	E
13	C	C	C	C	C	C	C	C	C	C	C	C	C	C
14	A	A	2	A	A	A	A	A	A	A	A	A	A	A
15	B	B	B	B	B	B	B	B	3	B	B	B	B	B
16	D	D	D	3	3	D	D	5	D	3	3	D	D	2
17	5	1	1	5	5	D	1	D	5	1	1	5	3	5
18	E	E	E	E	E	E	E	E	E	E	E	E	E	E
19	A	A	A	A	A	A	A	A	A	A	A	A	A	A
20	E	E	E	E	E	E	1	E	E	E	E	E	E	E
21	3	3	3	3	3	3	3	3	3	3	3	3	3	A
22	E	4	E	3	3	3	4	1	E	3	1	E	E	4
23	A	A	A	A	A	A	A	A	A	A	A	A	A	A
24	C	C	C	C	C	C	C	C	C	C	C	C	4	C
25	E	E	E	E	E	E	E	E	E	E	E	E	E	E
26	E	E	E	E	E	E	E	E	E	E	3	E	E	E
27	E	E	E	E	E	E	E	E	E	E	E	E	E	E
28	2	2	2	2	2	2	2	2	2	A	2	2	2	2
29	A	A	A	A	A	A	A	A	A	A	A	A	A	A
30	B	B	B	B	B	B	B	B	B	B	B	B	B	B
31	E	E	E	E	E	E	E	E	E	E	E	4	E	E
32	D	D	D	D	D	D	D	D	1	D	D	D	D	D
33	1	1	1	1	1	1	1	1	1	1	1	C	1	1
34	C	C	C	C	C	C	C	C	C	C	C	C	C	4
35	C	C	C	C	C	C	C	C	C	C	C	C	C	C
36	B	B	B	B	B	B	B	B	B	B	B	B	B	B
37	2	2	2	2	2	2	2	2	2	2	2	2	2	2
38	5	5	C	C	1	C	C	1	C	2	5	5	C	5
39	E	E	E	3	E	E	E	E	1	3	E	1	E	E
40	D	D	D	D	D	D	D	D	3	D	D	D	D	D
41	C	C	C	1	C	C	C	C	C	C	C	4	C	1
42	4	4	4	4	4	4	4	4	4	4	E	4	E	E
43	A	3	A	3	A	A	A	A	3	3	A	3	A	A
44	E	E	E	E	E	E	E	E	E	E	E	4	E	3

**Legend:**

The test identifiers that belong to tests in similar test clusters are highlighted with a different color to indicate the cluster. There are two clusters in this data, one with 2 tests and the other with 13 tests. The most frequent incorrect response is highlighted using gold. The second most frequent incorrect response is highlighted using yellow. The least frequent incorrect response is highlighted using gray.

If the student met or exceeded the TAKS standard the test score is highlighted in blue.

The reader should visually compare the sections of the table to verify the strong clumping of incorrect answer choices. Especial attention should be directed at the strong clumping of the correct answer choice with the right hand portion of the table (e.g., Table A-2) as compared with the left hand portion of the table (e.g., Table A-1, where the similar tests were not detected).

## Appendix B – Description and Illustration of Aberrance

The aberrance statistic used in Caveon’s Data Forensics estimates test-taking modalities. The word “modality” refers to the mode in which the test is answered by the student. Normally, a student would take the test in the single mode corresponding to his or her knowledge or proficiency. However, if the student gets help on some questions and not others, then the student is taking the test in more than one “mode.” The aberrance statistic is based in Item Response Theory, but since it does not attempt to measure lack of fit or goodness of fit it is technically not a person-fit statistic. Simulations done at Caveon show this statistic is a very powerful detector of test-taking modalities.

Technically, the aberrance indicator estimates whether a student is answering some test items at a much higher knowledge or proficiency level than other test items. Conceptually, if a student misses easy items but gets difficult items correct, the student is demonstrating two different test-taking modalities. When the student misses the easy items it appears as if the student has no or little knowledge. When the student answers the difficult items correctly it appears as if the student has high knowledge.

Bimodal test-taking is a symptom of several behaviors, some of which would be considered test fraud or cheating. The aberrance statistic determines whether a student is taking the test in a consistent manner within the knowledge framework of the exam. If teachers have focused on certain portions of the test material the students will show expertise in those areas but inability in other areas, the aberrance statistic will detect many of these situations. If teachers give “extra assistance” to students while they take the exam, again the aberrance statistic will be sensitive to this behavior. If students use crib notes that cover a portion of the exam material, they will exhibit bimodal test taking.

The plots below show how this statistic distinguishes bimodal test taking behavior. The statistic estimates a low and high ability for every student. Using the low and high ability level estimates, the statistic estimates the probability that the test item was answered at either the low or high ability level. When no bimodality is present the computer algorithms are unable to separate the low and high ability estimates. This is shown in the Non-Aberrance Illustration #1 (Figure B-1). Non-Aberrance Illustration #2 (Figure B-2) shows a test where some degree of bimodality is present, but it is not statistically significant.

Three series are depicted in the plots. The  $P(x|low)$ <sup>10</sup> and  $P(x|high)$  series plot the probability of selecting the chosen response given the low or high ability estimate. When the low and high theta (e.g. ability) estimates are very close, the probabilities will be quite close to each other. The final series,  $P(low|x)$ <sup>11</sup>, plots the probability that the item was answered as a low ability test taker. The item data is sorted using the values of

---

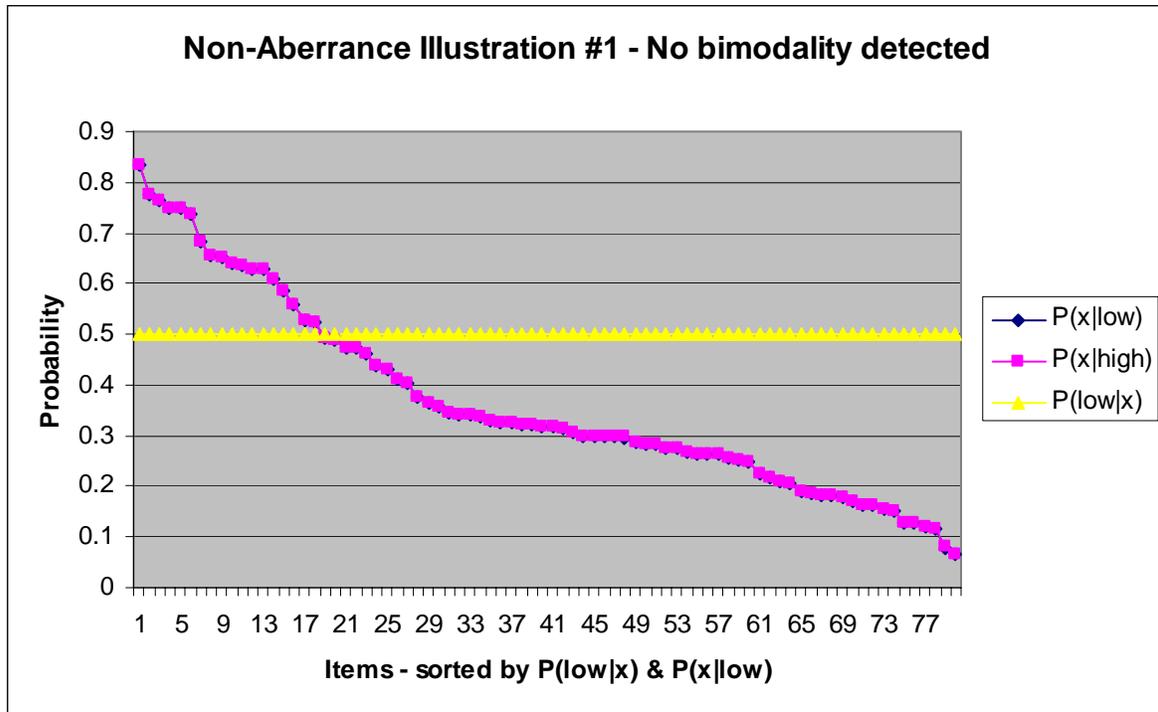
<sup>10</sup> The probability expression  $P(x|low)$  is read, “The probability of response x given that the student is answering in the low mode.”

<sup>11</sup> The probability expression  $P(low|x)$  is read, “The probability the student is answering in the low mode given the observed response is x.”

$P(\text{low}|x)$  to help visualize the differences as accentuated by the probability analysis. Without sorting the data, the series are very noisy and visually difficult to process.

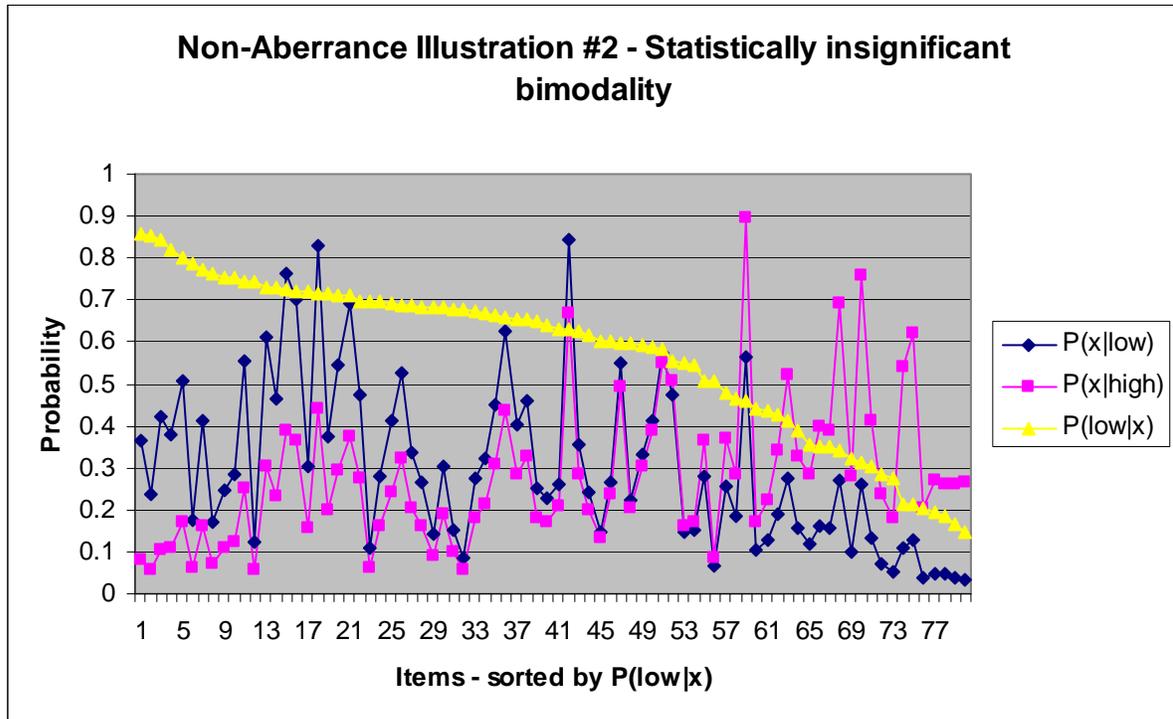
When bimodal test taking is present, a distinct separation is seen in the item set. The statistic then favors the high or low ability estimate as the mode of responding quite strongly. This is shown in the Aberrance Illustrations. As aberrance (or bimodality) increases, the separation becomes stronger.

**Figure B-1: Non-Aberrance Illustration #1**



Since no bimodality was present in the above data, the magenta line,  $P(x|\text{high})$ , is placed on top of the blue line,  $P(x|\text{low})$ . This is because the statistical algorithm could find no inconsistency that related to bimodality on this test. Consequently, the  $P(\text{low}|x)$  line is level at .5 (i.e., the high mode – low mode split is a 50-50 proposition).

Figure B-2: Non-Aberrance Illustration #2



In Figure B-2, bimodality was not detected and there is not significant separation between the probability lines for the high and low test-taking modes. The yellow line shows some discrimination across the items set for the low and high modes, but the degree of separation is within expected variation.

Figure B-3: Aberrance Illustration #1

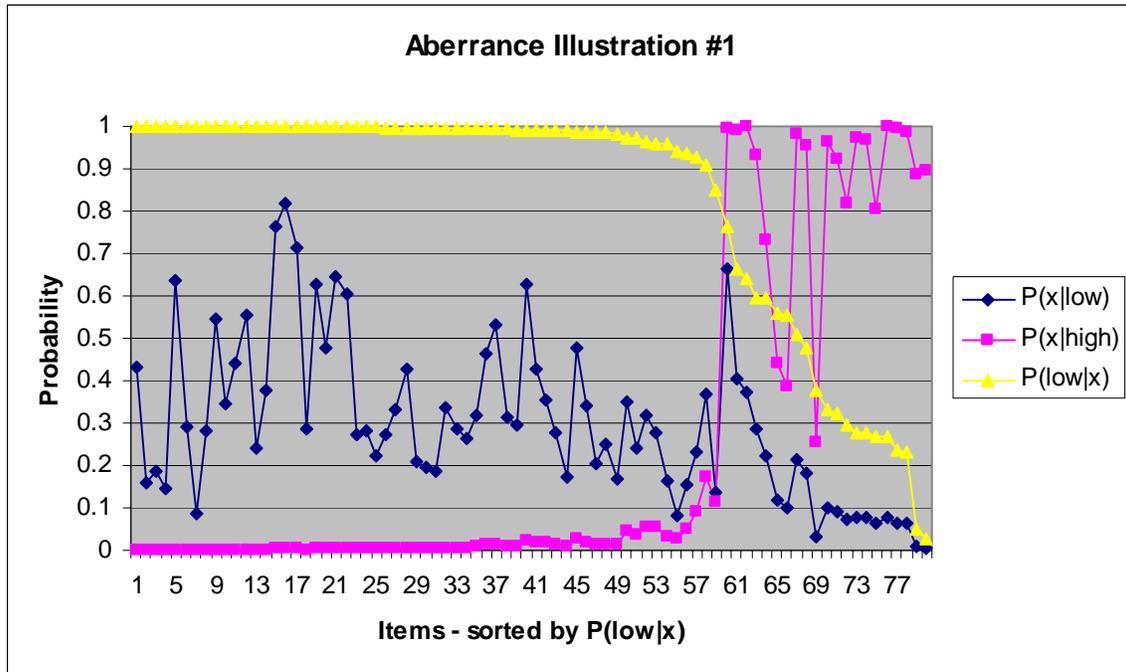
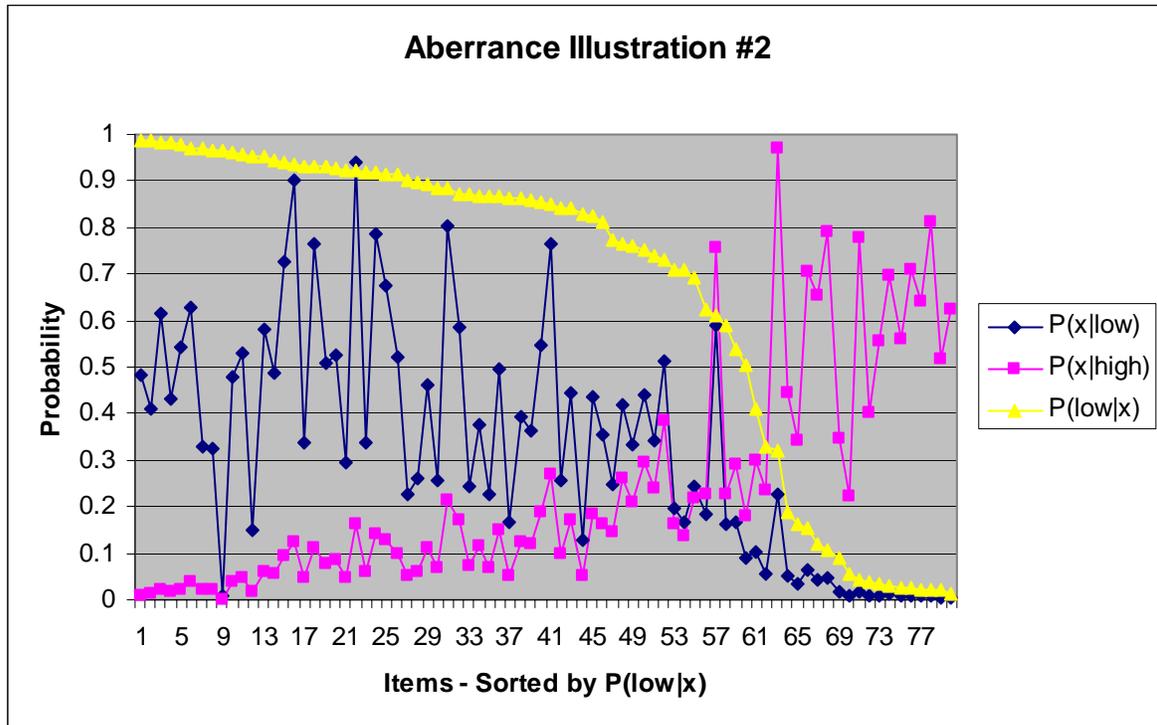


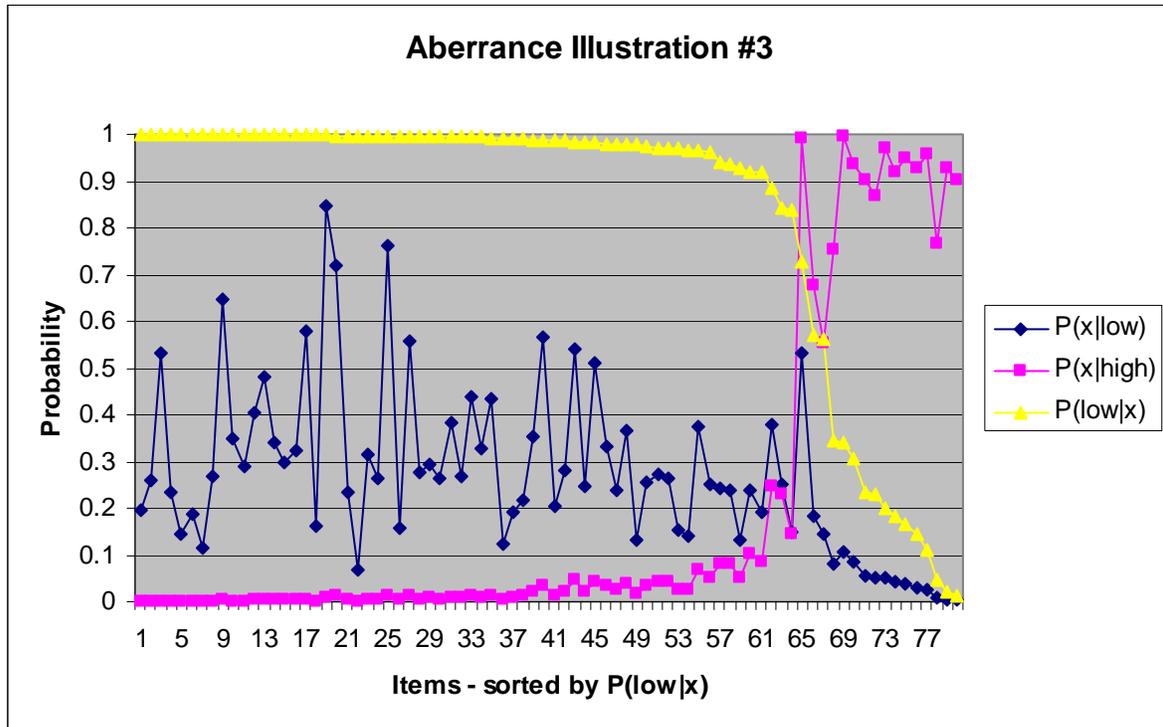
Figure B-3 above illustrates a very strong separation between the high and low test-taking modes. Low values of the pink line (e.g.,  $P(x|\text{high})$ ) indicate the student responded incorrectly to the question. High values indicate the student responded correctly to the question.

Figure B-4: Aberrance Illustration #2



In Figure B-4 above, the yellow line is not as sharply discriminating as the yellow line in Figure B-3. The degree of aberrance is less in Figure B-4 when compared to the aberrance in Figure B-3. This student's test is showing aberrance but the student is not a "high-performer."

Figure B-5: Aberrance Illustration #3

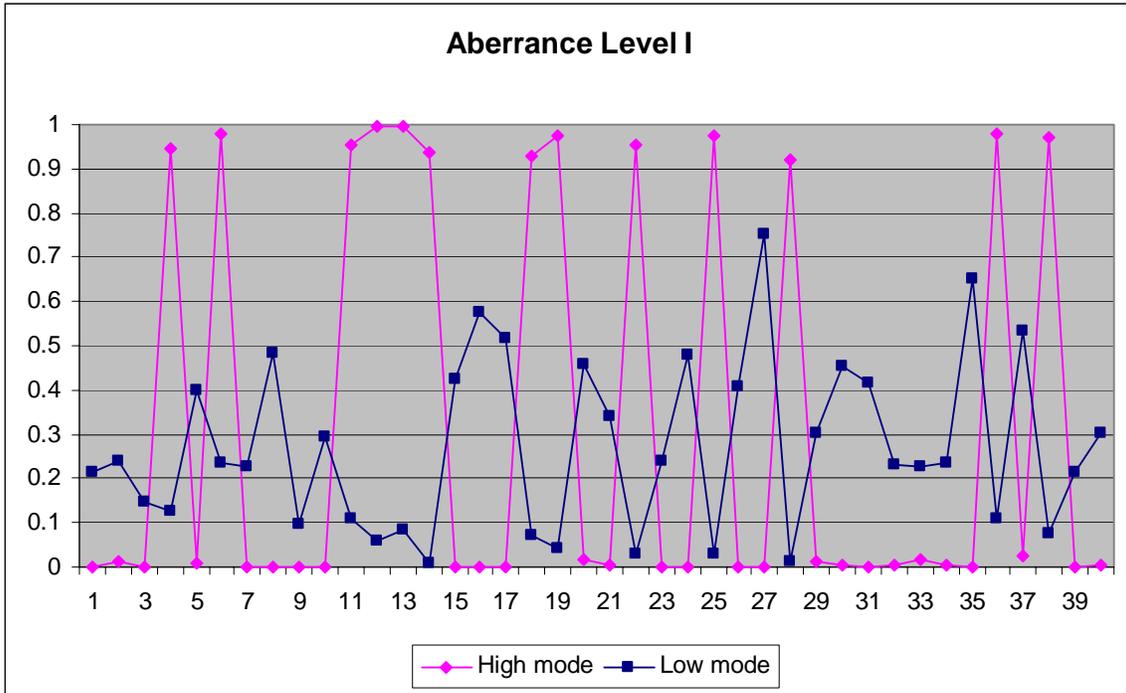


In the Figure B-5, the student has been able to answer some very difficult questions correctly, but has not shown that the responses to those difficult questions are representative of the student's actual proficiency on the exam.

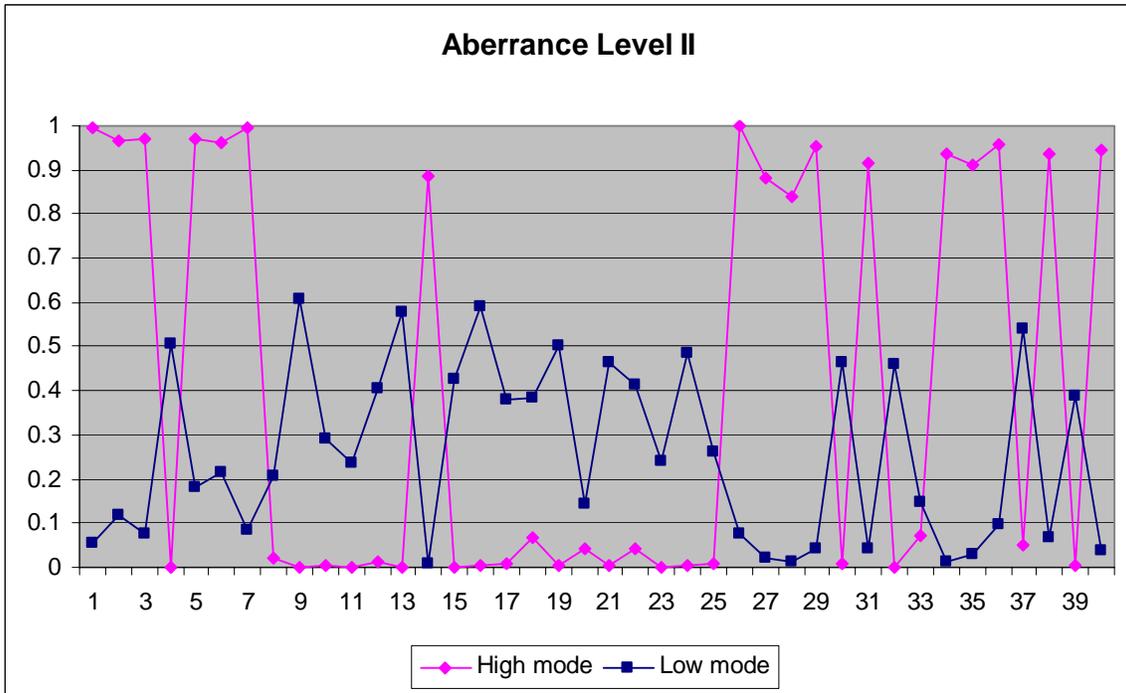
An additional perspective of aberrance can be gained by comparing data where the relative amount of low- versus high-mode performance varies. This has been done in Figures B-5 through B-10. Beginning in Figure B-5 the low-mode dominates and then through each figure there is a smaller amount of low-mode responding in favor of a larger amount of high-mode responding. These are labeled Levels I through VI for convenience.

