Caveon Data Forensics Pilot Report

Pilot Analysis of the Spring 2005
Texas Assessment of Knowledge and Skills Test Administration

Prepared by

Dennis Maynes
Chief Scientist

Caveon Test Security

Table of Contents

Report Overview

This report provides Data Forensics analysis of the Spring 2005 Texas Assessment of Knowledge and Skill (TAKS) examinations administered by the Texas Education Agency (TEA). The purpose of the analysis is to identify statistical inconsistencies[1] which may be associated with testing irregularities, such as answer copying, and lax test security.

Definition 1: Testing Irregularity

Test irregularities refer to events where students, classrooms or schools have obtained higher test scores than would have been achieved if the irregularity had not happened. Such irregularities might arise from unfair access to the test content and answer copying, improper collaborative efforts that compromise test security, sharing the test content or teaching the actual test items, or inappropriately changing answers to raise scores.

We emphasize that a statistical association between inconsistencies and testing irregularities does not imply that the identified inconsistencies are due to testing irregularities. Each identified inconsistency must be evaluated on its own merits in determining its cause.

Executive Summary

The analysis was conducted using the results of English/Language Arts (ELA) tests for grades 10 and 11, Math tests for grades 3 through 11, Reading tests for grades 3 through 9, Science and Social Studies tests for grade 11. (Hereafter, these tests are referred to as the TAKS tests, even though an analysis of all TAKS tests was not performed). A summary breakdown of the numbers of tests, classrooms, schools and districts included in the study is provided in Table 1.

Table 1: Summary of Studied Test Results

	Math	Reading/ELA	Science	Social Studies
Students	2,512,890	2,511,433	227,412	229,574
Classrooms	69,661	73,793	2,613	2,624
Schools	7,095	7,112	1,628	1,633
Districts	1,245	1,246	1,100	1,099

[1] A set of one or more test responses is statistically inconsistent when the observed values are seen to be extremely different than the expected values for those values. This report refers to statistical inconsistencies as anomalous testing data or observations. A common euphemism that describes statistical anomalies is "outlier." Statistical detection of outliers and anomalies is based upon statistical tests where the probability value of the observed value is extremely small.

Caveon Data Forensics analyzes test response data using patent-pending methods and systems to identify statistically inconsistent test response patterns from among the total population of test responses. These algorithms detect anomalous patterns which may be associated with different forms of testing irregularities. Besides testing irregularities, the statistical inconsistencies that are identified in this report may be attributed to other sources including "environmental factors" such testing interruptions and student preparation, and "behavioral factors" such as illness and fatigue.

The reader should remember that anomalous data, by definition, are unusual and rare. The inconsistencies that are described in this report represent a very small fraction of all TAKS 2005 test results. We emphasize the point to discourage the reader from using this report to conclude that testing irregularities were widespread during the TAKS 2005 administration. The data do not support this conclusion.

The analyses in this report identify several types of statistical inconsistencies which indicate that testing irregularities may have occurred:

- **Very similar** test responses on different test records with a low probability of occurring by chance. Very similar test responses could indicate that two or more students did not independently answer the test questions.
- **Multiple marks** on answer sheets that generally occur through smudging the answer sheet or erasing and changing answers. Multiple marks (sometimes referred to as "erasures") are measured by the scanning software. An extreme number of multiple marks could indicate that a student's answer sheet was inappropriately modified.
- **A larger score gain** in a student's score than TAKS scores from prior years would predict. An extremely higher than expected score gain could indicate that the student's performance does not measure the student's knowledge.
- **Aberrant or unusual response patterns** that indicate the student's performance on the exam is inconsistent with demonstrated knowledge. Inconsistencies, such as students missing very easy questions while answering the difficult questions correctly, may suggest that the student may have had unfair access to portions of the test content.

Four major findings with associated recommendations are discussed in this report:

1. Statistical inconsistencies were found in a small percent of classrooms[2] and schools. The analyses detected at least one of these anomalies in 1% (702 of 73,793) of the classrooms and 8.6% (609 of 7112) of the schools. This analysis cannot identify the cause of the statistical inconsistencies, although potential causes are discussed later in the report.

[2] Classroom is being used in lieu of the actual batch identifier that was provided by each school. Procedures for assigning batch identifiers vary between the schools. Sometimes the batch identifiers are assigned to all students within the grade. At other times, the batch identifiers are assigned corresponding to teacher. Given this variability in assignment, "classrooms" as used in this report is an approximation of the actual classroom.

2. Of the four types of inconsistencies that were measured, high similarity between the exams was the most prevalent. Several alternative explanations for highly similar exams (including possible answer-copying and sharing of answers between students) are presented later in the report.
3. Statistical inconsistencies due to large score gains were detected in only a very small percentage of classrooms (about 0.3% for Math and 0.1% for Reading). However, the analysis of higher than expected score gains indicated that when a statistically significant large number of students within classrooms and schools experience high gains from prior years to the current year, pass rates within these groups approach 100%.