The data for Figures B-5 through B-10 are from TAKS Reading Grade 5 tests. The item data are presented in the order in which the items are given on the test.

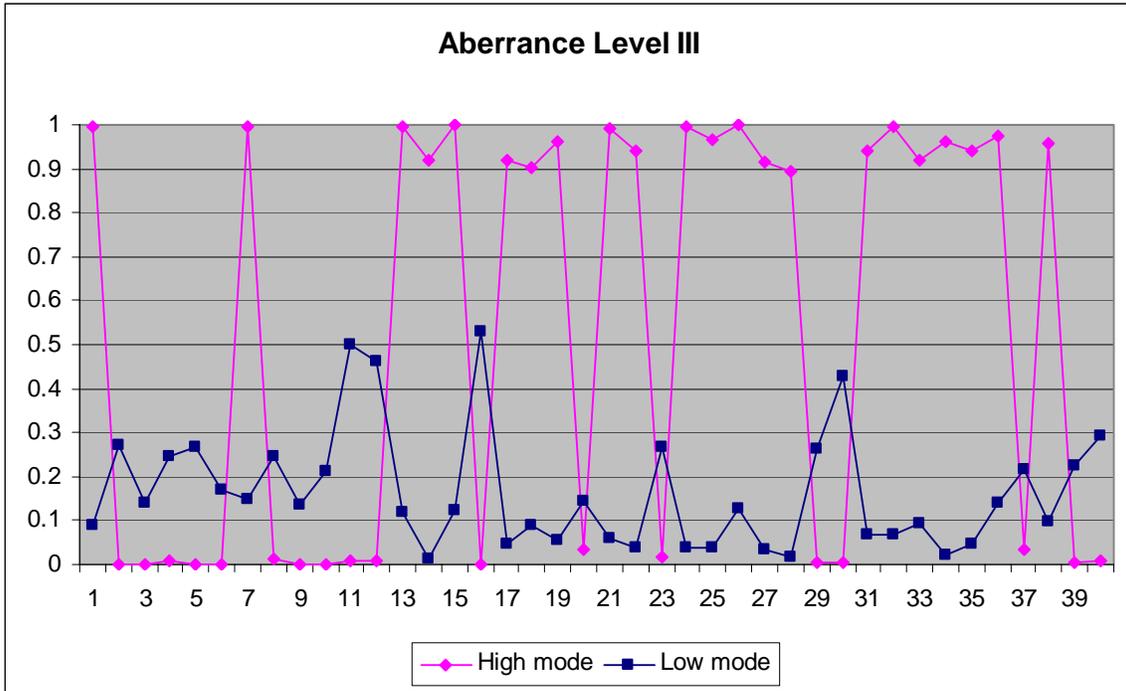
**Figure B-5: Aberrance Level I**



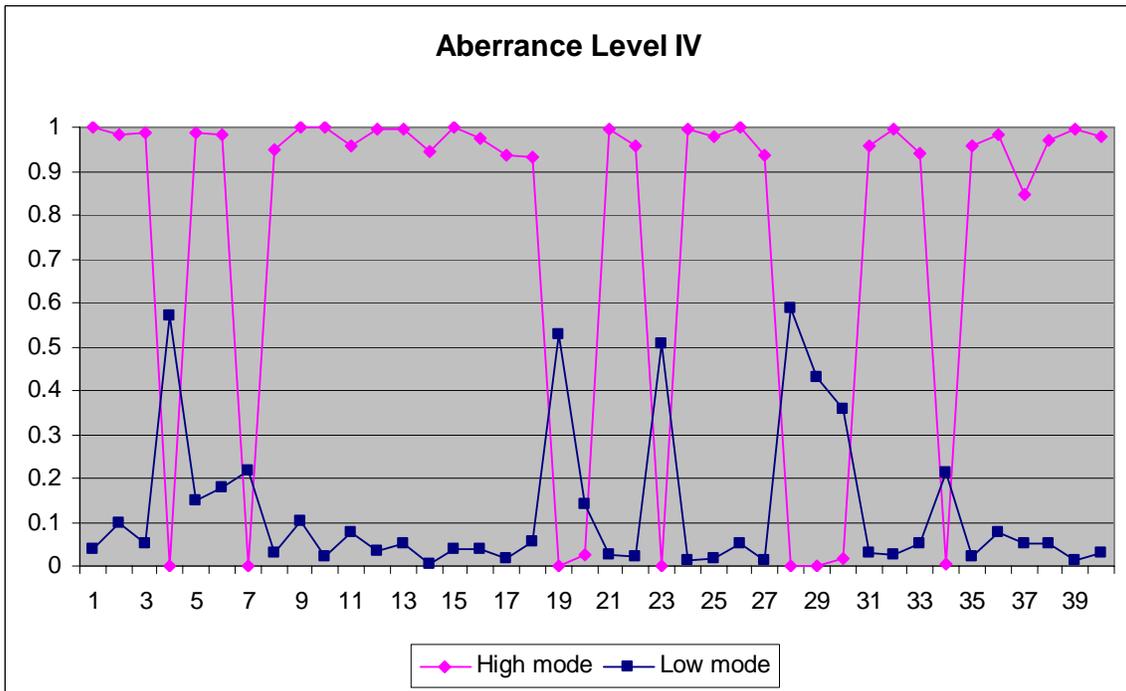
**Figure B-6: Aberrance Level II**



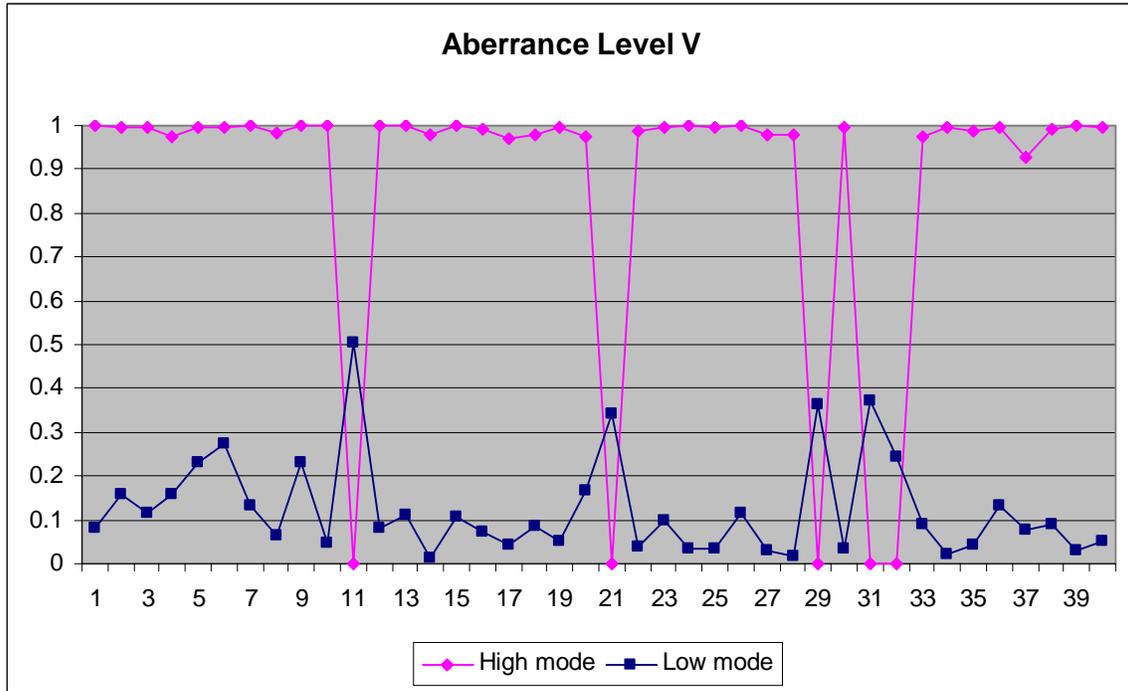
**Figure B-7: Aberrance Level III**



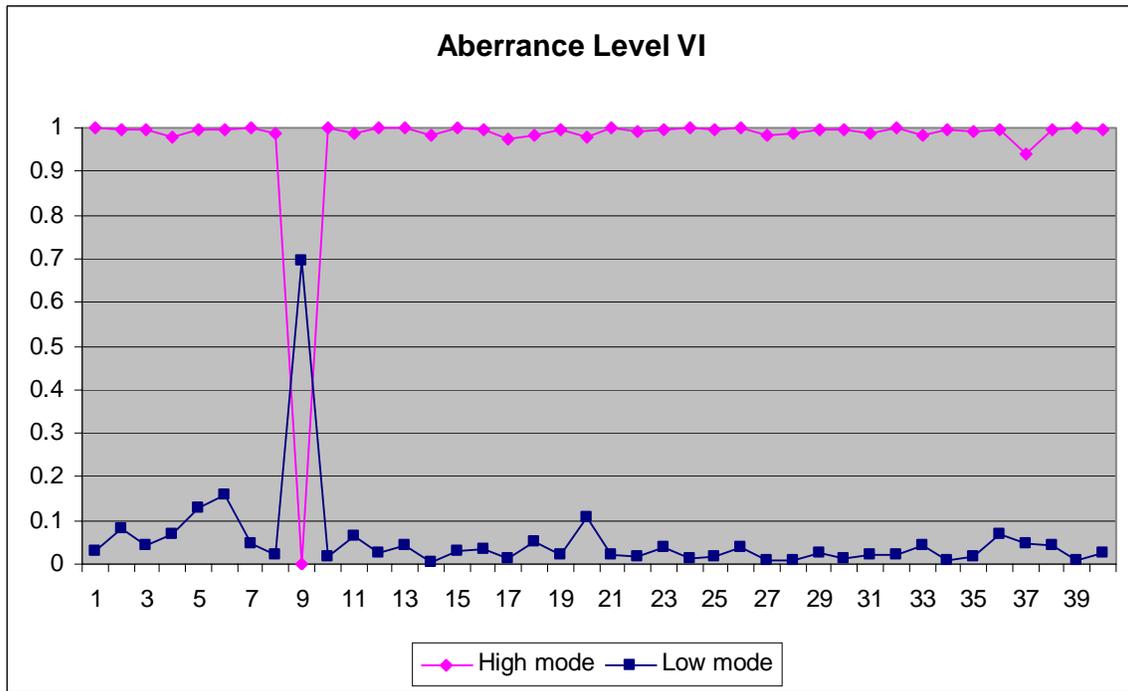
**Figure B-8: Aberrance Level IV**



**Figure B-9: Aberrance Level V**



**Figure B-10: Aberrance Level VI**



The different levels correspond to the different kinds of aberrance that could be seen in test taking. For example, Level I corresponds to the kind of aberrance that would be seen by a student who answered difficult questions by lucky guessing or using a partially

effective guessing strategy. At each level, the degree of low-mode proficiency lessens. At the highest level, Level VI, the student has only missed one question, but that incorrect answer choice was aberrant because the answer that was selected is very improbable given the student's demonstrated proficiency on the rest of the exam. In this case, we would say that the student had blundered. The intermediate levels can be ascribed to different degrees of content mastery possessed by the student.

Of course, guessing and blundering happen all the time (as do other behaviors that trigger aberrance). The important element of Data Forensics is to capture the normal level of these types of aberrance and then make inferences when excessive amounts of aberrance are seen that the aberrance is caused by something the students or teachers are specifically doing.

## Appendix C – Statistical Baseline

This appendix provides the statewide aberrance, similarity, multiple mark, and high gain score rates<sup>12</sup> and how those rates were obtained. The rates were computed for each grade and subject (Grade 10 ELA rates were combined across test forms). Tables C-1, C-2 and C-3 below list the statewide pass, aberrance, similarity, multiple mark, and high gain score rates<sup>13</sup> for the Spring 2005 TAKS Math, Reading/ELA, Science and Social Studies tests in this analysis.

This analysis relies heavily upon reliable baseline rates for the four statistical indicators. Using upper 95% critical value for the test statistics, statewide rates using these critical values were obtained. These rates are shown in Tables C-1, C-2, and C-3. Under assumptions of independence, the statewide rate is an estimate of the population proportion of administered tests for each statistic that exceeds the associated threshold value. This becomes the base rate for the analysis.

**Table C-1: Statewide percentage rates for TAKS Math, Spring 2005**

Grade	n <sup>14</sup>	Pass Rate	Aberrance Rate	Similarity Rate	Multiple Mark Rate	Gain Score Rate
3	274,481	81.9	5.7	1.2	1.6	
4	277,700	81.2	5.5	1.4	2.6	5.7
5	280,257	79.2	5.7	1.5	3.6	6.1
6	289,510	72.0	4.6	2.2	3.0	5.9
7	293,432	63.8	3.8	3.6	3.9	5.5
8	290,359	60.7	4.4	3.8	1.2	5.3
9	316,564	56.3	3.6	4.6	1.0	5.1
10	264,603	58.4	3.9	4.4	1.0	5.1
11	225,984	81.0	4.5	6.1	1.5	5.5

**Table C-2: Statewide percentage rates for TAKS Reading/ELA, Spring 2005**

Grade	n	Pass Rate	Aberrance Rate	Similarity Rate	Multiple Mark Rate	Gain Score Rate
3	269,398	89.0	6.6	0.9	1.2	
4	272,913	79.4	5.7	1.2	2.3	7.2
5	276,261	75.1	6.1	1.5	3.9	5.8
6	287,940	85.2	6.0	1.9	2.7	5.7
7	292,922	80.9	6.2	2.5	3.9	5.5

<sup>12</sup> These rates measure the percent of tests where aberrance, similarity, excessive multiple marks and high gain scores was measured. These four statistical indicators are designed to detect testing irregularities, but they do not directly measure testing irregularities. The rate should not be interpreted as an indication of the amount of cheating that is actually present.

<sup>13</sup> Gain scores were not available for computation in grade 3. Gain scores for grade 4 used the prior year as covariates. Gain scores for all other grades used two prior years if the data were available. If the data were not available gain scores were not computed. Gain scores for Science and Social Studies in grade 11 were based upon one prior test score.

<sup>14</sup> The numbers of tests administrations are slightly lower than previously reported numbers. This is due to the discarding of between .1 and .2% of the tests having excessive numbers of unanswered questions. Unanswered questions do not seem to be a source of testing irregularity. Most of these unanswered questions corresponded to blank answer sheets.

8	291,448	83.1	6.3	2.2	2.1	6.1
9	321,602	82.0	7.6	1.6	2.7	5.0
10	269,709	67.2	10.1	26.8	1.1	4.9
11	229,228	87.9	9.7	25.1	0.8	5.0

**Table C-3: Statewide percentage rates for TAKS Grade 11, Spring 2005**

Subject	n	Pass Rate	Aberrance Rate	Similarity Rate	Multiple Mark Rate	Gain Score Rate
Math	225,984	81.0	4.5	6.1	1.5	5.5
Reading/ELA	229,228	87.9	9.7	25.1	0.8	5.0
Science	227,412	80.8	4.4	8.2	3.3	5.3
Social Studies	229,574	94.5	5.6	6.0	1.5	5.6

The thresholds for determining aberrance were determined by simulating 50,000 tests for each form and then using the upper 95% value from the distribution.

The thresholds for determining high similarity between test responses were set using the theoretical distribution of the number of identical correct and number of identical incorrect responses when two tests are compared and assuming that they are statistically independent. Since this is a multivariate test, an ordering rule was imposed in order to obtain consistency and maintain the type I error level of the hypothesis test. The upper 95% value from the distribution was used. The empirical results for all the tests except ELA grades 10 and 11 as compared to the actual TAKS data are close to these values. The ELA grade 10 and 11 similarity rates are higher than the nominal 5% level. This is probably due to the design of these tests that create dependence among the test items.

The thresholds for determining excessive multiple marks (i.e., wrong-to-right answer changes) were set using the theoretical distribution of wrong-to-right answer changes, other answer changes, and no answer changes, as specified by statewide rates on each form. A multivariate ordering rule for evaluation of the tail area was used in order to ensure consistency in probability interpretation. The upper 95% value from the distribution was used. The empirical distribution appears to be somewhat tighter than the theoretical distribution. This is most likely due to the assumption of that multiple mark rates are common across all items (This assumption was required because the multiple mark data was provided as counts for each test). The uni-directional test of the multivariate distribution protects against inferring that excessive wrong-to-right answer changes occur when a student changes a lot of answers (many of which are right-to-wrong answer changes).

The thresholds for determining unusual gain scores were set using the standardized residuals and then assuming the residuals are normally distributed. This is known to be a poor assumption in practice and a t-distribution is preferred. However, the degrees of freedom of the regression are so many because the regressions are computed for all tests in the state that there is no difference between the t-distribution and the normal

distribution. Table C-4 provides the number of tests that were analyzed and the link rates<sup>15</sup> for the gain score computations.

**Table C-4: Linkages for Gain Score Models**

Subject	Grade	Tests	Linked Tests	% Linkage
Math	3	274,481		
Math	4	277,700	246,413	88.7
Math	5	280,257	225,353	80.4
Math	6	289,510	236,612	81.7
Math	7	293,432	246,145	83.9
Math	8	290,359	246,988	85.1
Math	9	316,564	258,413	81.6
Math	10	264,603	221,770	83.8
Math	11	225,984	200,649	88.8
Reading/ELA	3	269,398		
Reading/ELA	4	272,913	243,766	89.3
Reading/ELA	5	276,261	225,201	81.5
Reading/ELA	6	287,940	234,075	81.3
Reading/ELA	7	292,922	244,255	83.4
Reading/ELA	8	291,448	246,832	84.7
Reading/ELA	9	321,602	261,552	81.3
Reading/ELA	10	269,709	224,038	83.1
Reading/ELA	11	229,228	200,040	87.3
Science	11	227,412	206,675	90.9
Social Studies	11	229,574	208,233	90.7

<sup>15</sup> No gain scores were available for Grade 3 test scores. A one year model was used for the Grade 4 tests. A one year model was used for the Grade 11 Science and Social Studies tests. Two year models were used for Math and Reading/ELA Grades 5 through 11.

## Appendix D – Counting Exceptions

An exception is detected for a classroom or school whenever the statistical indicators (i.e., aberrance, similarity, multiple marks, and unusual gains), combined together to form the overall statistical index, are statistically anomalous. In order to control for environmental and endemic effects the statewide aberrance, similarity, multiple mark and unusual gain score rates were used as the baseline rate for each test (See Appendix C for these rates).

In order to determine if the values of the statistical indicators for a classroom or school are anomalous the probability of observed value or large is transformed into a negative logarithm. It is well known that these transformed values are approximately distributed as chi-square random variables with 2 degrees of freedom (The distribution is approximate because the binomial distribution is discrete). Under assumptions of independence the sum of these chi-square variables is also distributed as a chi-square variable with degrees of freedom equal to the sum of the degrees of freedom of the random variables in the sum.

The extreme value distribution of the resulting chi-square variable is used as an outlier test on observation (specifically extremeness from the hypothesized distribution). This distribution controls the Type I error rate (more commonly known as alpha) at the experiment level for all the statistical tests of hypotheses. The Type I error rate is set so that by chance alone, only 1% of the time would one observation be rejected from among the entire set of observations that are being tested.

### ***Discussion of the Extreme Value Distribution***

Control of alpha at the experiment-wide level is achieved using the extreme value distribution (i.e., distribution of the maximum order statistic). The nominal level of the extreme value distribution was set at .01. This level is chosen such that the maximum order statistic would only exceed the associated critical value 1 time in 100 when the null hypothesis is true.

Technically, the distribution function of the maximal order statistic<sup>16</sup> is used:

$$F_n(y_n) = [F(y_n)]^n$$

If the probability of the maximal order statistic is .01 or less of exceeding the observed statistic, the critical value can be expressed in terms of the original distribution function, as shown algebraically below:

---

<sup>16</sup> Hogg, R. V. and Craig, A. T. Introduction to Mathematical Statistics, Fourth Edition. Macmillian Publishing Co. (1978), pp. 154-161, Section 4.6, "Distributions of Order Statistics."

$$P(y_n > c) = 1 - F_n(c) = [1 - F(c)]^n = 1 - \alpha = .99$$

$$\ln(1 - F(c)) = \frac{\ln(1 - \alpha)}{n}$$

$$1 - F(c) = \exp\left\{\frac{\ln(1 - \alpha)}{n}\right\} = \alpha_n$$

Therefore, if the probability of the observed value is less than  $\alpha_n$  then the probability of the maximum order statistic exceeding the observed value is .01 or less.

## Appendix E – Analysis of Pass Rates and Statistical Indicators for Math

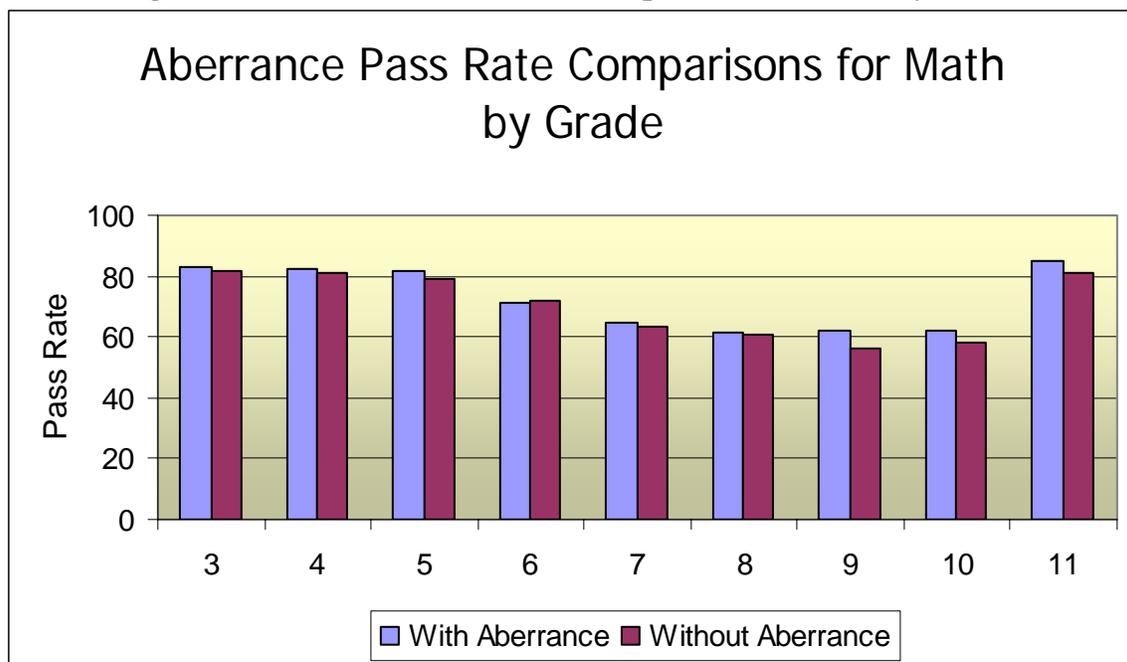
This appendix provides a detailed analysis of the plots in Appendices F and G for the Math test. The analysis is a discussion of “main effects” and “interaction effects” that are typical in Analysis of Variance interpretations. Each of the four statistical indicators is analyzed and interpreted.

The difference in pass rates is plotted for tests that exceed the threshold for the statistical indicator versus tests do not exceed the threshold for the statistical indicator. In general these plots do not show large effects. A clearer interpretation of the data results by splitting the classrooms into extreme and non-extreme classrooms using the statistical indicator (see Definition 11).

The Math data are now presented for each of the four statistical indicators. The plots and data for Math and the three other subjects (i.e., Reading/ELA, Science and Social Studies) are found in Appendices F and G.

The pass rate comparisons for the grades in the states are first presented. These plots show the overall effect for the state-wide test administrations that are present as a result of the detected statistical inconsistencies. Figure E-1 shows the difference in pass rates for the Math tests where aberrance is detected versus the Math tests where aberrance is not detected.

**Figure E-1: Aberrance Pass Rate Comparisons for Math by Grade**

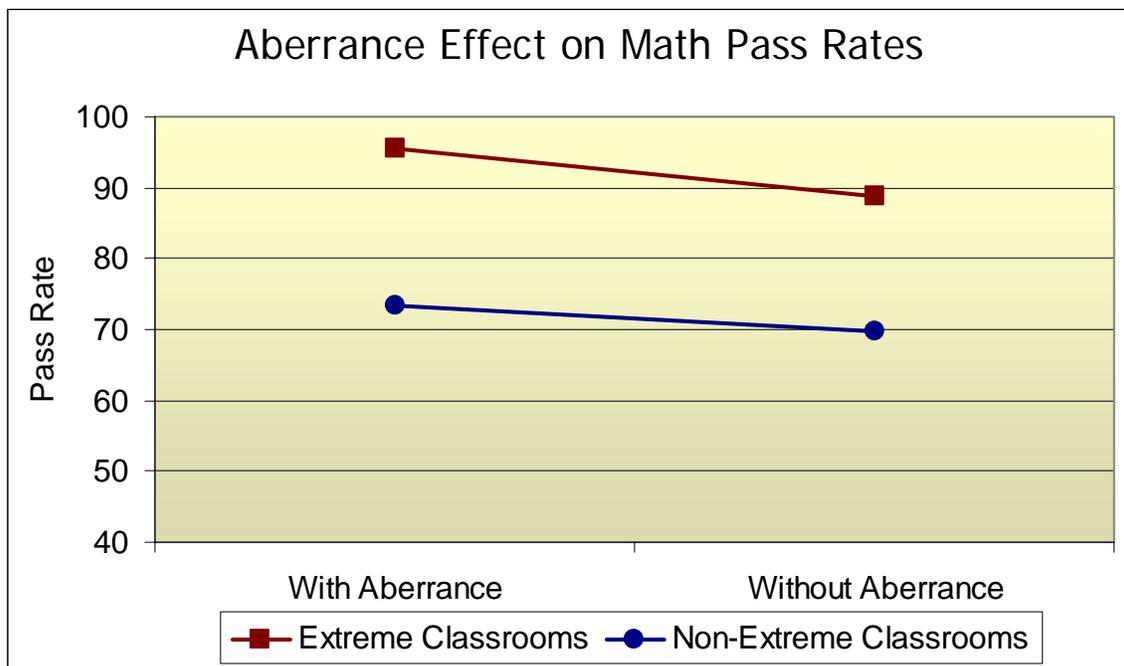


There is a slight increase in pass rates for all grades, except grade 6. The largest increases are observed in grades 9, 10 and 11. These effects seem rather small, given previous

experience and the nature of what is being measured. A greater understanding of the effects is obtained by examining the pass rates in “extreme” and “non-extreme” classrooms.

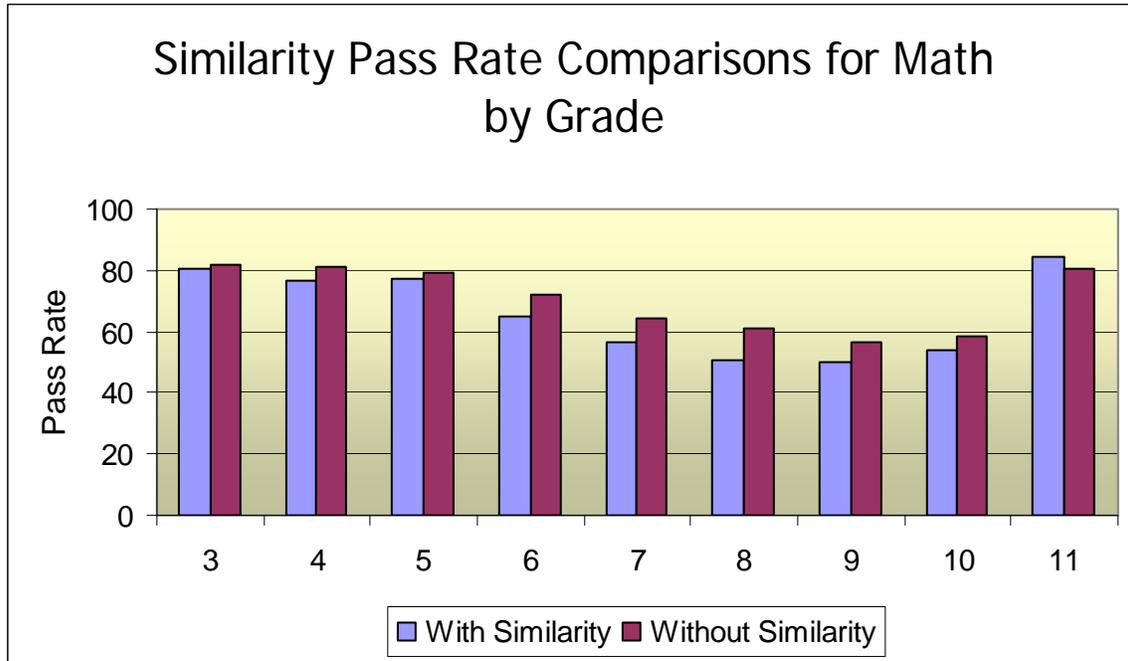
Figure E-2 compares the pass rates for the extreme and non-extreme aberrance classrooms. The figure shows a classic statistical main effect. There is a higher pass rate in the extreme classrooms (red line). The lines are almost parallel, indicating that the difference in pass rates between aberrant and non-aberrant tests is about the same in both the extreme and non-extreme aberrance classrooms. In other words, pass rates are higher for all the tests in the extreme classrooms, as opposed to just the tests detected with aberrance in the extreme classrooms.

**Figure E-2: Aberrance Effect on Math Pass Rates**



Following the presentation pattern above, Figures E-3 and E-4 illustrate the effects of very similar test responses on the Math tests.

**Figure E-3: Similarity Pass Rate Comparisons for Math by Grade**



The pass rates for tests where similarities were detected are noticeably lower than the pass rates for tests without similarities at all grades except 11. A more complete understanding is obtained in Figure E-4 where the pass rates are analyzed using the extreme and non-extreme classroom groups.

**Figure E-4: Similarity Effect on Math Pass Rates**

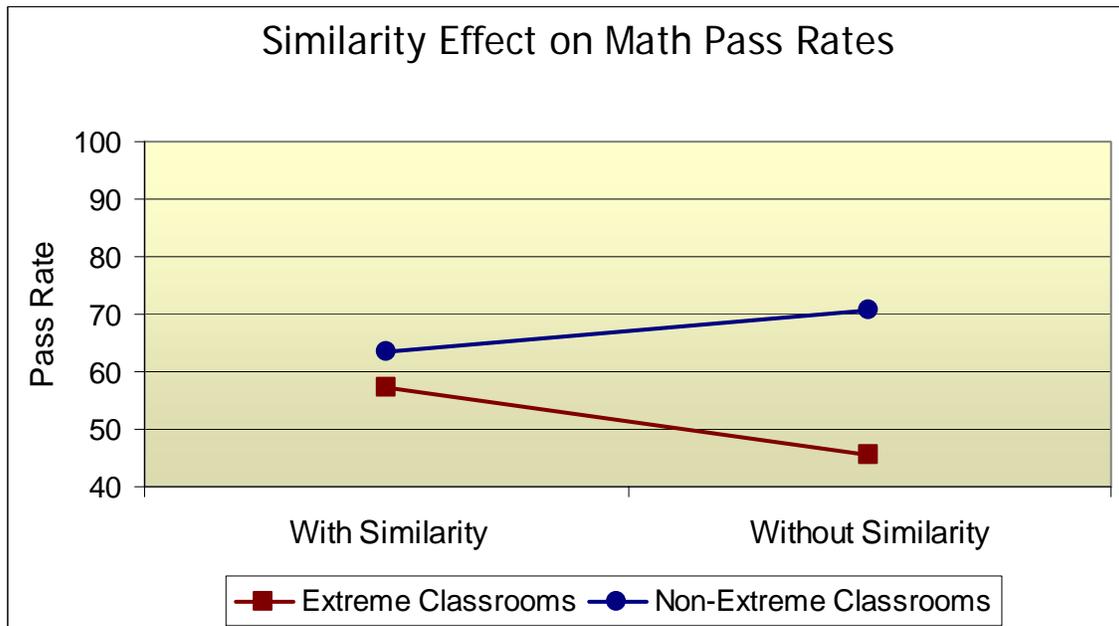
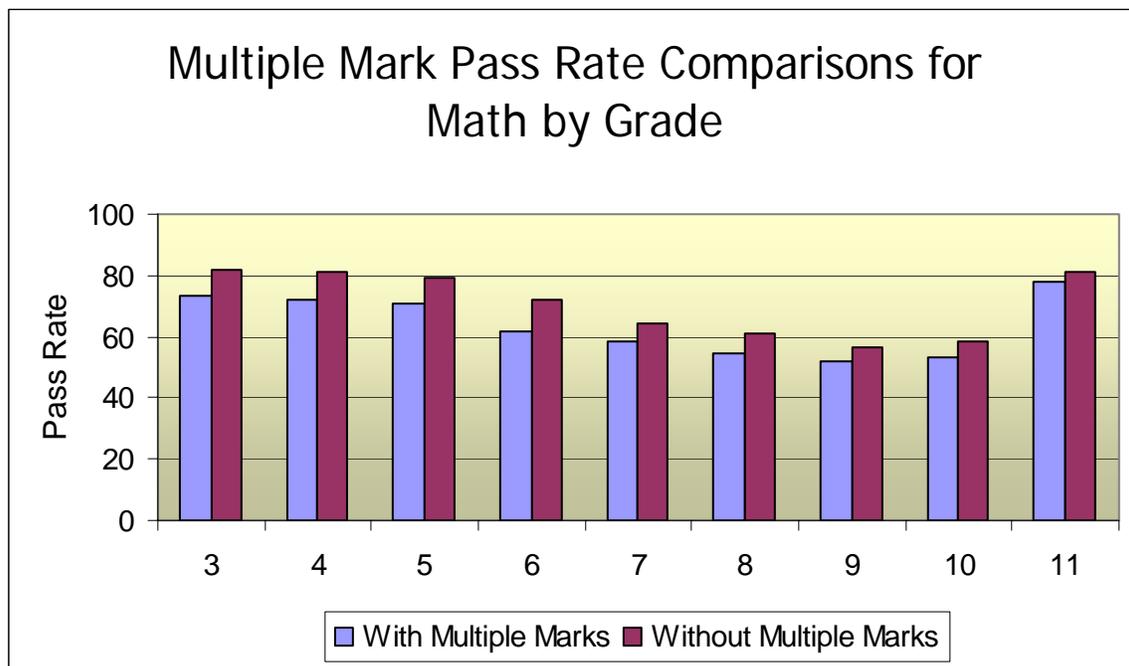


Figure E-4 reveals a strong interaction pattern. Similarity interacts with the extreme and non-extreme classroom groups. The non-extreme classrooms constitute nearly all the classrooms in the state and the pass rate effects in these classrooms are consistent with Figure E-3 (the pass rates are lower for very similar tests). However, in extreme classrooms the pass rate is higher for very similar tests than the pass rate for tests that were not similar. The difference in pass rates is shown by the red line in Figure E-4. In comparison, the blue line (non-extreme classrooms) shows a lower pass rate for very similar tests than the pass rate for tests that were not similar. The actual interaction effect is a 12% pass rate gain due to similarity within extreme classrooms.

Figures E-5 and E-6 present the data for the excessive multiple marks on the Math tests.

**Figure E-5: Multiple Mark Pass Rate Comparisons for Math by Grade**



The pass rates are noticeably lower in all grades when excessive multiple marks are present. This seems surprising since erasing and answer changing is a technique that a few educators have utilized in the past to raise student test scores. A more complete understanding is obtained from Figure E-6 where the pass rates in the extreme and non-extreme classrooms are compared.

**Figure E-6: Multiple Marks Effect on Math Pass Rates**

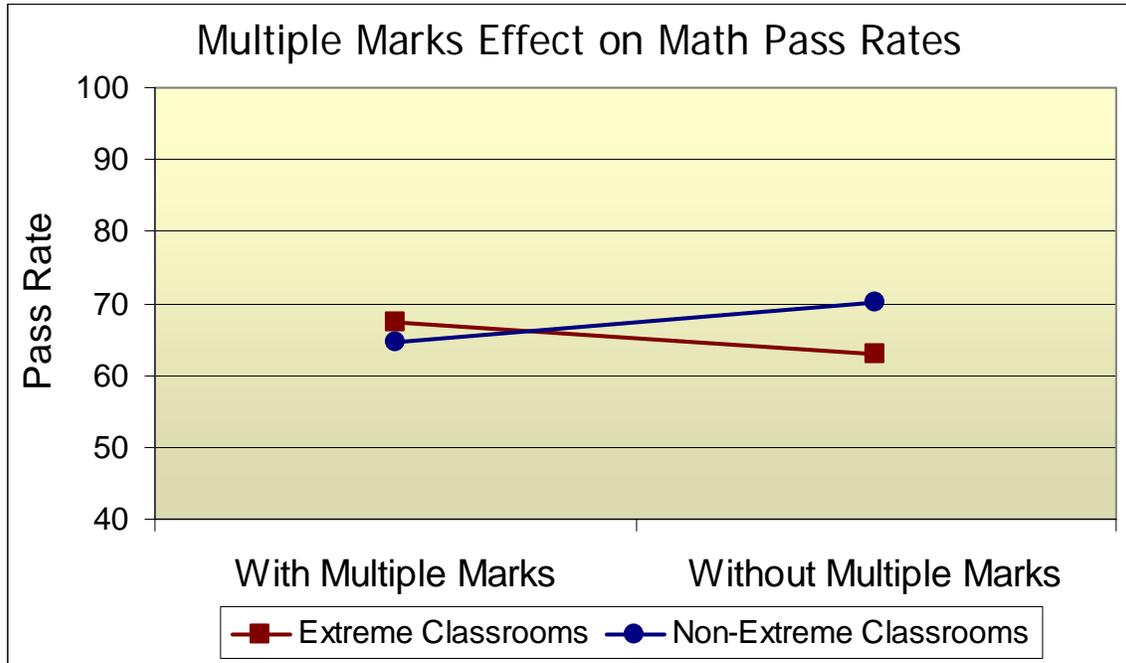
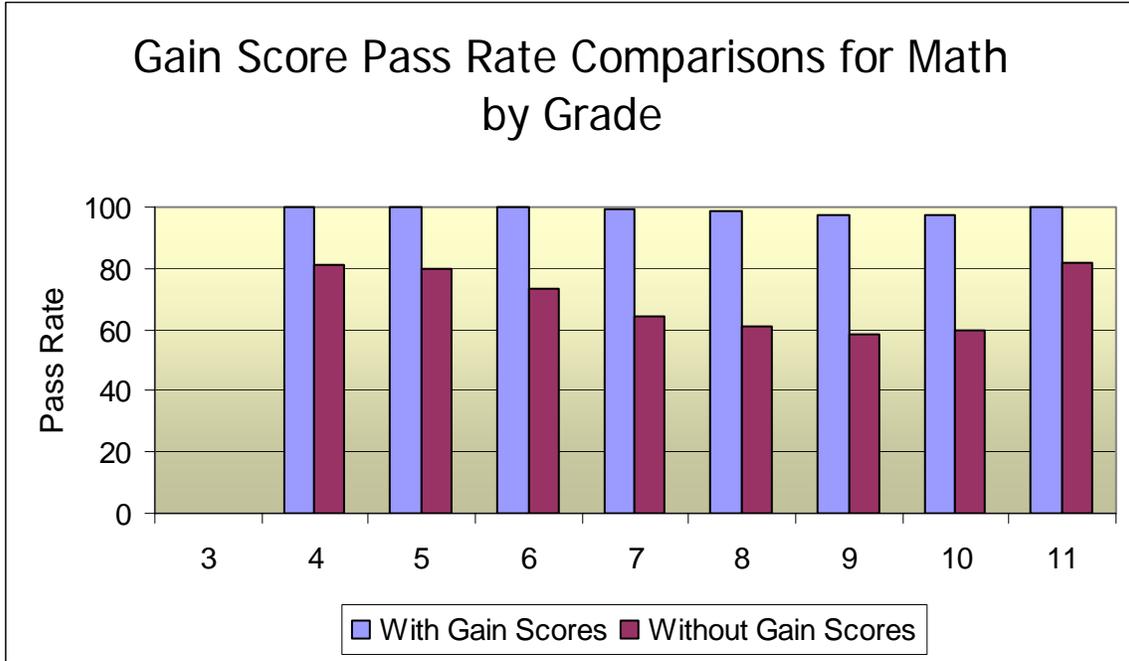


Figure E-6 shows a classic interaction pattern with crossed lines where the relationship between the pass rates and the extreme or non-extreme classrooms reverses itself depending upon whether the tests have high multiple marks or not. In this regard the data clearly show that pass rates are higher for tests having high multiple marks where answer changes from wrong to right are excessive (i.e., in extreme classrooms).

The data for the final statistical indicator, unusual gains on the Math tests, is shown in Figures E-7 and E-8.

**Figure E-7: Gain Score Pass Rate Comparison for Math by Grade**

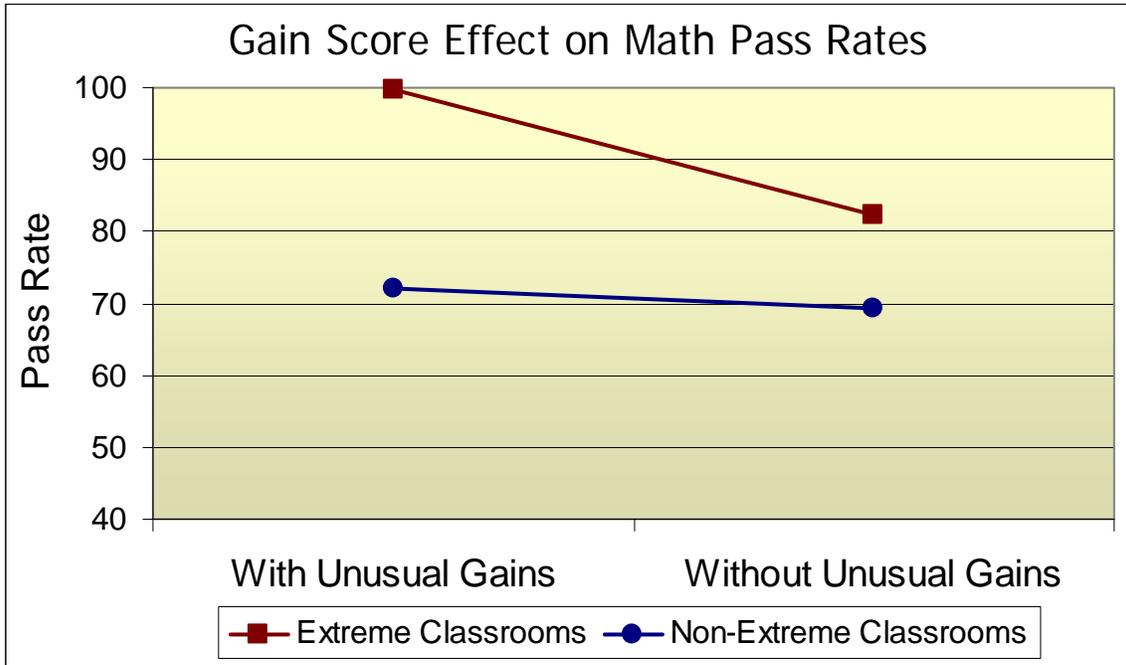


(Note: Gain scores were not computed for grade 3 since there are no grade 2 TAKS tests.)

There is a pronounced effect of high gain scores on pass rates for all grades. This effect is confounded with the definition of a high gain score, since nearly always a high gain will result in meeting or exceeding the TAKS standard. However, the size of the effect is very large. The overall difference in pass rates is at least 20% and in some grades it is as high as 40%.

Figure E-8 provides a deeper view into the data.

**Figure E-8: Gain Score Effect on Math Pass Rates**



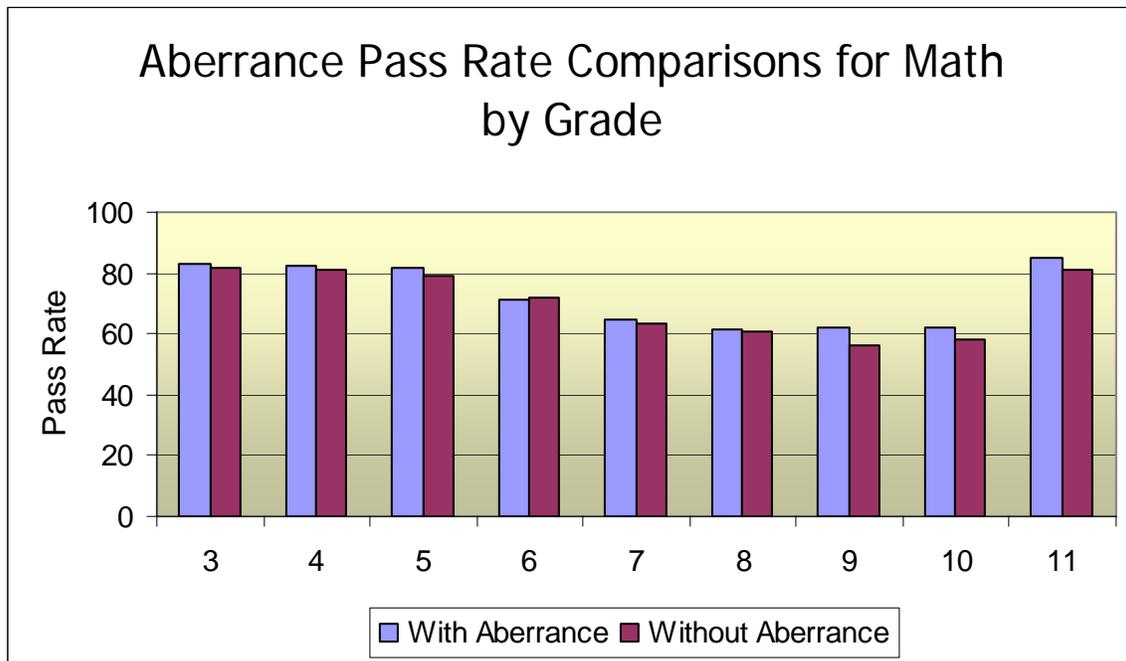
There are clearly higher pass rates observed in the extreme classrooms for both students with and without unusual gains. The pass rate for the students that do not have unusual gains is about 13% higher than the pass rate for students without unusual gains in non-extreme classrooms (82% vs. 69%).

Extreme classrooms where the number of students with high gains is much greater than would be expected using the statewide rates are definitely very different than the other classrooms. Only two explanations for these plots seem viable: excellent instruction or testing irregularities.

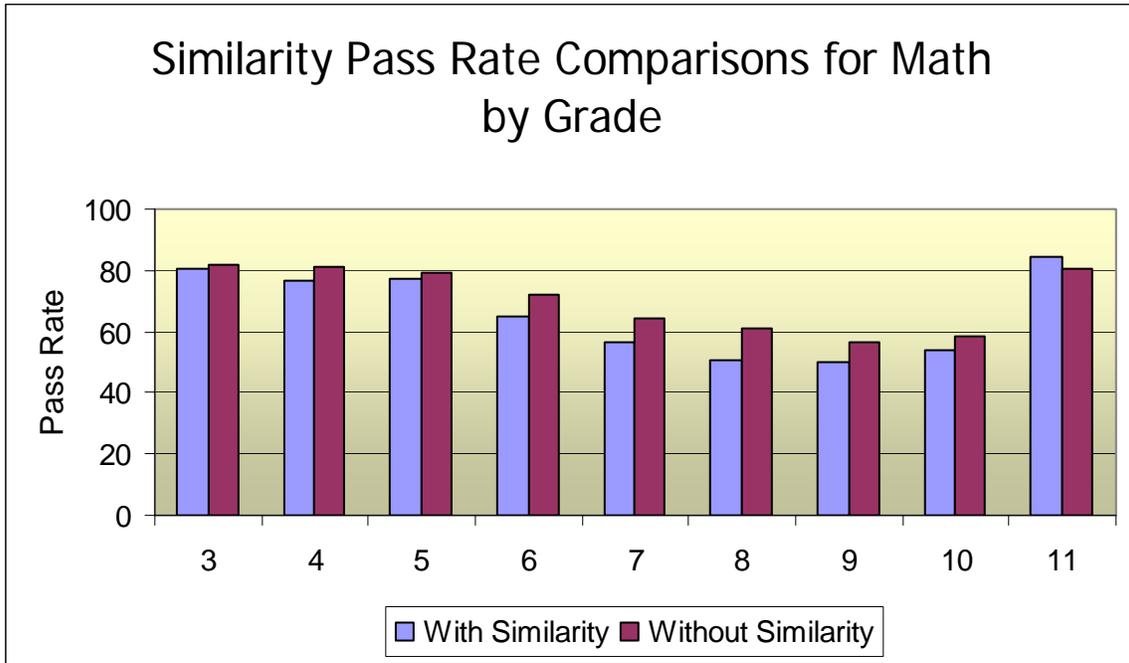
## Appendix F – Grade-level Pass Rates for the Four Statistical Indicators

These plots compare pass rates for the tests where the value of the statistical indicator exceeded the threshold (which was set at the theoretical 5% level) with pass rates for tests where the value of the statistical indicator did not exceed the critical threshold. If there is no association between the statistical indicator and the pass rate, then pass rate differences between the two groups of tests will be due to normal variability. Presumably, the statistical indicators measure quantities that increase when testing irregularities are present. If this is so, the pass rate should be higher for tests which exceed the threshold than for tests that do not exceed the threshold.

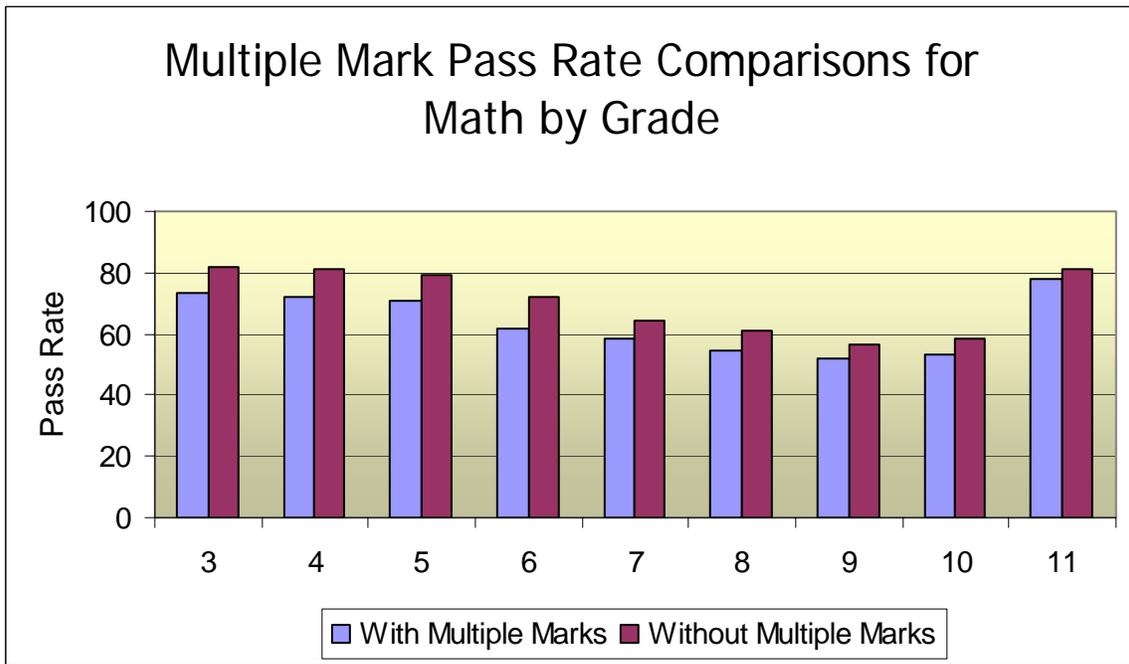
**Figure F-1: Aberrance Pass Rate Comparisons for Math by Grade**



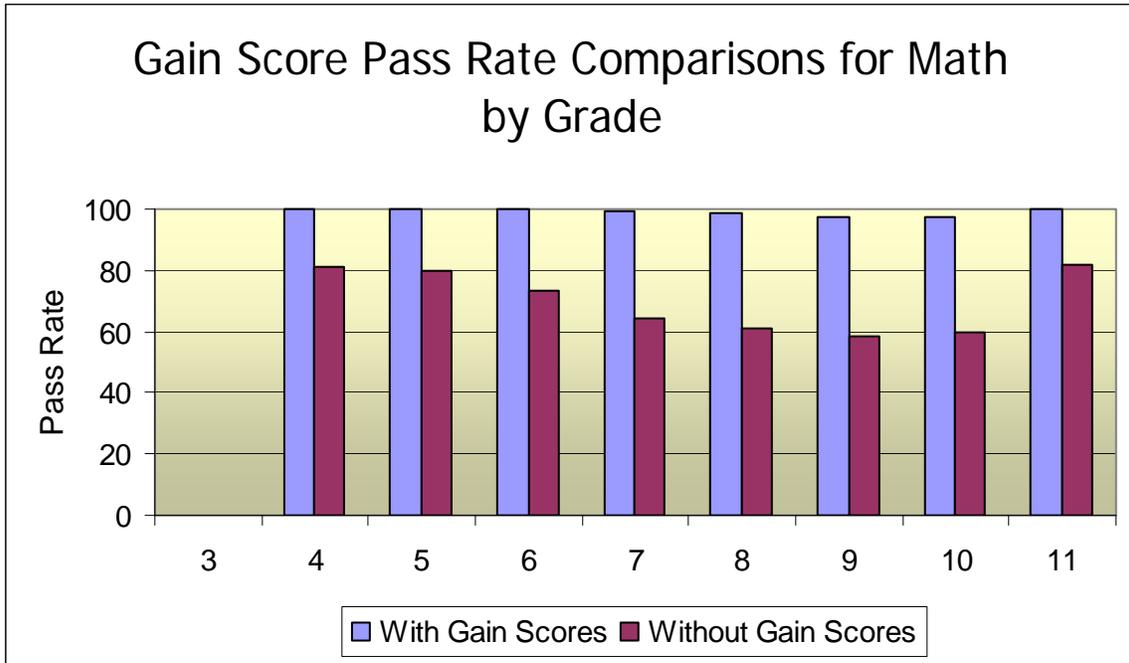
**Figure F-2: Similarity Pass Rate Comparisons for Math by Grade**



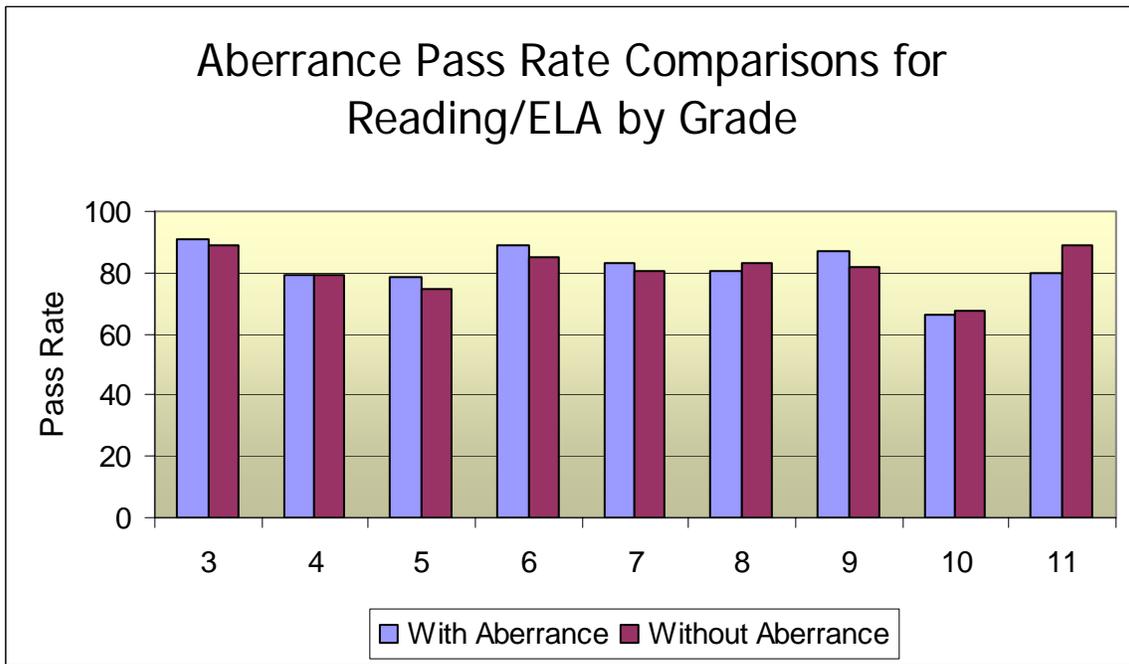
**Figure F-3: Multiple Mark Pass Rate Comparisons for Math by Grade**



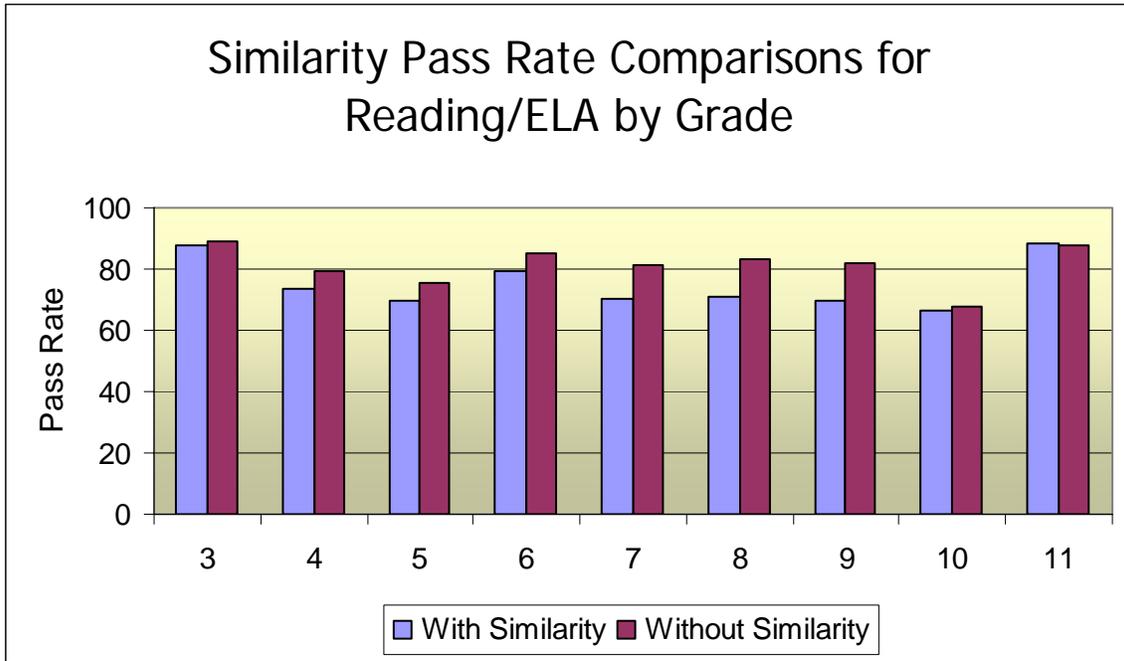
**Figure F-4: Gain Score Pass Rate Comparisons for Math by Grade**



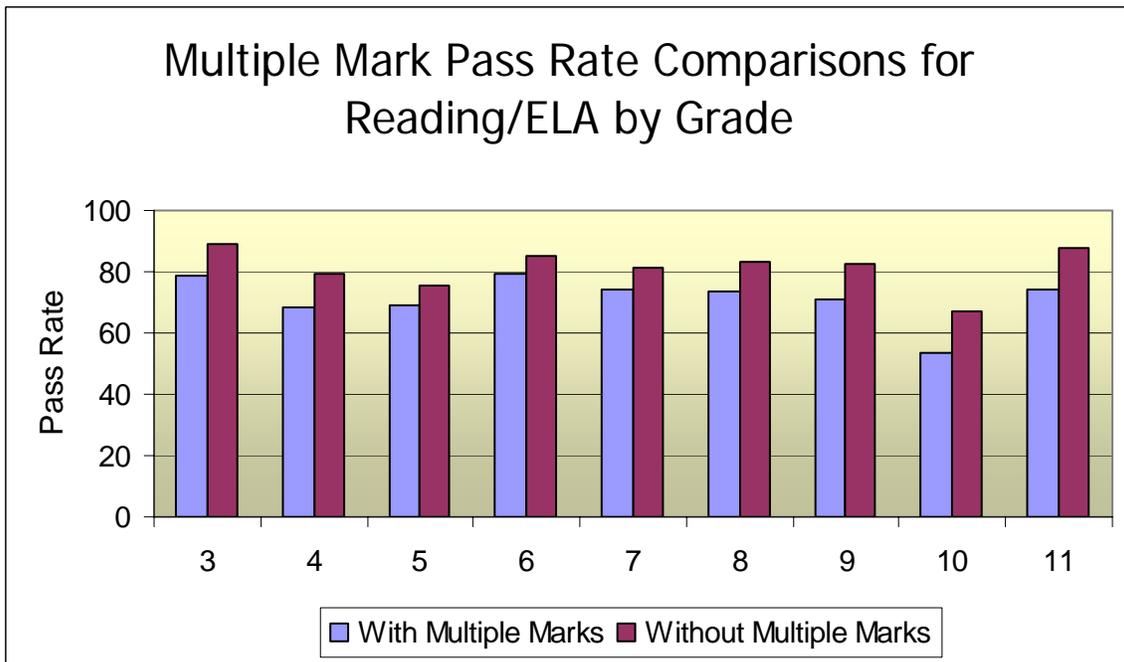
**Figure F-5: Aberrance Pass Rate Comparisons for Reading/ELA by Grade**



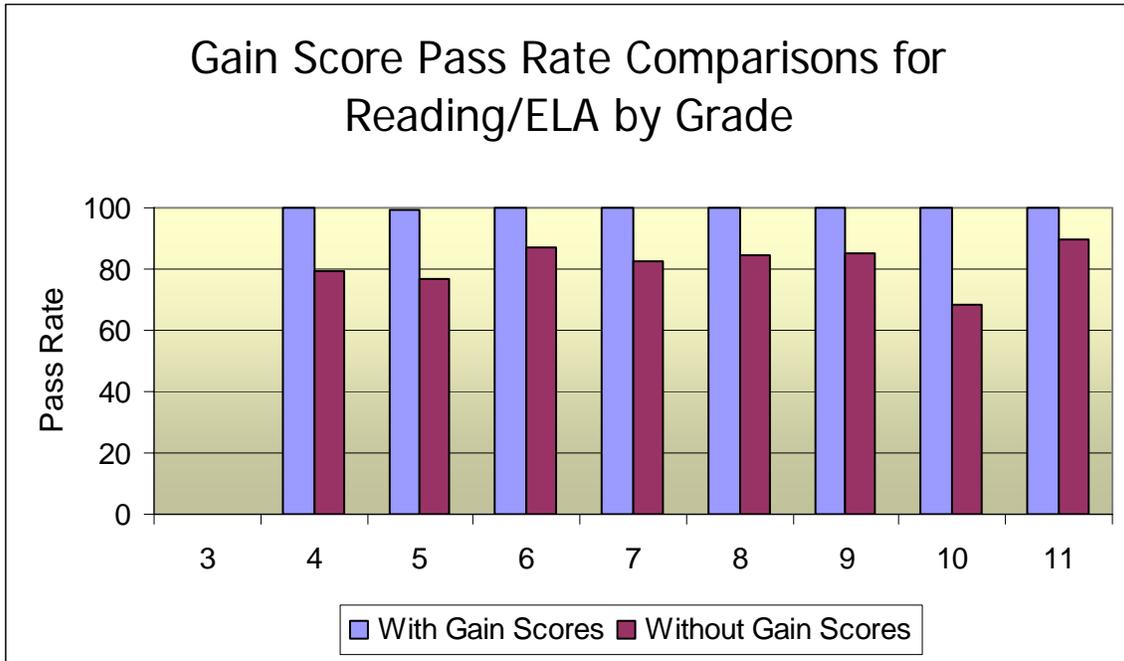
**Figure F-6: Similarity Pass Rate Comparisons for Reading/ELA by Grade**



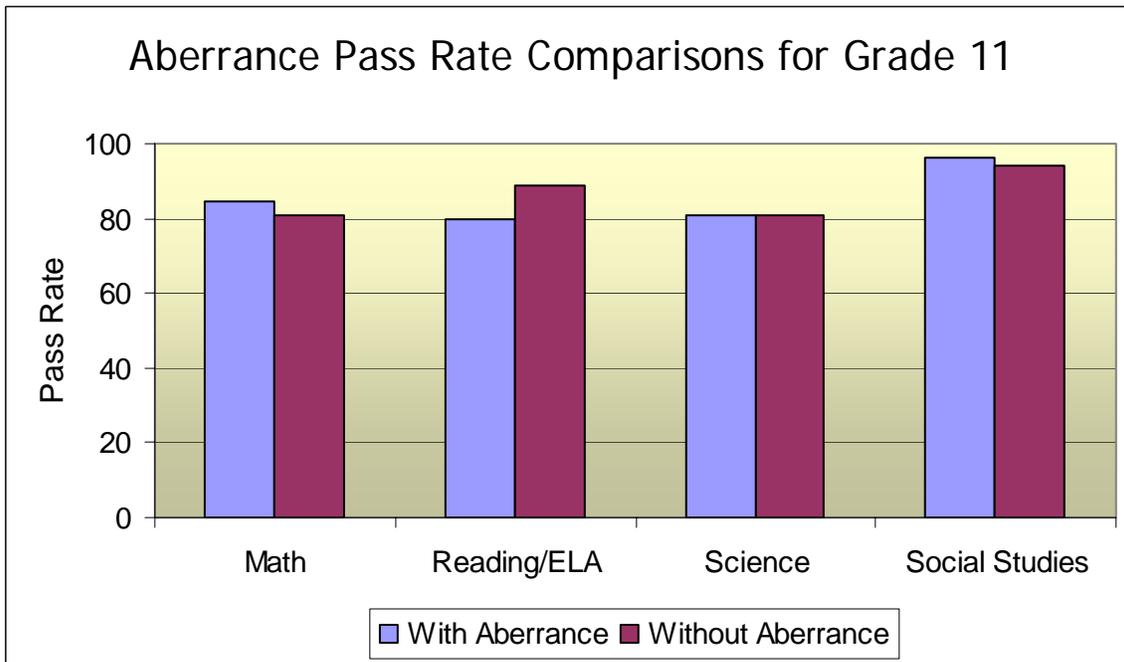
**Figure F-7: Multiple Mark Pass Rate Comparisons for Reading/ELA by Grade**



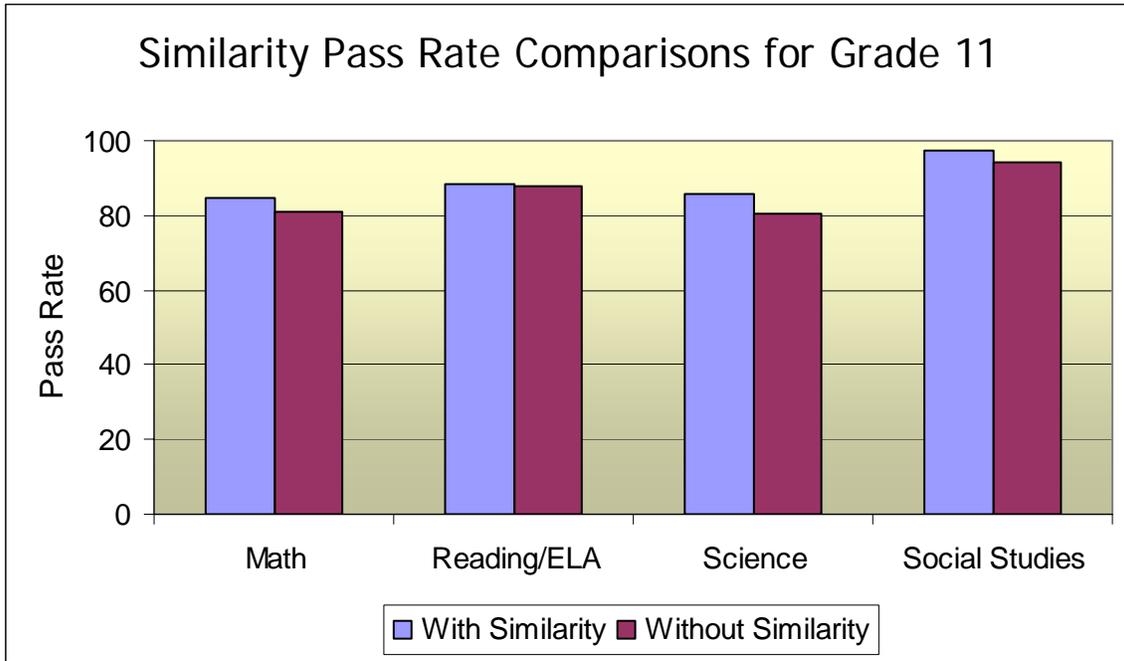
**Figure F-8: Gain Score Pass Rate Comparisons for Reading/ELA by Grade**



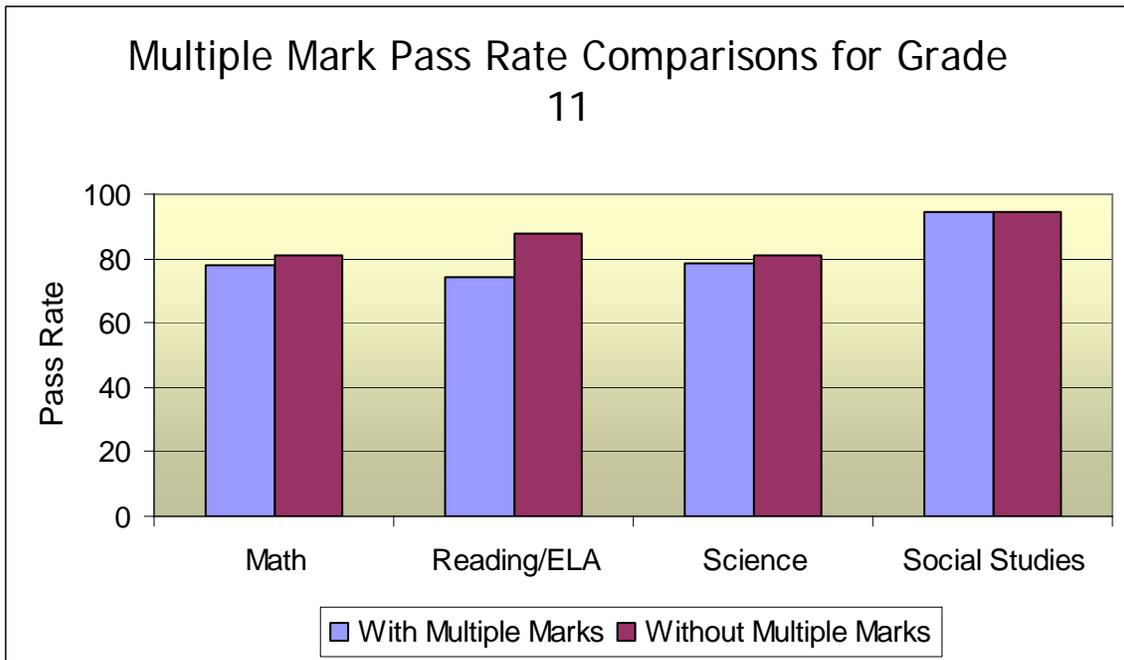
**Figure F-9: Aberrance Pass Rate Comparisons for Grade 11**



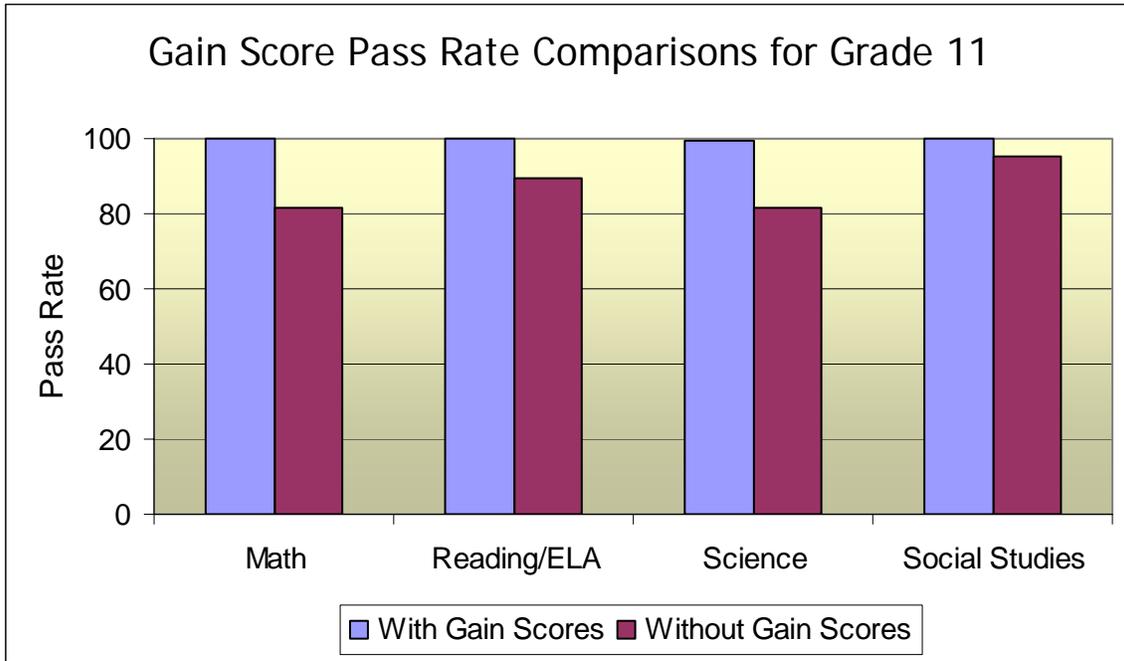
**Figure F-10: Similarity Pass Rate Comparisons for Grade 11**



**Figure F-11: Multiple Mark Pass Rate Comparisons for Grade 11**



**Figure F-12: Gain Score Pass Rate Comparisons for Grade 11**



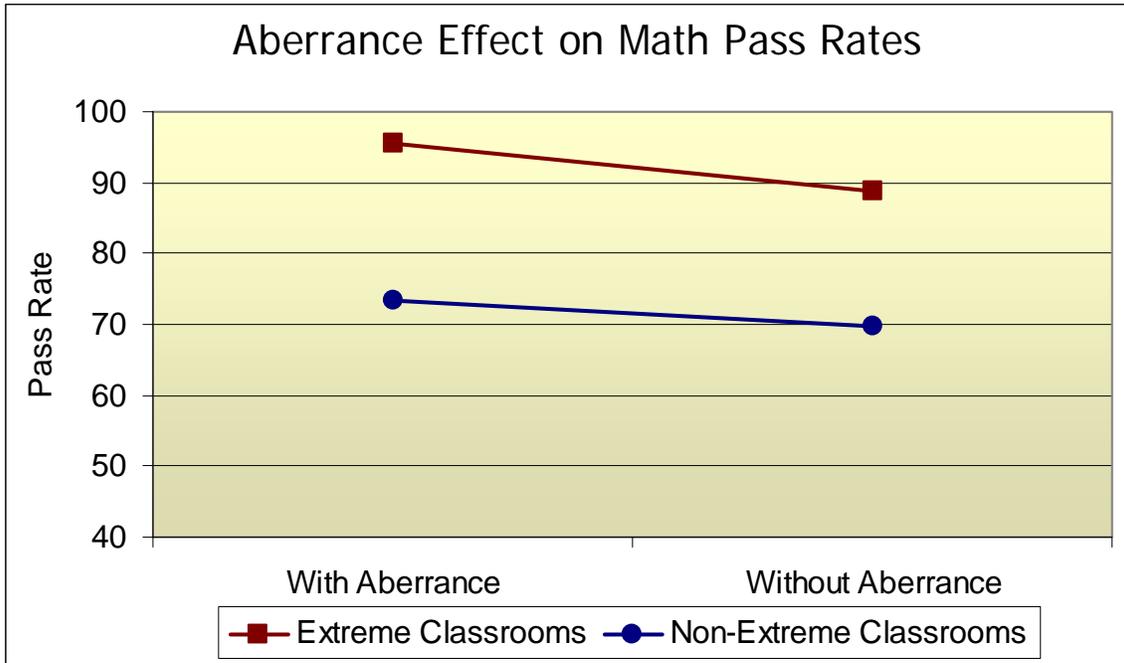
## **Appendix G – Interaction Effects of Statistical Indicators and Pass Rates**

These plots illustrate the association between the statistical indicators of aberrance, similarity, multiple marks, and unusual gains and pass rates. The plots are constructed by assigning the classrooms into two groups. All classrooms that exhibit an extreme value of the statistical indicator are assigned into the “Extreme Classrooms” group. All remaining classrooms are assigned into the “Non-Extreme Classrooms” group. The tests within these two groups are categorized into two groups. All tests that exceeded the critical threshold of the statistical indicator are assigned in the “with” group. All remaining tests are assigned in the “without” group.

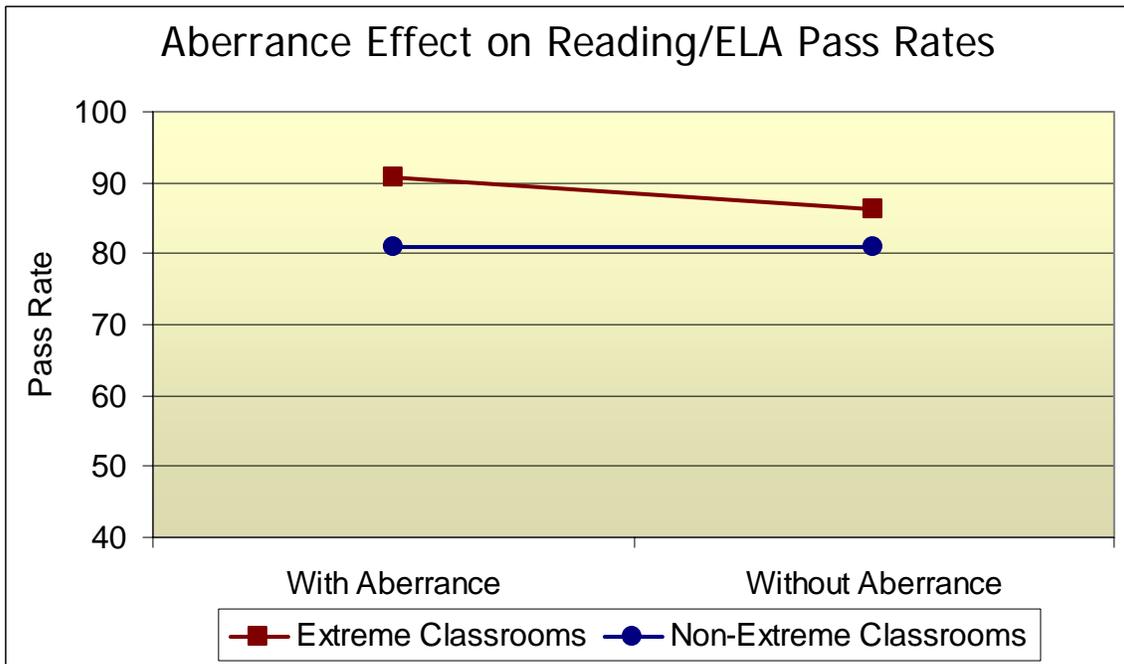
Every test is assigned into one of four categories: Extreme Classroom with the statistical inconsistency, Extreme Classroom without the statistical inconsistency, Non-Extreme Classroom with the statistical inconsistency, and Non-Extreme Classroom without the statistical inconsistency. If the value of the statistical indicator has no association with pass rate then we would expect the pass rates for the four groups to be the same. If the value of the statistical indicator always has a positive effect on pass rate, then we would expect the two groupings with the statistical inconsistency to show an increase over the paired group of tests in the respective classroom group.

However, the values of the statistical indicators are subject to variability and the values naturally arise at approximate the 5% level in the data without any testing irregularity being present. Therefore, the comparison of interest is comparing the pass rates for tests with statistical inconsistency within the extreme classroom group against the pass rates for tests without the statistical inconsistency within the extreme classroom group. These plots show these comparisons.

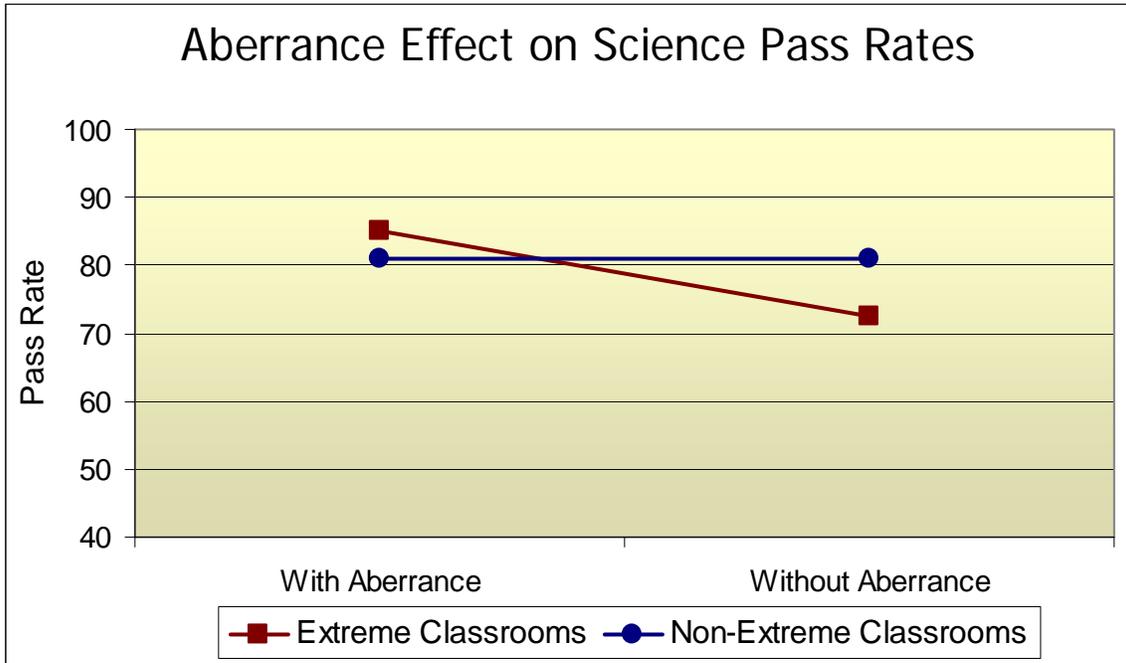
**Figure G-1: Aberrance Effect on Math Pass Rates**



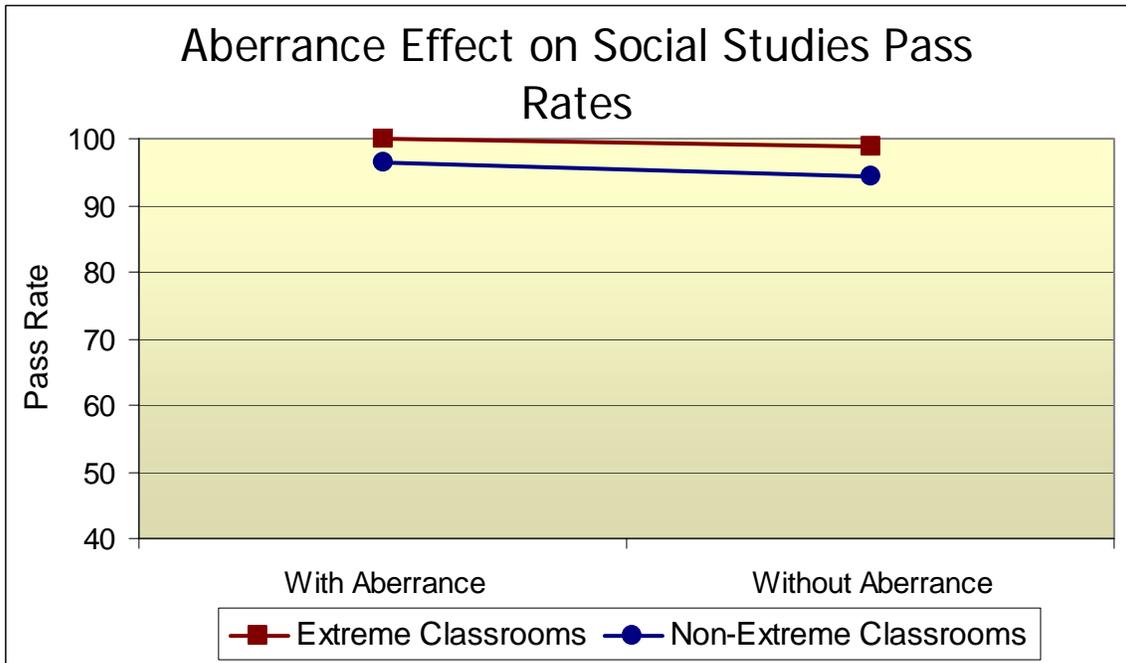
**Figure G-2: Aberrance Effect on Reading/ELA Pass Rates**



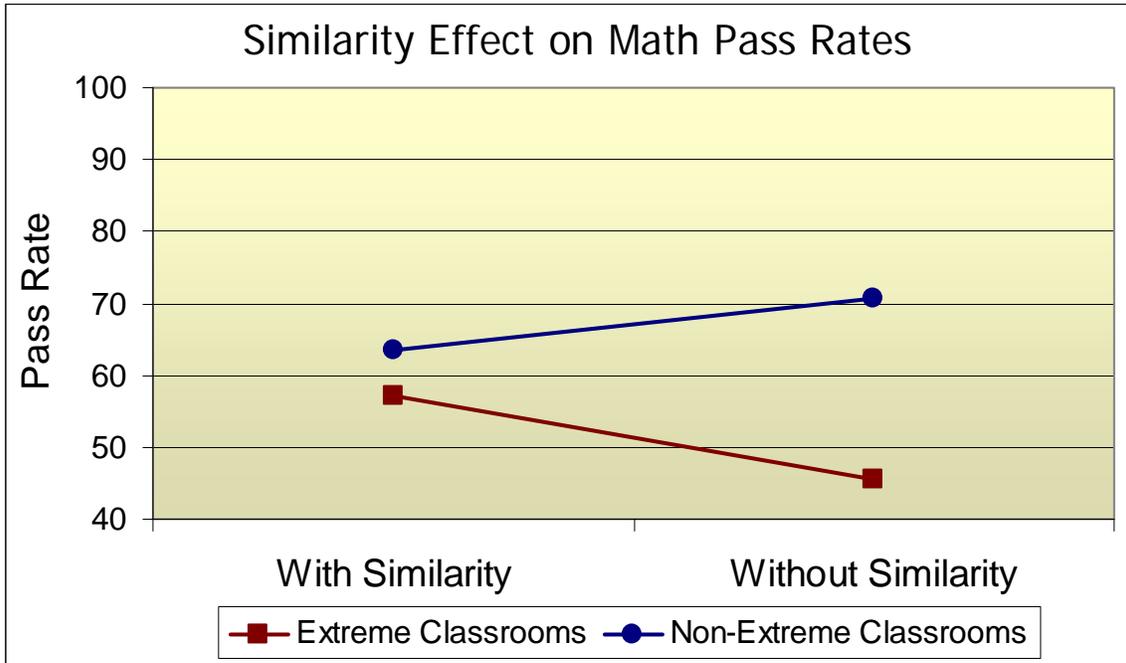
**Figure G-3: Aberrance Effect on Science Pass Rates**



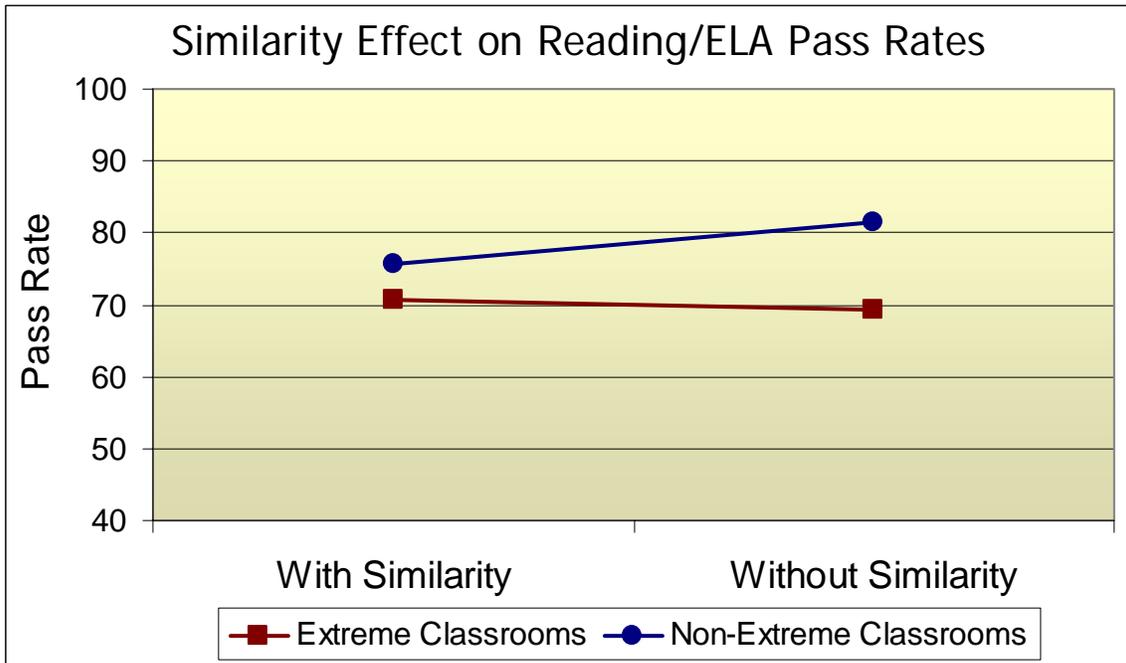
**Figure G-4: Aberrance Effect on Social Studies Pass Rates**



**Figure G-5: Similarity Effect on Math Pass Rates**



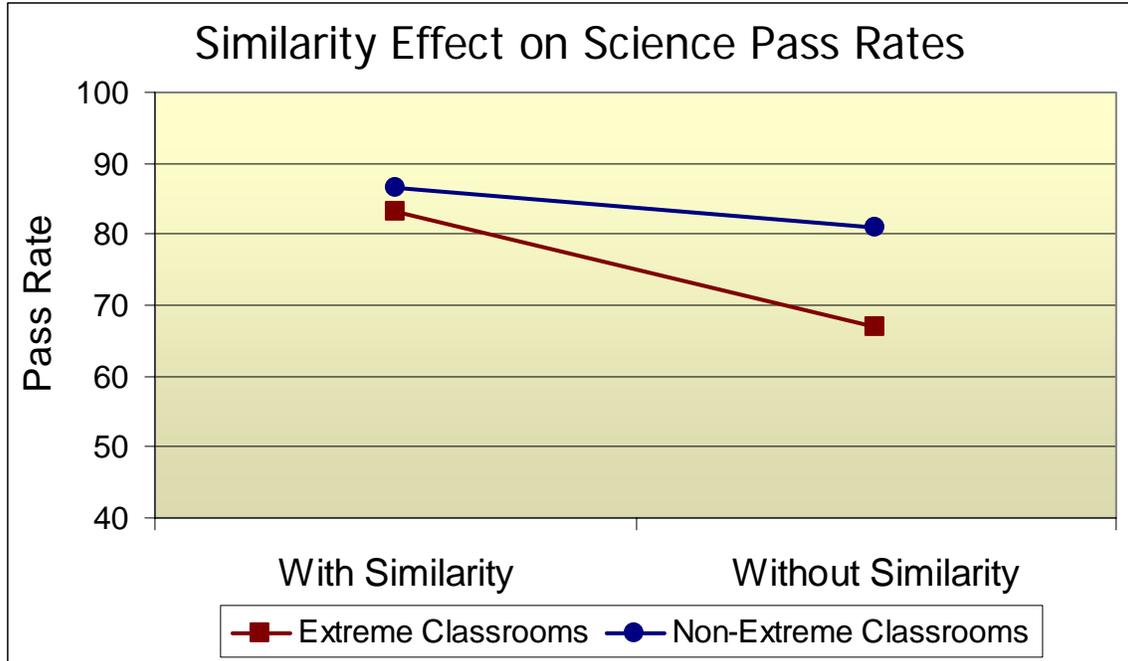
**Figure G-6: Similarity Effect in Reading/ELA Pass Rates**



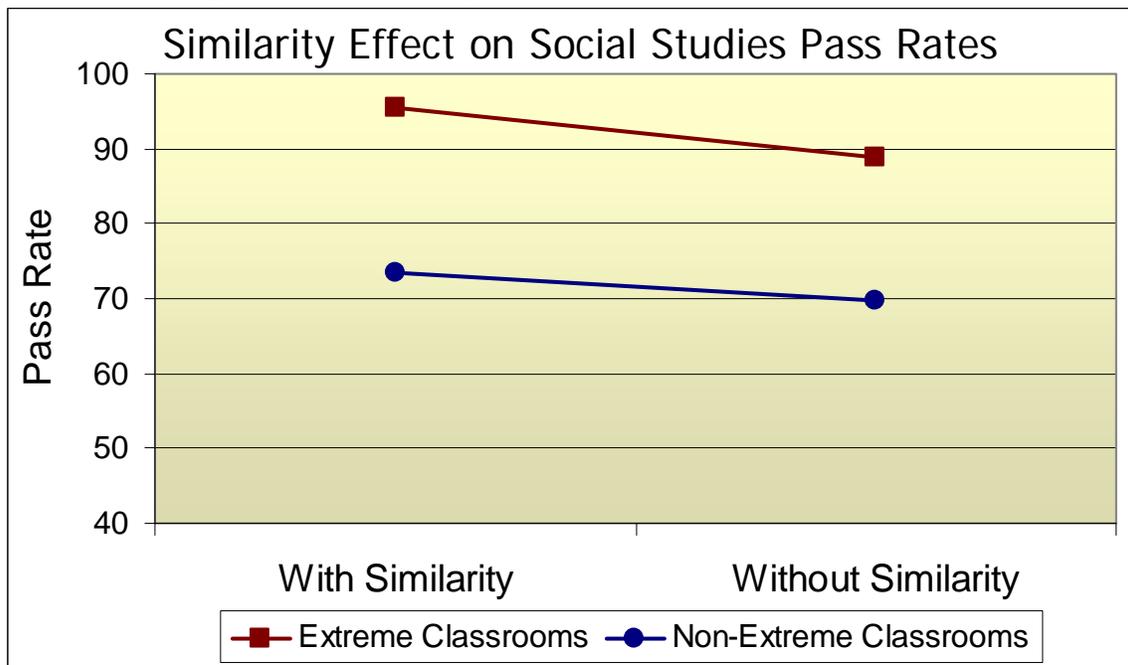
The above data shows that in non-extreme classrooms the pass rate is about 6% lower for highly similar tests as compared to the pass rate for tests where similarity is not detected. There appears to be higher rates of similarity among lower performing students. This situation is reversed in the extreme classrooms, since the pass rate is about 1% higher for

tests where similarity is detected as compared to the pass rates for tests where similarity is not detected.

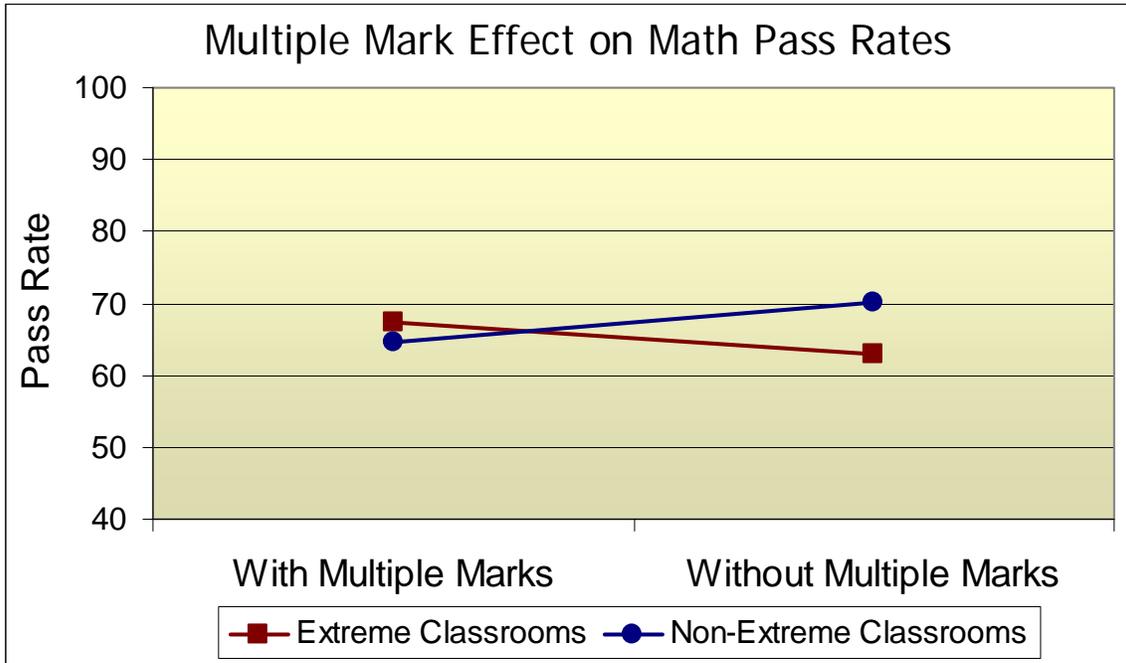
**Figure G-7: Similarity Effect on Science Pass Rates**



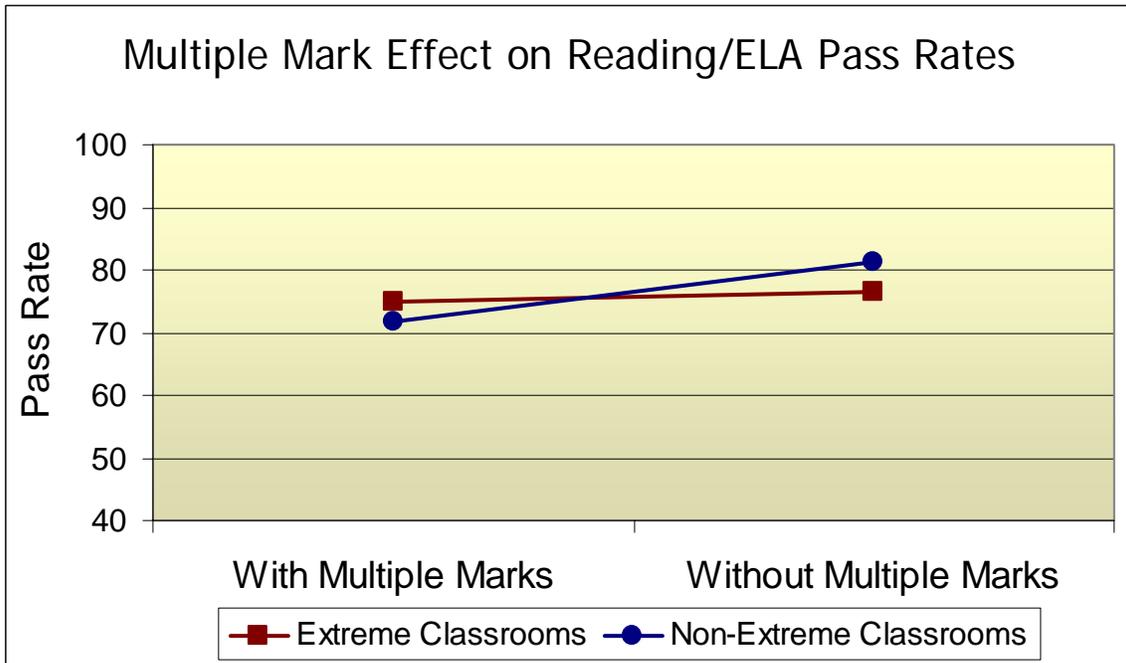
**Figure G-8: Similarity Effect on Social Studies Pass Rates**



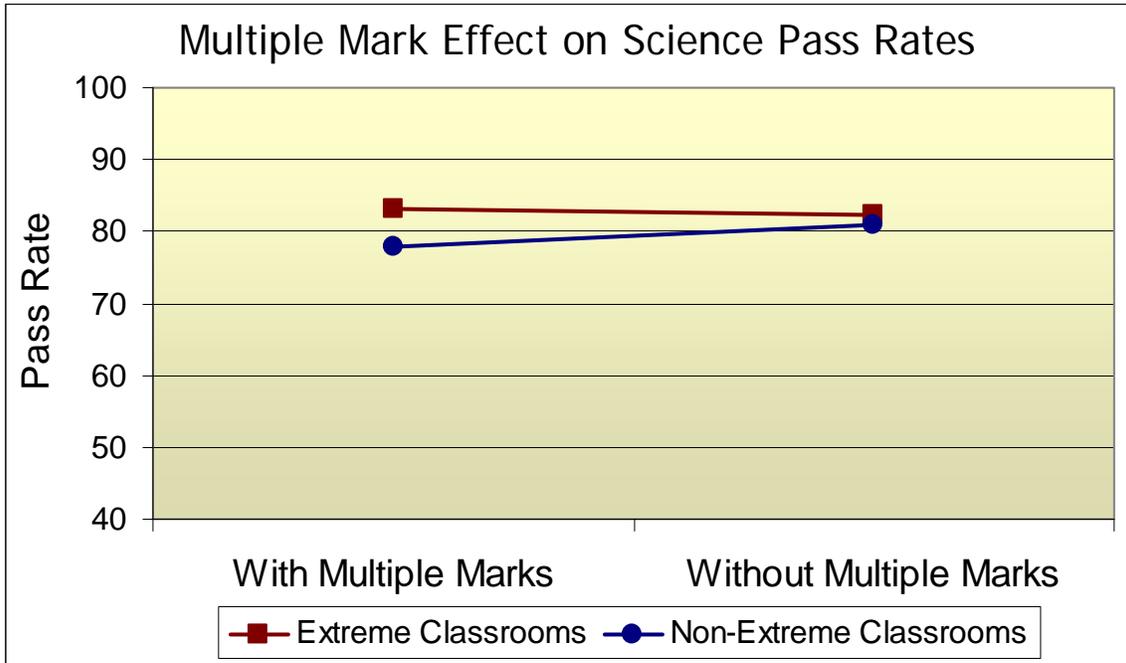
**Figure G-9: Multiple Mark Effect on Math Pass Rates**



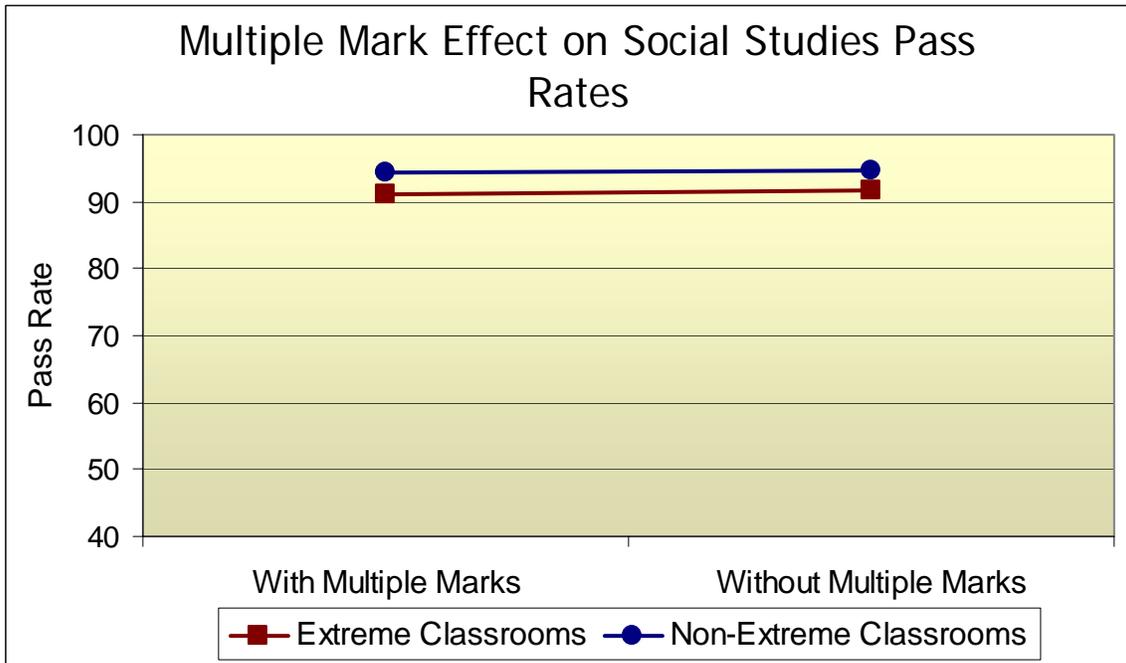
**Figure G-10: Multiple Mark Effect on Reading/ELA Pass Rates**



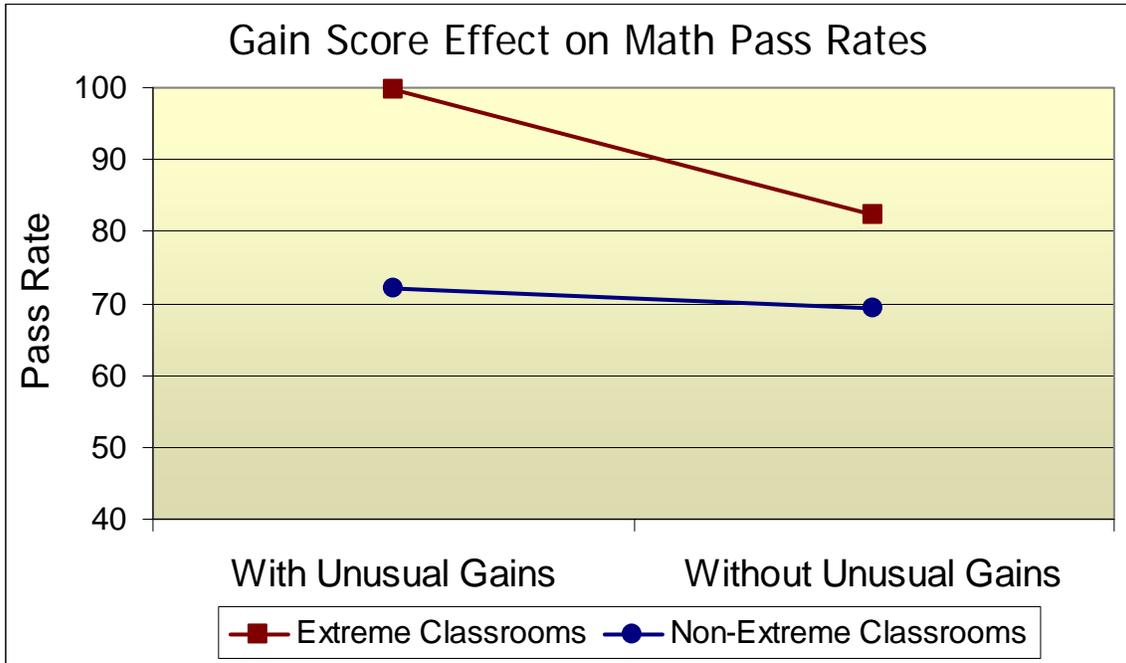
**Figure G-11: Multiple Mark Effect on Science Pass Rates**



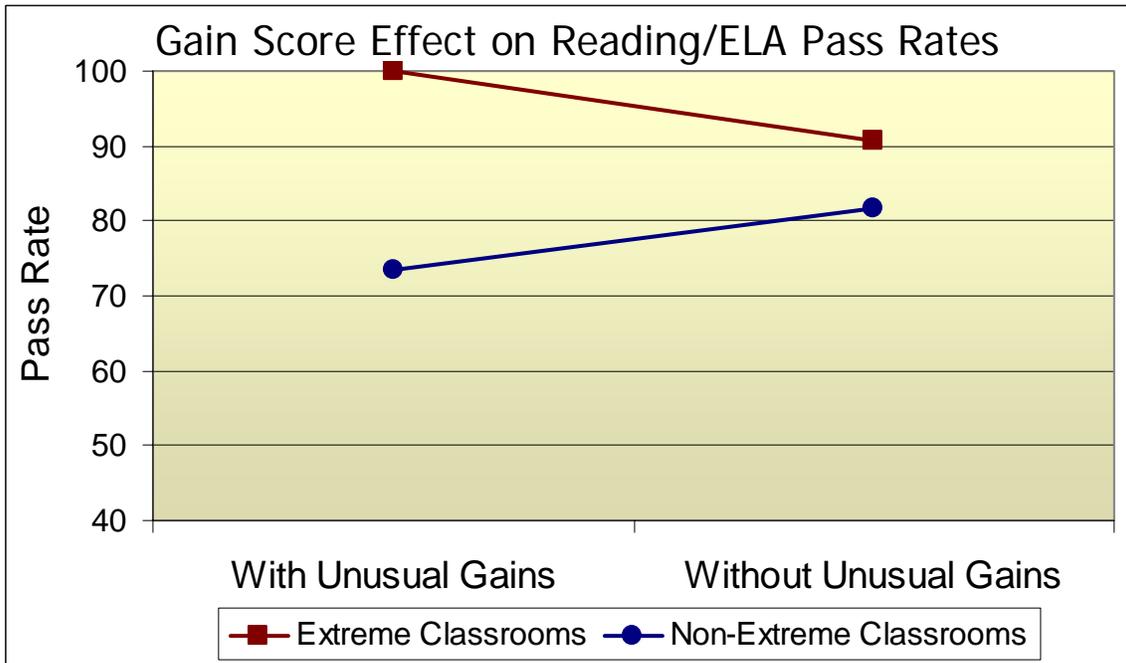
**Figure G-12: Multiple Mark Effect on Social Studies Pass Rates**



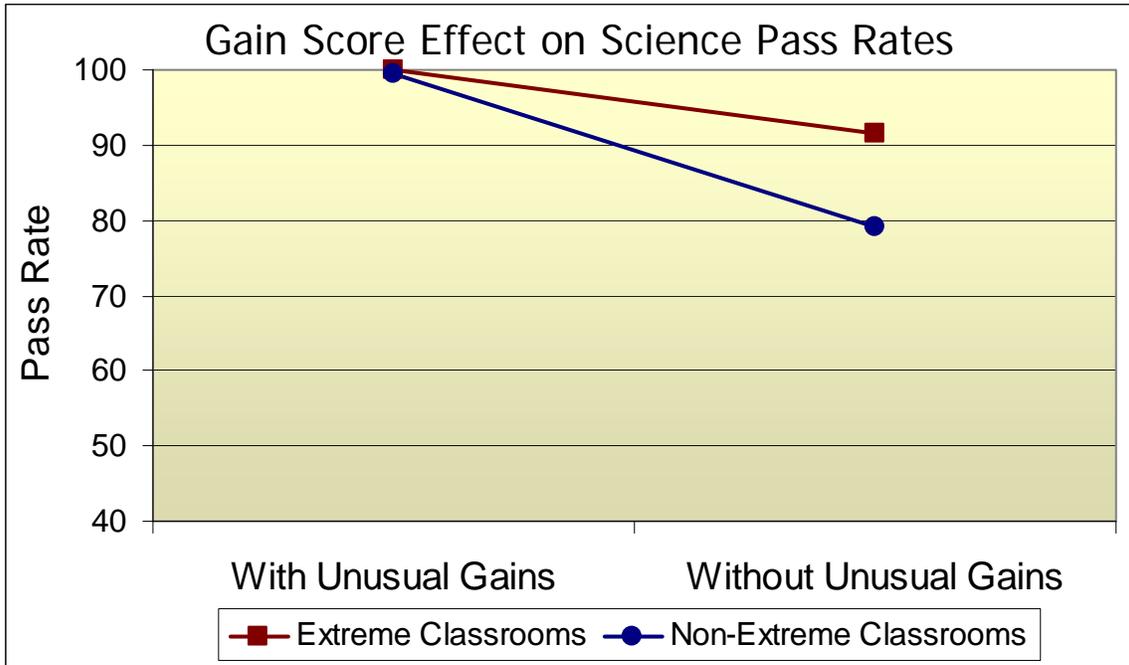
**Figure G-13: Gain Score Effect on Math Pass Rates**



**Figure G-14: Gain Score Effect on Reading/ELA Pass Rates**

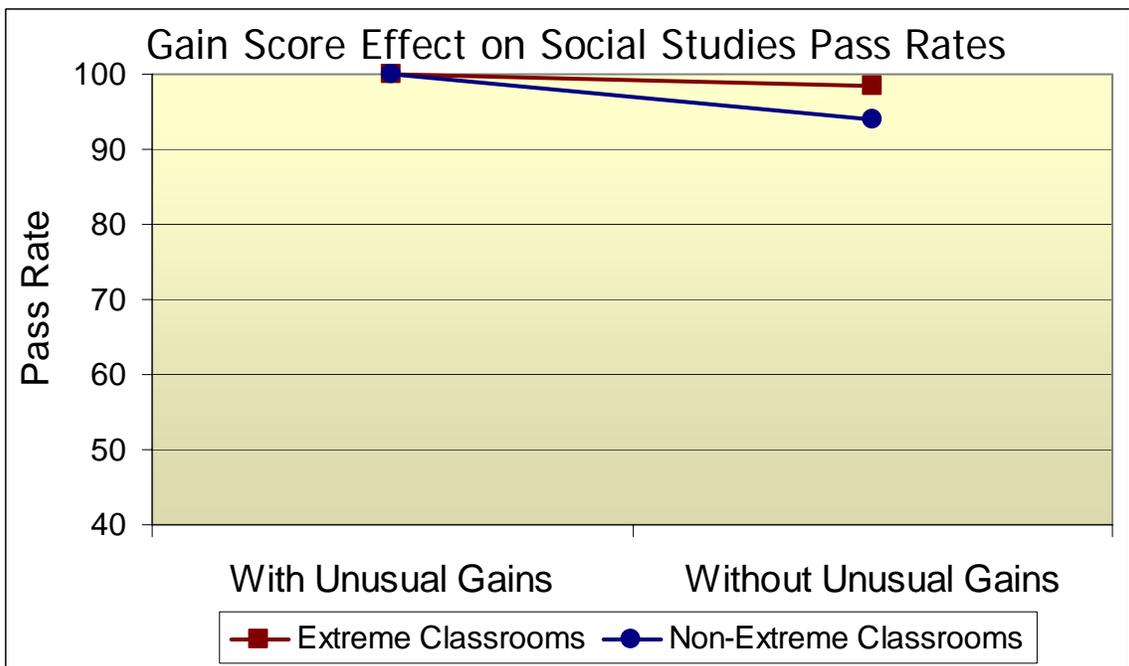


**Figure G-15: Gain Score Effect on Science Pass Rates**



The above figure shows a strong interaction effect, but the predominant effect is that the pass rate for extreme classrooms is higher than the pass rate for non-extreme classrooms. The large increase of the pass rate when unusual gains are present is strongly related to what an unusual gain means. In nearly all cases an unusual gain means the student will meet or exceed the TAKS standard.

**Figure G-16: Gain Score Effect on Social Studies Pass Rates**



## Appendix H – Probability Regions for Exception Concentration Categories

Figures 12 and 13 (in the main body of the report) provide histograms of districts where concentrations of school exceptions are present. A description of the five concentration categories along with probability regions (see Figures H-1 through H-8) for the categories is provided within this appendix.

The categories of very low, low, medium, high and very high are constructed using probability ranges<sup>17</sup> of observing the reported number of exceptions or anomalies within the district. A binomial distribution is assumed and the associated statewide rate is used as the binomial proportion. The concentration category is defined by the probability of the observing the number of detected exceptions within the district or more (i.e., an upper tail probability is used).

If  $x$  is the number of observed exceptions and if  $n$  is the district size (i.e., number of schools or classrooms) then the probability that the number of exceptions is greater than or equal to  $x$  can be computed using a binomial distribution. For example, if  $p(k \geq x)$  is less than or equal to .0001 the district would be placed in the “Very High” concentration category for exceptions.

The probability ranges used for the five categories are:

- Very low – The probability is greater than .75.
- Low – The probability is less than .25, but greater than .05
- Medium – The probability is less than .05, but greater than .002
- High – The probability is less than .002, but greater than .0001
- Very High – The probability is less than .0001

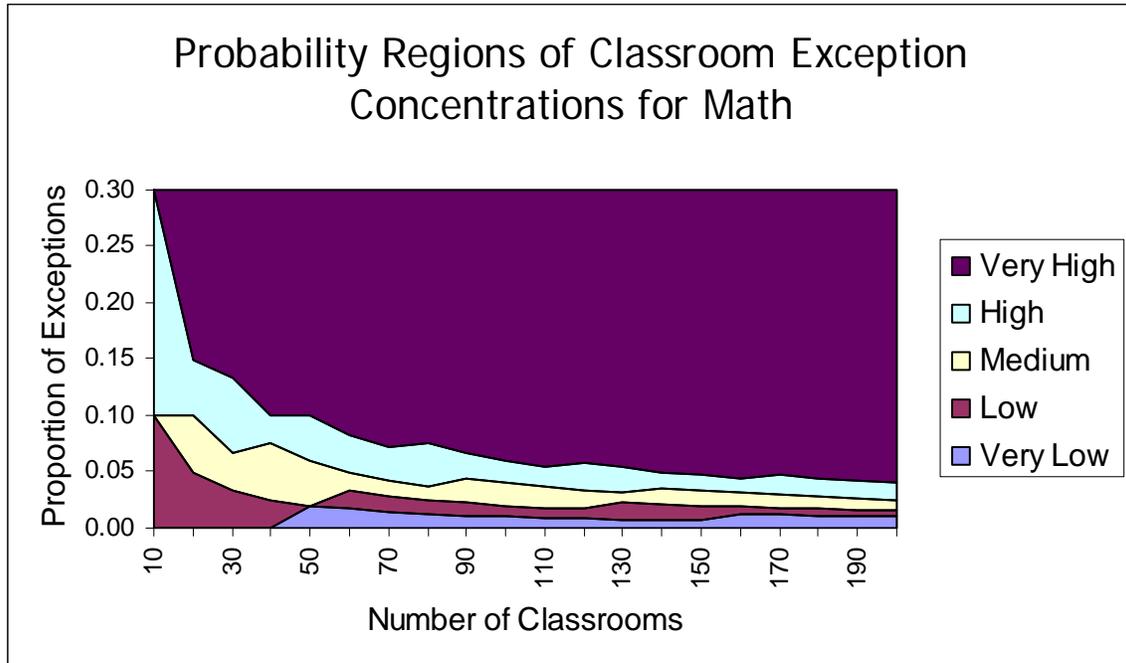
The probabilities depend upon three variables: the statewide rate, the number of schools or classrooms within the district, and the observed number of exceptions. These three variables define a probability region for each of the five concentration categories. The figures in this appendix illustrate the probability regions for the eight combinations of subject area (i.e., Math, Reading/ELA, Science, and Social Studies) by exception type (i.e., School and Classroom).

The probability regions have jagged boundaries due to the discrete nature of the binomial distribution. The graphs are read by using the school district size of schools or classrooms along the horizontal axis and then by using the observed proportion of exceptions along the vertical axis.

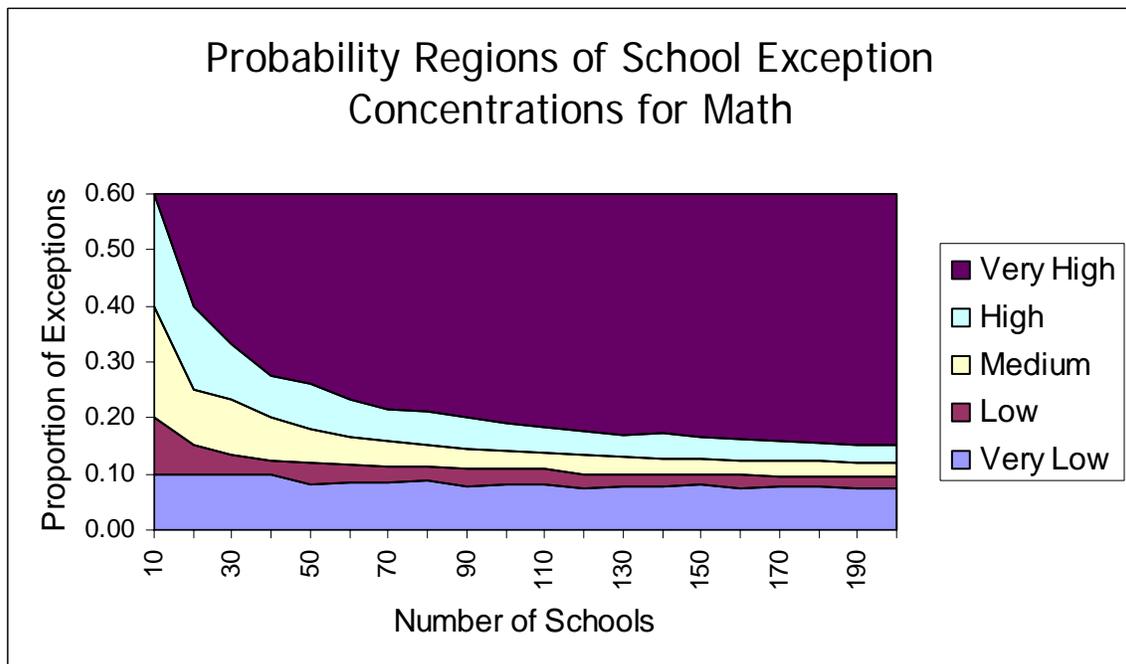
---

<sup>17</sup> In order to create the concentration categories a statistical probability device is used. The reported “probabilities” should not be construed as actual probabilities, since we expect to see clustering and clumping of the data, even if the anomalies are completely random. The statistical device is very useful in identify clusters of anomalies that are tighter than those predicted using the statewide proportions of anomalous results (previously reported in Table 3).

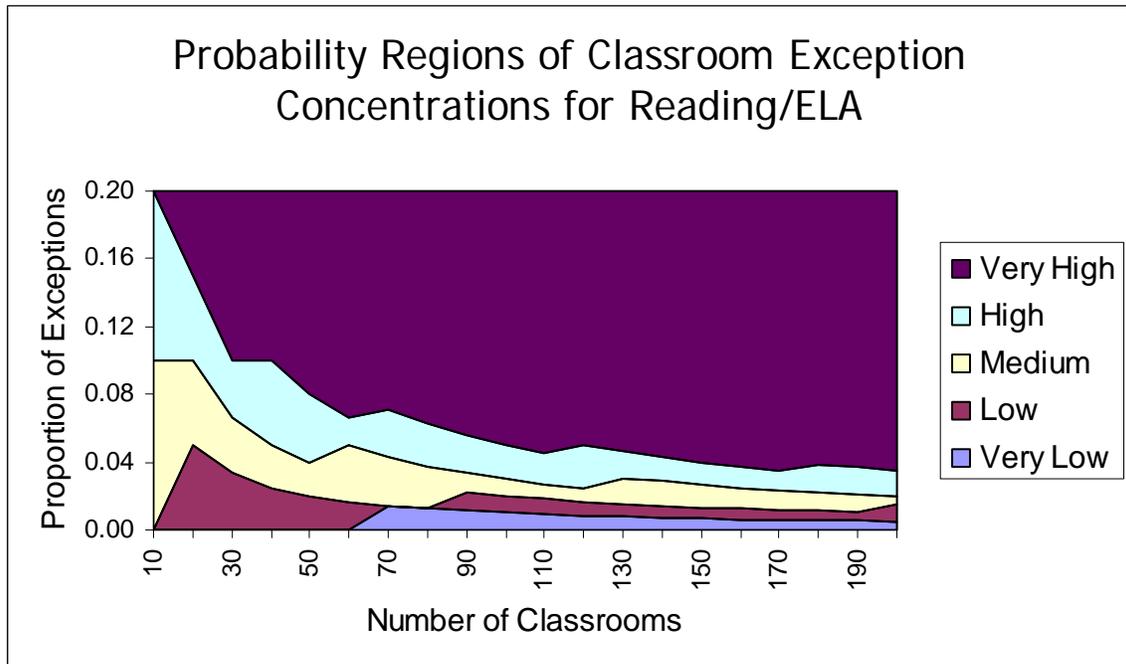
**Figure H-1: Probability Regions of Classroom Exception Concentrations for Math**



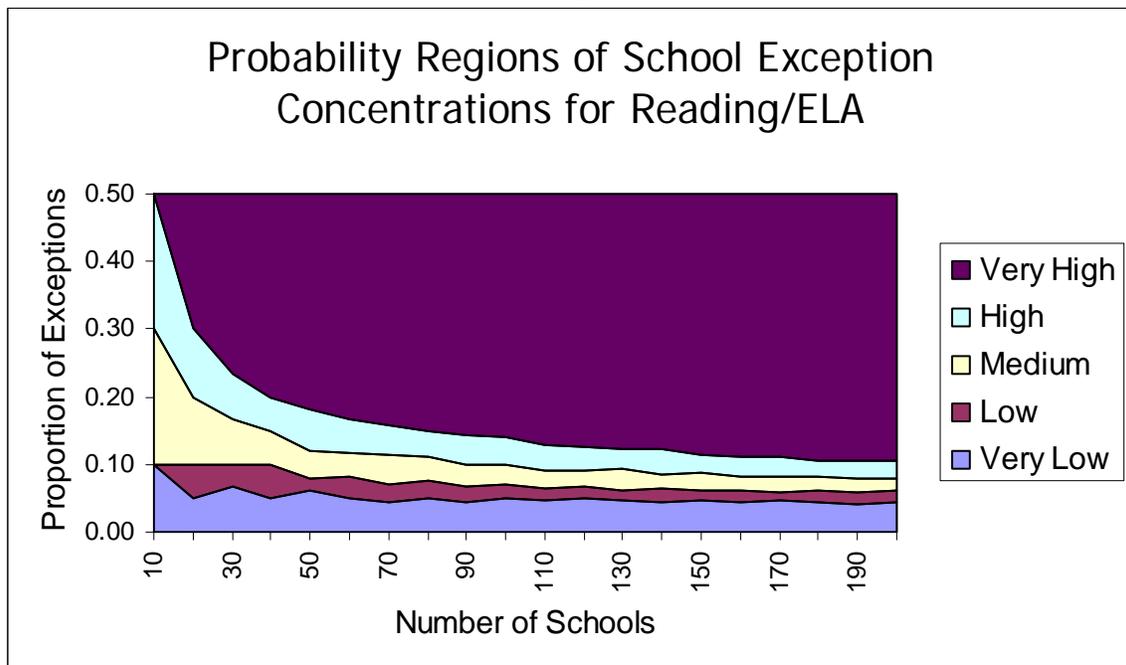
**Figure H-2: Probability Regions of School Exception Concentrations for Math**



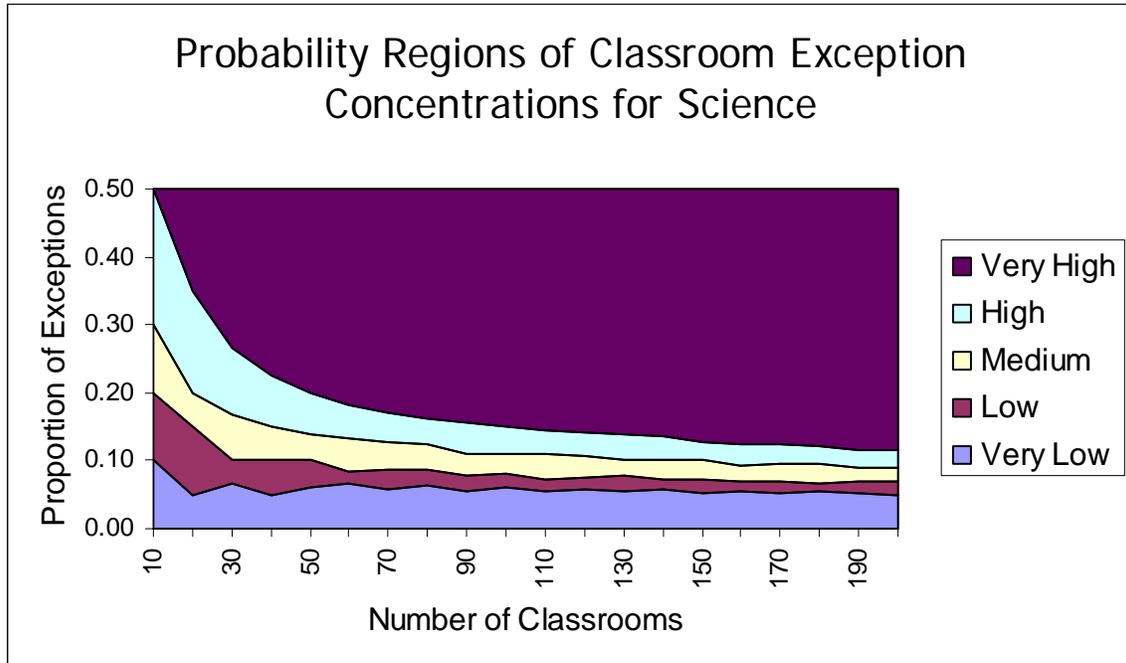
**Figure H-3: Probability Regions of Classroom Exception Concentrations for Reading/ELA**



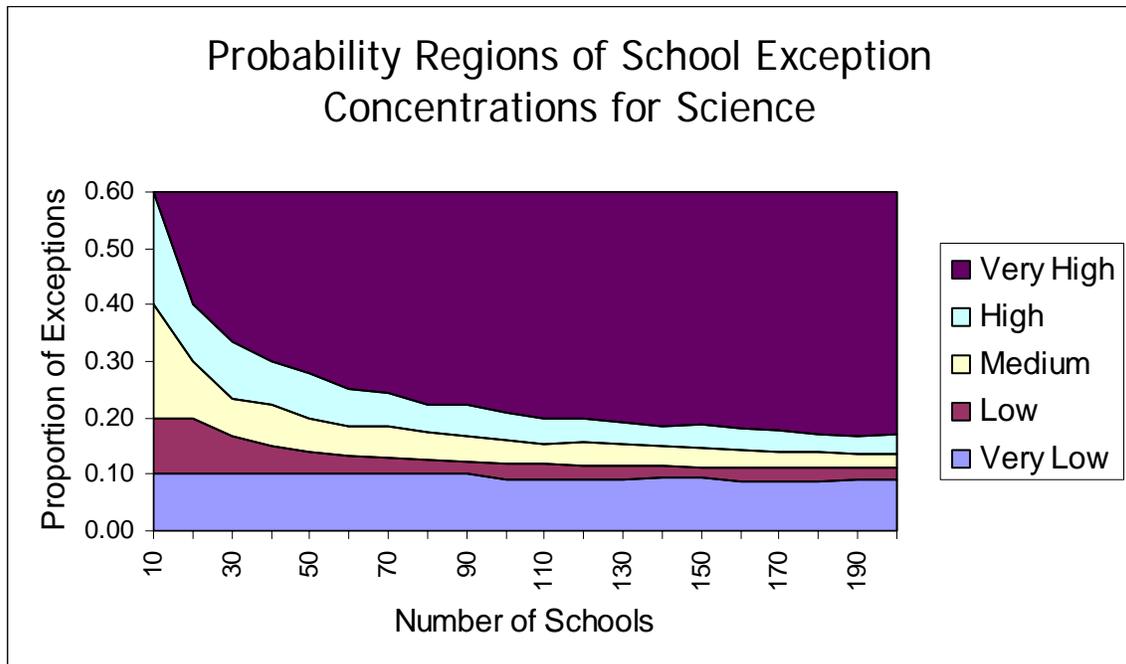
**Figure H-4: Probability Regions of School Exception Concentrations for Reading/ELA**



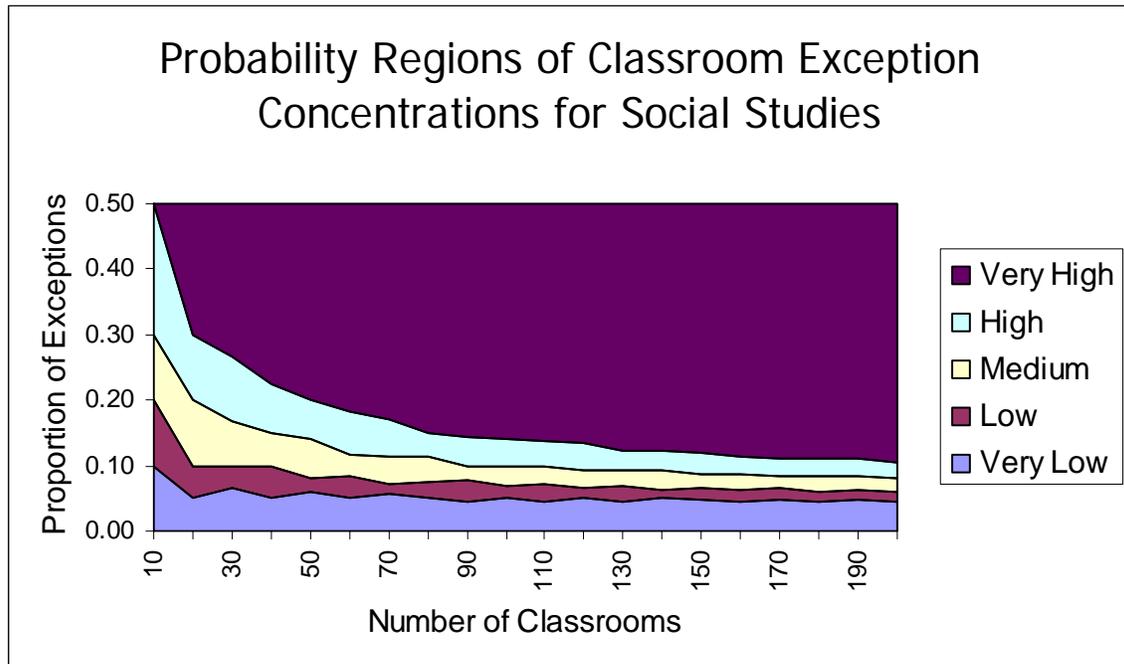
**Figure H-5: Probability Regions of Classroom Exception Concentrations for Science**



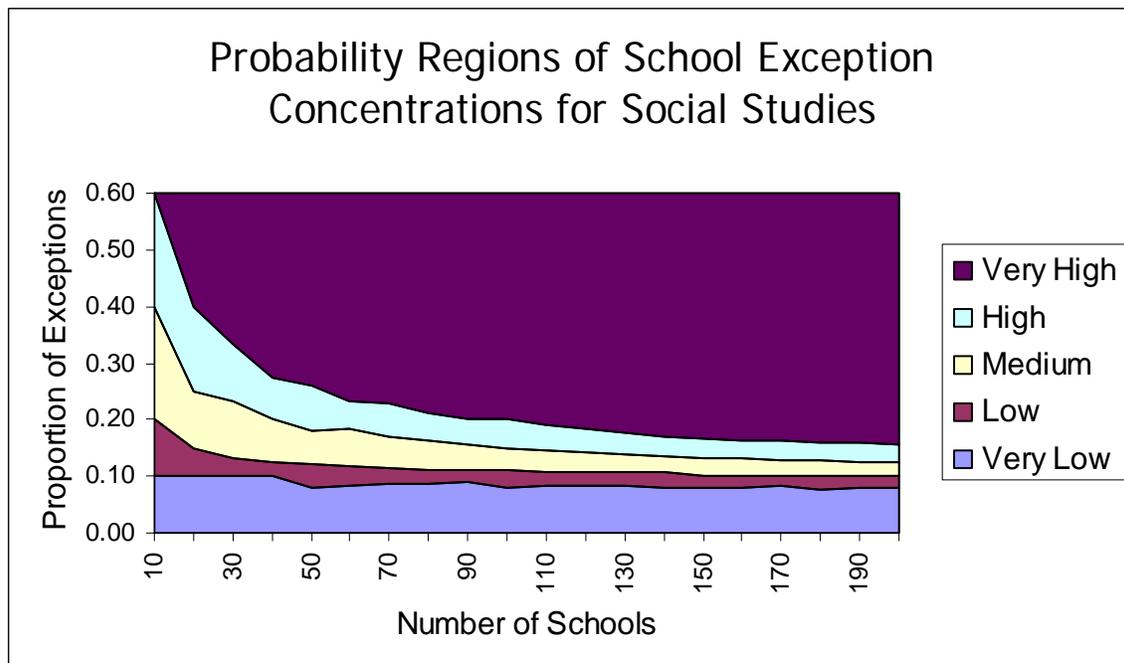
**Figure H-6: Probability Regions of School Exception Concentrations for Science**



**Figure H-7: Probability Regions of Classroom Exception Concentrations for Social Studies**



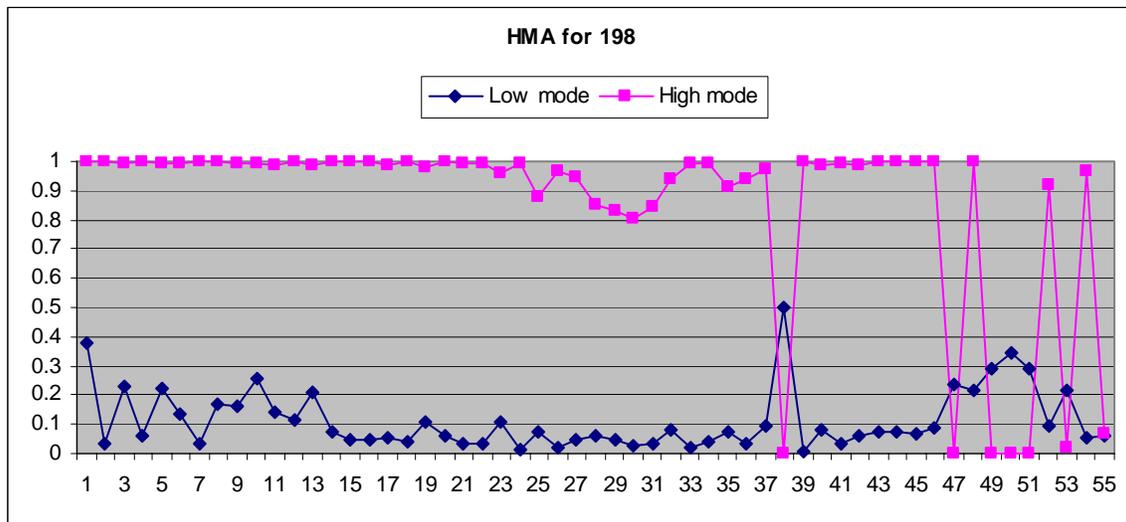
**Figure H-8: Probability Regions of School Exception Concentrations for Social Studies**



## Appendix I – Aberrance Illustrations for Case V

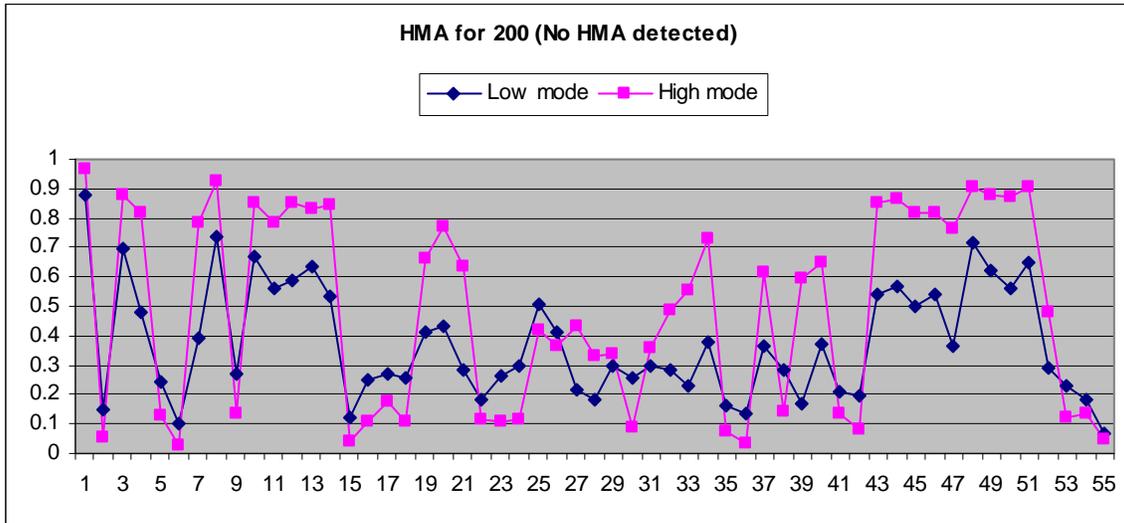
These data are from Case V. There are 9 tests for this class, 6 of which are aberrant. The data were selected to show aberrance in the context of live data. It is instructive to depict the aberrance of the 9 tests. The bimodal model estimates the amount of low-mode and high-mode aberrance on the test. The probabilities of selecting the chosen response for each test item are plotted for the two performance modes. If these probabilities are sufficiently close, then bimodal aberrance is not present. The reader should visually compare the two plotted lines of probabilities. The tests having bimodality show large differences between the probabilities for the test items.

**Figure I-1: Aberrance for 198**



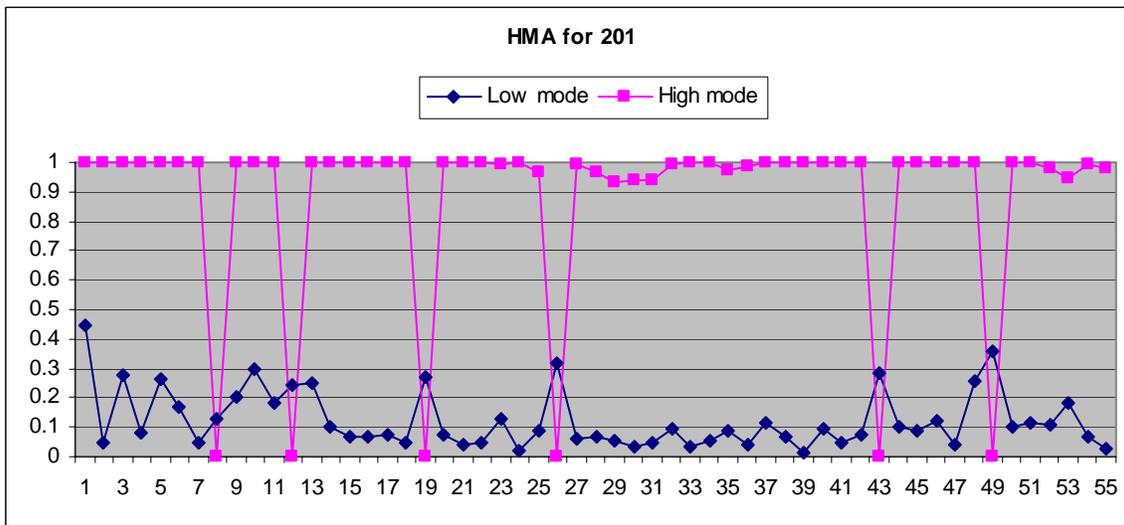
In Figure I-1 above, it appears as if the student became tired or fatigued at the end of the test. It is also possible that the student ran out of time and had to guess on the remaining questions. When the high-mode line dips the student is choosing a very improbable incorrect answer choice. If the probability is very close to zero, given all the other answer choices, then bimodal test taking will be detected. Very difficult questions are shown when the high-mode line does not touch the upper bound of 1, as is seen in questions 27 through 33. The student's scale score was 2376.

**Figure I-2: Aberrance for 200 (No aberrance detected)**



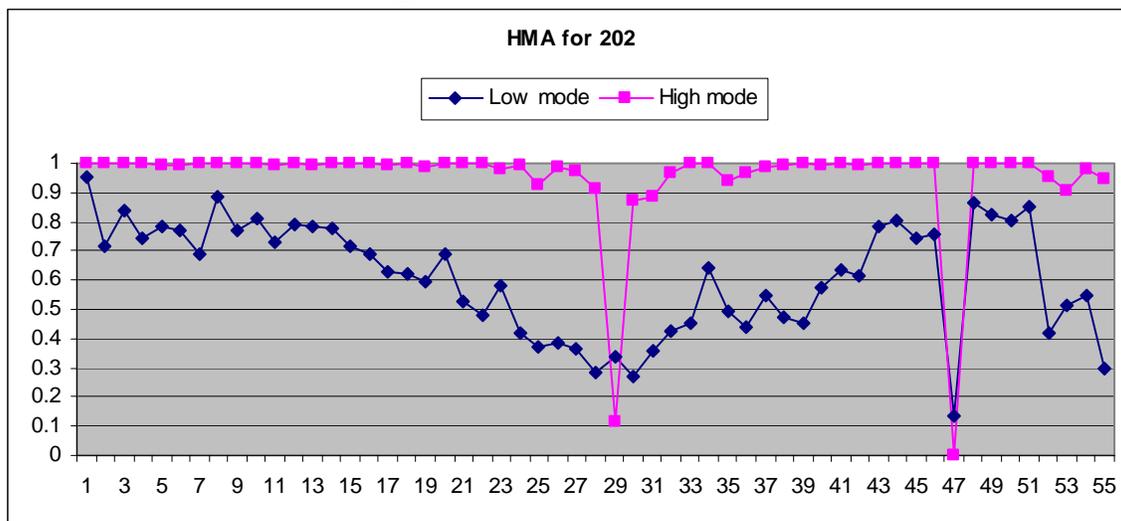
In Figure I-2 above the student appears to not be doing very well on the exam, but aberrance was not detected since the estimated low-mode and high-mode probabilities are very close to each other. This student is answering in a manner that is consistent with demonstrated ability. The student's scale score was 2133.

**Figure I-3: Aberrance for 201**



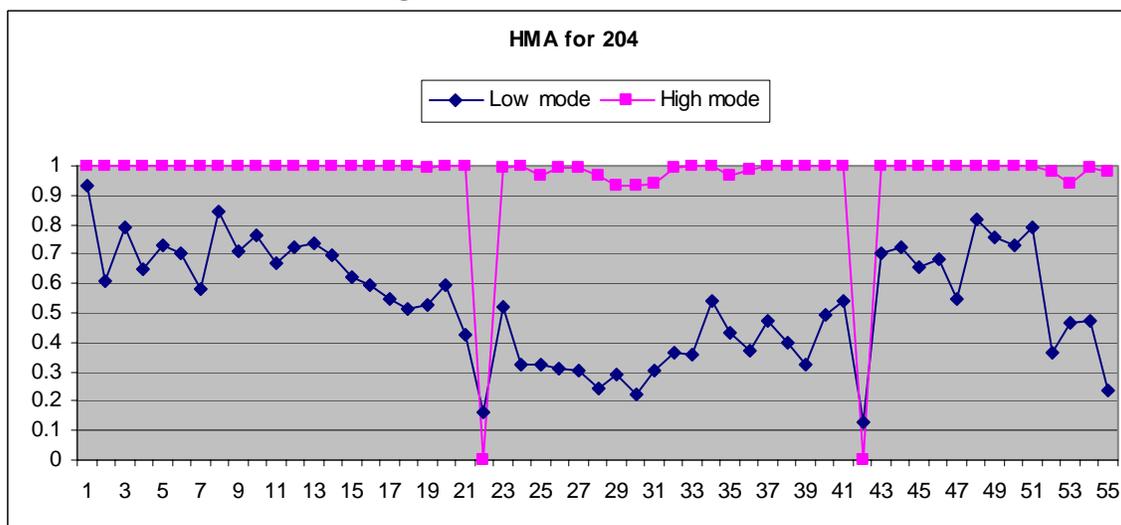
In Figure I-3 above, the student has answered 6 questions incorrectly with an answer choice that was extremely improbable given the demonstrated ability on the other questions. The probability of the incorrect answer choices are above the guessing level and indicate that the student doesn't really know the answers to these questions. This test is the second most aberrant test in this set of 9 tests. The most aberrant test is 219. The student's scale score was 2400.

**Figure I-4: Aberrance for 202**



In Figure I-4 above, the student’s low mode is reasonably high, when compared with the bimodality plots for the other students. The answer to question 47 was very improbable for this student in the high mode. It was even unusual for the student in the low mode. However, even though the answer to question 29, was incorrect, it was a reasonable choice at the high mode. This is the reason that the low mode was somewhat high. It would seem unusual to detect bimodality on the basis of only two incorrect answer choices. Question 47 is a very highly discriminating question. Those who answer this question incorrectly on average score 59%, while those who answer correctly on average score 79%. This 20% difference between the groups means the item distinguishes very well between low- and high-performing students. The above answer looks like a blunder on the part of the student. The student’s scale score was 2557.

**Figure I-5: Aberrance for 204**



In Figure I-5 above the student has chosen two rather low probability answer selections for a knowledgeable student as demonstrated by the other answers, but they are somewhat reasonable for a student who has moderate knowledge. The student’s scale score was 2557.

**Figure I-6: Aberrance for 207 (Low-mode aberrance)**

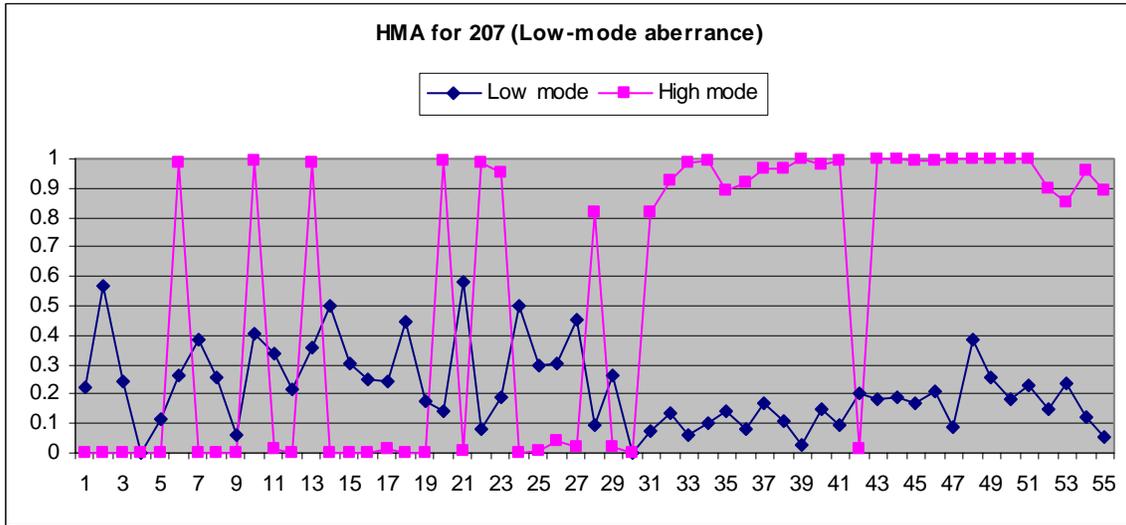
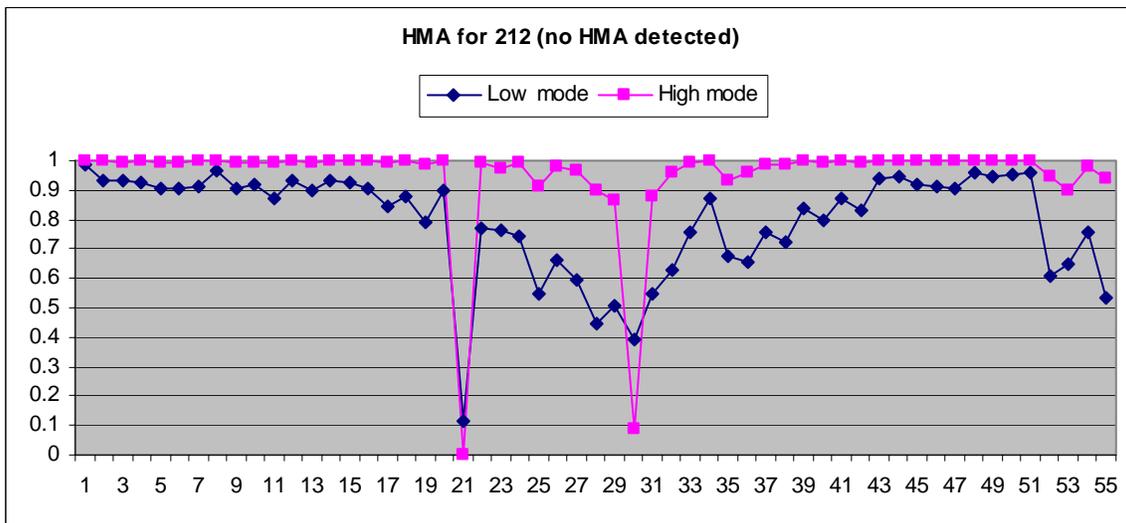


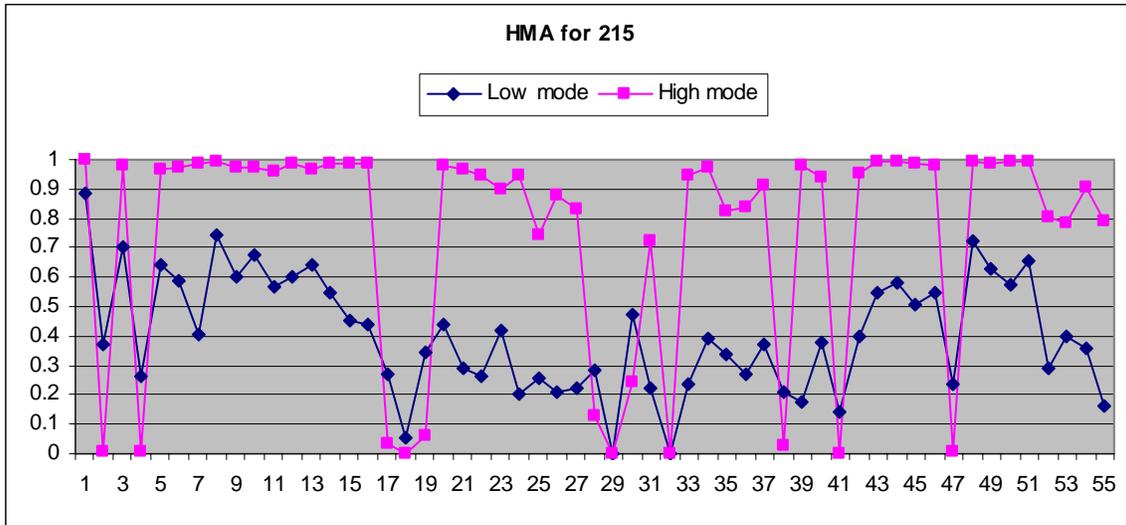
Figure I-6 demonstrates a low-mode aberrant test. Low-mode aberrance is generally associated with poor preparation or at the extreme end, lucky guessing. The student’s response pattern is very unusual, since the student is doing poorly, with “lucky guesses” on the first half of the exam and then completely switches and does very well on the second half of the exam. This does not appear to be a normal test-taking pattern. The student’s scale score was 2133.

**Figure I-7: Aberrance for 212 (no aberrance detected)**



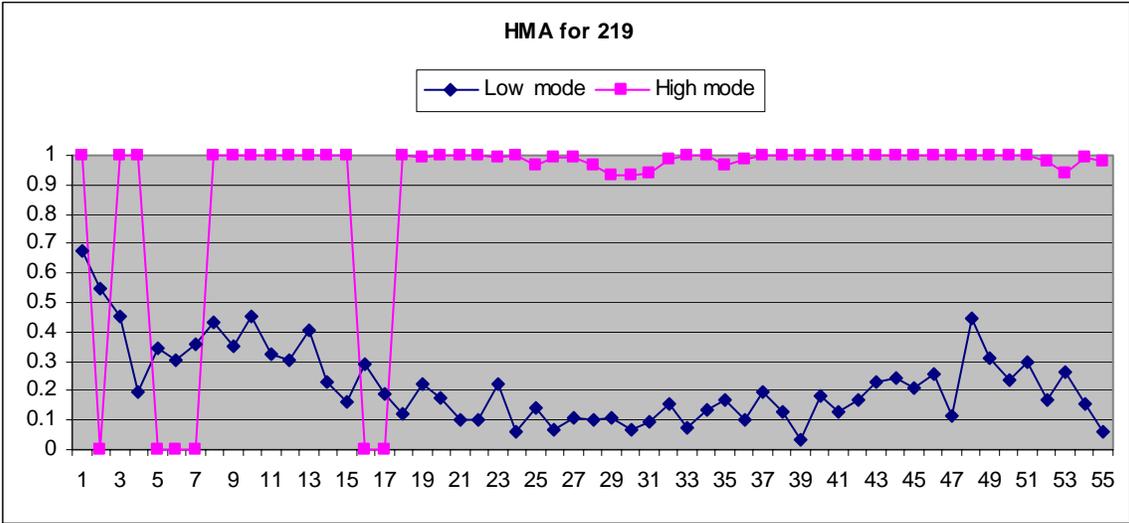
In Figure I-7, the aberrance did not exceed the preset threshold. Therefore, no aberrance was detected. The statewide aberrance rate for the selected threshold on this exam was only 5.6%. It is instructive to contrast Figures I-7 and I-5. In both cases only two questions were missed. In the case for Figure I-7, the selected choices were consistent with a person who would score high instead of very high. For this reason the low-mode and high-mode probability lines are quite close together. The student's scale score was 2557.

**Figure I-8: Aberrance for 215**



In Figure I-8 there are a lot of incorrect answers. Even for a student performing at this level, many of the incorrect answers are quite improbable and indicate that the student's performance is inconsistent on the exam. The score for this student is not a good indication of the student's actual knowledge. This may be a case of receiving some inappropriate assistance. The student's scale score was 2284.

Figure I-9: Aberrance for 219



In Figure I-9, the most extreme case of aberrance in this classroom is shown. This is very extreme because of the large number of improbable answer choices that were made at the beginning of the exam. The student’s scale score was 2400.

## Appendix J – Illustration for Case VI – Highly similar tests

The data in Table J-1 are 186 tests that were found to be highly similar in the analysis of Case VI. Excessive values are highlighted according to the Legend for Table J-1 below.

The important information from this table is the development of the similar test clusters. The size and number of similar test clusters indicate that any testing irregularities that might be present are probably the result of pairs and small groups of students sharing or copying answers. This indicates a possible security breach during the administration of the tests.

### Legend for Table J-1

Test Identifier <sup>18</sup>	If the test was aberrant, the Test Identifier is highlighted in gold.
Score 2004 & 2005	If the test had an unusual gain score, the 2005 and 2004 scores are highlighted in aqua.
Wrong to Right Answer Changes and Other Answer Changes	If the test had excessive multiple marks, the values in the wrong-to-right answer changes and other answer changes columns are highlighted using purple.
Cluster	Clusters of more than two tests are highlighted in alternating yellow and green.
Cluster	Clusters of more than two tests are highlighted in alternating yellow and green.
Similarity Indicator	Identical tests are highlighted in the Similarity Indicator column using red.

**Table J-1: Test Similarity Data for Case VI**

Test Identifier	Grade	Score 2005	Score 2004	Score 2003	Wrong to Right Answer Changes	Other Answer Changes	Cluster	Matches	Similarity Indicator
3281116	10	1948	1871	1864	0	0	2909	3	5.3
3281114	10	1948	2050	2029	0	1	2909	3	6.9
3281109	10	1935	2012	2100	0	0	2909	3	6.9
3281123	10	1923	1931		0	0	2909	3	6.4
3281238	10	2031	1897	1966	0	0	2910	1	2.5
3281134	10	1972	2012	2000	0	1	2910	3	4.1
3281133	10	1972	2029	2100	0	0	2910	1	4.1
3281132	10	1923	1897	1957	1	1	2910	1	1.7
3281149	10	1935	1897		0	0	2911	1	9.8
3281147	10	1853	1823	1761	0	0	2911	1	9.8
3281155	10	2066	2000	2100	1	0	2912	1	5.5

<sup>18</sup> Test Identifiers were assigned sequentially as the data were processed and have no direct association with students, schools or classrooms.

3281170	10	2055	1842	1957	0	0	2912	1	5.5
3281187	10	1923	1931	1895	0	0	2913	1	3.8
3281192	10	1923	1947	1927	0	0	2913	1	3.8
3281208	10	2223	2110	2057	0	0	2914	2	3.1
3281189	10	2179	2077	1983	0	0	2914	3	10.9
3281191	10	2152	1911	1896	2	0	2914	3	10.9
3281200	10	2055	1897	1942	0	0	2914	2	2.6
3281531	10	2019	2050	1949	1	1	2915	1	10.6
3281204	10	2019	2077	2015	0	0	2915	1	10.6
3281497	10	2078			0	0	2916	1	3.0
3281210	10	2066	1980	2000	0	1	2916	1	3.0
3281227	10	1960	2050	2043	0	0	2917	1	5.3
3281215	10	1960	1861	2043	0	0	2917	1	5.3
3281243	10	2179	2145	2176	0	0	2918	1	1.8
3281352	10	2139	2050	2043	0	0	2918	1	1.8
3281288	10	2364	2283	2338	0	0	2919	1	2.8
3281271	10	2256	1897	1878	0	0	2919	1	2.8
3281279	10	2338	2651	2247	0	0	2920	1	3.3
3281274	10	2315	2219	2145	1	0	2920	1	3.3
3281300	10	2055	1914	2115	0	0	2921	1	3.0
3281312	10	2008			0	2	2921	1	3.0
3281330	10	2066	2145	2071	0	0	2922	1	12.3
3281323	10	2031	1823	2130	0	0	2922	1	12.3
3281344	10	2102	2061	1942	0	0	2923	1	10.4
3281340	10	2100	1947	2115	0	0	2923	1	10.4
3281357	10	1882			0	0	2924	1	9.6
3281359	10	1882	1842		0	0	2924	1	9.6
3281371	10	2054	2219		0	0	2925	3	12.7
3281388	10	2054			5	5	2925	3	19.9
3281373	10	2031	1964	1877	2	1	2925	3	19.9
3281368	10	2019	2026	1994	0	0	2925	4	12.7
3281378	10	1910	1803	1911	0	1	2925	1	2.5
3281369	10	2054	2000	2000	0	0	2926	1	3.4
3281370	10	2031	1964	1927	1	0	2926	1	3.4
3281406	10	2256	2283	2193	0	0	2927	1	3.3
3281405	10	2127	2012	2015	0	0	2927	1	3.3
3281533	10	2152	2400	2210	2	0	2928	2	2.4
3281529	10	2054	1980	2071	0	0	2928	3	19.6
3281534	10	2054			0	0	2928	2	19.6
3281535	10	2031	1842	1942	0	0	2928	3	18.5
3561194	11	2243	2205	2050	0	0	3214	1	2.3
3561200	11	2106	1961	1877	1	0	3214	1	2.3
3561211	11	2100	2065	2050	0	0	3215	1	2.1
3561232	11	1993		1914	0	0	3215	1	2.1
3561238	11	2189	2150	2016	0	0	3216	1	11.4
3561236	11	2129	2042	1819	0	0	3216	1	11.4
3561511	11	2289	1973	1931	23	5	3217	6	11.9
3561442	11	2289	2191		1	0	3217	6	11.9

3561433	11	2289	2150	2152	1	1	3217	6	11.9
3561506	11	2289	1885	1819	0	0	3217	6	11.9
3561243	11	2289	1949	2050	0	0	3217	5	3.7
3561259	11	2273	1897	1931	4	2	3217	6	10.7
3561436	11	2229	1996	1931	0	1	3217	5	3.4
3561245	11	2258	2124	2189	0	0	3218	1	6.5
3561248	11	2189	2042	1966	0	1	3218	2	6.5
3561288	11	2164	2065	2152	1	0	3218	1	1.8
3561251	11	2016	2029	2100	0	0	3219	1	2.7
3561249	11	1981	1879	1931	1	2	3219	1	2.7
3561256	11	1958	1885	1877	0	0	3219	2	2.7
3561269	11	2058	2054	1931	0	1	3220	1	7.7
3561272	11	2027	2019	2033	0	0	3220	1	7.7
3561283	11	2215	2163	2152	0	0	3221	3	15.9
3561276	11	2215	1856		0	0	3221	3	15.9
3561273	11	2189	1911		0	0	3221	3	10.4
3561274	11	2106	2077	1877	2	2	3221	3	3.8
3561342	11	2038	2077	2100	0	0	3222	1	2.2
3561329	11	1958			2	0	3222	1	2.2
3561346	11	2164	2112	2016	1	0	3223	1	14.6
3561337	11	2129	2007	1931	0	0	3223	1	14.6
3561402	11	2273	2237	2292	0	0	3224	1	2.1
3561393	11	2129	1823	1949	0	0	3224	1	2.1
3561397	11	2258	2237	2100	0	0	3225	1	2.0
3561399	11	2229	1984	1966	1	0	3225	1	2.0
3561405	11	2015	1924	1877	0	0	3226	1	10.0
3561427	11	2015	1984	1819	0	0	3226	1	10.0
3561410	11	2072	2050	2066	0	0	3227	1	3.8
3561415	11	2027	2077	2050	0	0	3227	1	3.8
3561439	11	2016	2054	2135	0	0	3228	1	3.7
3561448	11	2016	1936	1966	0	0	3228	1	3.7
3561454	11	2129	1996	1931	0	0	3229	1	9.1
3561453	11	2083	2007	1931	0	0	3229	1	9.1
3561464	11	2106		1931	0	0	3230	2	12.9
3561462	11	2038	2088	1931	0	0	3230	2	7.8
3561459	11	2038	1911	1949	1	0	3230	2	12.9
3561498	11	2058	1885	2100	0	3	3231	1	4.2
3561497	11	2016	1996		0	0	3231	1	4.2
2956090	9	2237	2307	2198	0	0	3678	1	1.6
2956730	9	2106	2260	2074	0	0	3678	1	1.6
2956820	9	2281	2400	2249	0	0	3679	1	1.8
2956658	9	2159	2057	2048	0	0	3679	2	2.4
2956804	9	2141	2057	2183	0	0	3679	2	1.8
2956960	9	2100	1980	2127	0	0	3679	1	2.1
2956097	9	2073	2260	2074	0	0	3679	3	2.4
2956542	9	2073	2077	2210	0	0	3679	1	1.7
2956127	9	2330	2364	2311	1	0	3680	1	5.7
2956102	9	2258	2057	1983	0	0	3680	1	5.7

2956109	9	2007			0	0	3681	1	11.6
2956121	9	1957	1995	1983	0	0	3681	1	11.6
2956118	9	2196	2148	2113	1	0	3682	1	1.7
2956694	9	1974	1995	2074	1	1	3682	1	1.7
2956182	9	2106	2039	2010	2	0	3683	3	5.7
2956183	9	2100	2085	2100	0	0	3683	3	5.2
2956181	9	2056	2015	1880	1	0	3683	3	9.4
2956166	9	2007	2100	2036	0	1	3683	3	9.4
2956184	9	2100	2029	2100	1	0	3684	1	1.7
2956168	9	1941	2039	2100	0	1	3684	2	3.3
2956173	9	1834	1861	2029	2	1	3684	1	3.3
2956186	9	1957	2039	2048	0	0	3685	1	1.7
2956502	9	1774	2077	2043	0	0	3685	1	1.7
2956834	9	1907	1919	1896	0	0	3686	1	2.5
2956244	9	1889	1836	2010	0	0	3686	1	2.5
2956393	9	2056	1965	2036	0	0	3687	1	2.1
2956302	9	1957	1935	2023	0	1	3687	1	2.1
2956347	9	2056	1995	2127	0	1	3688	1	12.5
2956355	9	2050	2057	2087	0	0	3688	1	12.5
2956356	9	2073	2148	2087	0	0	3689	1	9.8
2956358	9	2056			1	1	3689	1	9.8
2956373	9	2023	1914	1927	2	0	3690	1	13.1
2956440	9	2023	1818	1927	0	1	3690	1	13.1
2956384	9	2237	2239	2127	0	0	3691	1	2.5
2956824	9	2159	2239	2154	0	0	3691	1	2.5
2956385	9	2123	2200	2231	0	0	3692	1	4.4
2956402	9	2023	1903	2023	0	1	3692	1	4.4
2956400	9	1974	1965	1862	0	0	3693	1	10.6
2956390	9	1907	2015	2010	0	1	3693	1	10.6
2956494	9	1957	2015	1862	0	0	3694	1	2.9
2956405	9	1957	1861	1877	0	0	3694	1	2.9
2956423	9	1853	1887	1956	0	0	3695	2	23.3
2956435	9	1853	1854		0	1	3695	2	23.3
2956408	9	1795		1927	0	0	3695	2	4.2
2956419	9	1889	1854	2036	0	0	3696	1	3.0
2956417	9	1853	1818	1911	0	0	3696	1	3.0
2956427	9	2050	2015	2048	1	3	3697	1	9.8
2956426	9	2007	2015	1927	0	0	3697	1	9.8
2956454	9	1871	1914	1957	0	0	3698	1	3.4
2956455	9	1834	1871	1970	0	0	3698	1	3.4
2956460	9	2196	1947	2000	0	0	3699	1	2.2
2956483	9	2159			0	0	3699	1	2.2
2956495	9	1907	1887	1942	0	0	3700	1	4.6
2956490	9	1871	1919	1912	0	0	3700	1	4.6
2956523	9	2141	2100	1983	0	0	3701	1	1.8
2956532	9	1974	2000	1805	0	0	3701	1	1.8
2956565	9	2050	1965	2048	0	0	3702	1	2.8
2956548	9	1974	2148	1983	0	0	3702	1	9.7

2956575	9	1957	2148	2228	0	0	3702	2	9.7
2956572	9	2100	2219	2183	0	0	3703	1	6.9
2956557	9	2050	2057	1927	0	0	3703	1	6.9
2956580	9	1974	2100	2113	0	0	3704	1	9.3
2956564	9	1924			0	0	3704	1	9.3
2956610	9	2023	1995	1983	0	0	3705	1	2.1
2956587	9	2000	2182	2100	0	0	3705	1	2.1
2956592	9	1957	2015	1997	0	1	3706	1	3.2
2956595	9	1774	1931	1911	0	0	3706	1	3.2
2956598	9	2177	1935	2061	0	1	3707	1	3.5
2956599	9	2100	1919	1970	0	0	3707	1	3.5
2956601	9	2141	2085	2231	0	1	3708	1	3.6
2956602	9	2056	2015	1997	1	2	3708	1	3.6
2956608	9	1941	2069	2023	0	0	3709	1	15.2
2956604	9	1924	2116	2048	3	5	3709	1	15.2
2956618	9	2007	1980	1970	0	0	3710	1	4.3
2956619	9	1974	2100	1896	0	0	3710	1	4.3
2956670	9	2258	2182	2168	0	1	3711	2	3.9
2956664	9	2141	1980	2010	2	0	3711	2	3.9
2956669	9	2123	1995	1972	0	0	3711	2	3.3
2956914	9	2056	2116	2048	0	0	3712	2	3.1
2956751	9	2007	2116	2061	0	1	3712	3	15.3
2956749	9	2000	1995	2048	0	0	3712	3	15.3
2956754	9	2000	1995	2100	1	0	3712	2	4.7
2956811	9	2177	2132	2100	0	0	3713	1	2.1
2956813	9	2177			0	1	3713	1	2.1
2956829	9	1941	1980	1912	0	0	3714	1	9.2
2956854	9	1941	2057	2036	0	0	3714	1	9.2
2956885	9	1941			1	0	3715	1	9.2
2956882	9	1941	2116	2036	0	0	3715	1	9.2

## Appendix K – Glossary of Terms Used in this Report

This glossary was prepared by collecting the terms defined in the text (using footnotes or definition boxes). Other relevant technical terms were included, also.

**Aberrance** - Aberrance in a set of test responses occurs when the student's response pattern on some questions is inconsistent with demonstrated knowledge for other test questions on the exam. The simplest example of aberrance is when the student is able to answer difficult questions correctly, but is unable to answer easy questions correctly. In addition to cheating other atypical behaviors contribute to aberrance. These other behaviors include fatigue, poor preparation, illness, running out of time, lack of motivation, guessing, differential test preparation (knowing some content well, but not knowing other content), and so forth. Hence, aberrance must be interpreted carefully.

**Anomalous** – (See Statistically Anomalous)

**Bimodal test taking** - Bimodal test taking is a form of aberrant test taking. The two modes are recognizable due to the test taker's inconsistency in responses. One mode will be associated with a higher ability level of than the other mode. If the predominant mode corresponds to the higher ability level, then the aberrance is known as high-mode aberrance (HMA). If the predominant mode corresponds to the lower ability level, then the aberrance is known as low-mode aberrance.

**Cheating** - Cheating refers to having and using pre-knowledge of the test content, or receiving unfair assistance in answering the test questions such as through answer copying or answer sharing.

**Classroom** - Classrooms are equated to batches of answer sheets that were returned by the testing personnel at the schools. The organization of batches varies. Generally, the smallest grouping of answer sheets is by classroom or teacher. However, some groupings are by grade or subject area.

**District** - For the purposes of this analysis a district is a unit that is uniquely identified using the district code. Districts could include units that are created for organizational purposes by the TEA which do not correspond to the normal view of a school district.

**Exception** - An exception is detected whenever the overall statistical index (which is a combination of the aberrance, similarity, multiple marks and gain score statistical indicators) is so large that the rate is deemed to be statistically greater than the statewide rate. Larger values of the statistical index correspond to more anomalous observations.

**Excessive Multiple Marks** - Excessive multiple marks occur when an unusually large number of answers on the answer sheet are changed from wrong to right.

Examples of testing irregularities that result in multiple marks are when the teacher helps the student realize that the answers initially chosen are wrong or should be changed to an answer given or suggested by the teacher, or when clean up of “stray” marks on filled-in answer sheets by test administrators includes the erasure and replacement of marks corresponding to incorrect answers in order to raise test scores.

**Extremeness** - A classroom or school is extreme for a particular statistical indicator (e.g., aberrance, similarity, multiple marks, and high gain scores) when the number of tests detected by the statistical indicator is extremely high as compared to the statewide rate for that indicator. The number is extremely high if the probability of the data is less than the experiment-wide alpha-controlled threshold.

**Gain scores** - Gain scores are computed for each student using that student’s scores for prior years. The gain scores are computed using an appropriate statistical model that predicts current performance using prior performance.

**High gain scores** - A high gain score is measured when a student’s test score is substantially higher than predicted based upon prior achievement using an appropriate statistical model.

**High-mode aberrance** - High-mode aberrance refers to bimodal test taking aberrance when the predominant ability mode exhibited by the test taker is the higher level of ability. At times, for the sake of convenience and brevity the term “High-Mode Aberrance” is replaced by the three letter acronym “HMA” in this Report.

**HMA** – (See High-mode aberrance)

**Item Response Theory** - Commonly known by the three letter acronym, IRT, Item Response Theory provides psychometric models for estimating response probabilities at varying levels of examinee ability.

**Low-mode aberrance** - Low-mode aberrance refers to bimodal test taking aberrance when the predominant ability mode exhibited by the test taker is the lesser level of ability. (See Bimodal test taking.)

**Nominal Response Model** - Caveon Data Forensics uses Nominal Response IRT (Item Response Theory) models in order to estimate aberrance and test similarity. These models allow probability computations for all the incorrect answer choices and are critical for establishing probabilities of the similarity indicator.

**Pass rate** - For the purposes of this analysis the term “pass rate” is used to mean “the rate of students who have met or exceeded the TAKS standard.” The term should not be construed to mean that the students have “passed” or “failed.”

School - A school is defined in the data as an organizational unit having a unique identifier of district and school code.

Similarity – Highly similar tests occur when students or educators participate in activities that result in greater similarity between the responses for two or more tests than would be expected if the tests were answered in a statistically independent manner (i.e., statistical independence allows the estimation of similarity between the tests under chance alone). High similarity arises when students copy answers from each other, when answers are changed in blocks so the same set of answers appear across multiple answer sheets, and when forbidden materials that provide answers to one or more of the test questions are displayed or provided in the testing area. This can also occur when students study together in pairs or groups.

Statistical index – The statistical index is a composite index of the statistical indicators for aberrance, similarity, multiple marks and unusual gains. It provides an objective probability assessment of the extremeness of an observation.

Statistical indicator – The statistical indicator provides a mechanism for counting the numbers of test administrations that are related to testing irregularities (such as answer copying and text messaging).

Statistically anomalous - An observation is statistically anomalous when the measured attributes are seen to be extremely different than the expected values for those attributes. A common euphemism to describe anomalous observations is “outlier.” Statistical practice for outlier detection or declaring an observation to be anomalous is usually based upon statistical tests where the probability value of the test statistic is extremely small. In this study, the probability values are approximately 1 in 1 million, or even more extreme, depending on the sample sizes being evaluated.

Teaching the test - “Teaching the test” is being used to indicate inappropriate disclosure of test content to students by educators. Intentional disclosure may be present, but if present it is more likely that test-specific problem formats and problem-solving techniques are being taught to the students.

Test content exposure – Test content exposure results when a test is administered so often or so frequently that the test content becomes well known. Test exposure can also occur when the test content is intentionally divulged by a person who has access to the test instruments or forms. Test coaching or teaching the test occurs when an educator divulges or exposes the test questions by teaching them to the class before the test is given. Another way the content can be exposed is by leaving forbidden materials (such as maps and multiplication tables) on the walls of the classroom where the test is given. Another aspect of exposure occurs when students or parents collaborate on the Internet to disclose the test content.

Testing irregularity – Testing irregularities are events that pose risk of security breach to an exam. Irregularities may occur before, after or during test administrations.

Unusual gains - Unusual gains occur when an unusually large number of high gains are present within a classroom or school. A high gain score is measured when a student's test score is substantially higher than predicted based upon prior achievement using an appropriate statistical model. Unusual gains are very unlikely and may be due to inappropriate coaching, wrong-to-right answer changing, and other testing irregularities. Alternative explanations of unusual gains must always be considered and include excellent teaching and improved access to resources assist students to achieve higher levels of performance.