Definition 2: 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 "Met Standard" level in 2005. The term should not be construed to mean that the students have "passed" or "failed."

4. A greater proportion of statistical inconsistencies were detected in the Math and Science tests than the Reading/ELA and Social Studies tests. High similarity between student tests was the largest contributor to this difference between the subjects.

Caveon recommends that TEA work with test districts across the state to establish an investigative process for the anomalous data findings for the classrooms and schools identified in this Report to the extent that its mandate and resources allow.

To aid in follow up investigations of the analyses summarized in this report, a priority ordering of different statistical inconsistencies and combinations of statistical inconsistencies was developed (see Table 8). According to this scheme, classrooms or schools where multiple statistical inconsistencies were detected should take higher priority in follow-up investigations. In addition, the concentrations of classroom and school exceptions within districts were summarized and classified into five categories (see Figures 12 and 13). Based on these classifications, districts with higher concentrations of classrooms and schools with having statistical inconsistencies should be considered higher priority candidates for further investigation.

Several recommendations for using the results of this report are offered, should TEA or districts wish to conduct follow-up investigations:

1. Use the exception lists for prioritizing the order of investigations.
2. Create summaries of the exceptions for each classroom and school to guide further investigations.

3. Use the similarity and gain score data as indicators of when it may be useful to inspect supporting documents (e.g., answer documents or instructional materials) during the investigations.

To deter testing irregularities going forward, a number of potential initiatives are suggested in this report, to the extent that TEA's mandate and resources allow. These include:

1. Review current security training and improve the security training for teachers and/or other school and district employees as appropriate.
2. Inform the public and school personnel that data is being collected and used in order to detect potential testing irregularities.
3. Assign additional monitors during the test administrations to those classrooms and schools where testing irregularities appear to be most prevalent.

Finally, continued monitoring of test administrations going forward will be valuable in evaluating the effects of actions taken to improve test security, including professional development for teachers, additional monitoring of test administrations, or any of the other actions possible. The Data Forensics analyses are sensitive to change or interventions designed to improve the security situation. That such actions are successful can be easily validated by comparing these analyses on a yearly basis.

Analysis Overview and Synopsis

The primary results of Caveon's Data Forensics analyses are contained in the body of this report, supporting statistical detail is provided in the appendices.

This Caveon Data Forensics analysis consists of four major investigations:

1. **Analysis of Schools and Classrooms**. Schools and classrooms were analyzed to detect statistical inconsistencies. These are provided in the section titled "Review of Schools and Classrooms." These analyses indicate 609 of 7,112 schools (8.6% of the schools), and 702 of 73,793classrooms (1% of the classrooms) where statistical inconsistencies in the test data indicate testing irregularities may have occurred.

2. **Effects on Pass Rates.** Analysis of the effect of potential testing irregularities on pass rates is provided in the section titled "Relationship between Statistical Inconsistencies and Test Pass Rates." This analysis answers the question: "What is the impact on test scores and pass rates when the measured statistical inconsistencies are present?"

   The overall results reveal a complex relationship between pass rates and test results when the measured statistical inconsistencies are present. Pronounced effects of these inconsistencies are related to the classroom and school performance levels. For example, excessive numbers of answer sheets with multiple marks appear to occur more often in lower-performing classrooms than in higher-performing classrooms.

3. **Distributions of Statistical Inconsistencies.** The concentration and distribution of the detected statistical inconsistencies are presented in the sections titled "Distribution of Exceptions within the Schools" and "Distributions of Exceptions within the Districts." These sections are intended to provide guidance for prioritizing the verification of this report's findings.

   In these analyses, statistical inconsistencies within schools were generally found in single classrooms (A "classroom" in some schools could consist of the entire grade. See footnote 2 above.). Very rarely did we find multiple classrooms within the same school having statistical inconsistencies. In 45% (Math) and 55% (Reading/ELA) of the schools where inconsistencies were detected, the anomalous test data was general and not classroom specific.

4. **Case Studies.** Specific examples of the nature of the extremeness of the statistical inconsistencies being detected are provided in the section titled "Selected Case Descriptions of Anomalous Data." These anecdotal descriptions illustrate the application of the various detection methods. Examples include one class where 15 of 27 answer sheets were identical and several of the answer sheets had excessive multiple marks indicating wrong-to-right answer changing; in another class 15 of 15 tests had excessive numbers of multiple marks, with one test having 36 of 40 answers being changed; of the 36 answer changes 32 answers were

changed from incorrect to correct. (These examples were provided because they are easily described. The case studies provide examples of all four kinds of statistical inconsistencies that are analyzed in this study.)

Measuring Statistical Inconsistencies in the Test Responses

Caveon Data Forensics measures four kinds of statistical inconsistencies in the test data that identify potential testing irregularities: aberrant or inconsistent test taking, very similar test responses, excessive multiple marks, and high gain scores.

Very similar test responses are measured by comparing pairs of test answer sheets and computing a statistical index of agreement between the answer sheets that indicates whether the answers are more similar between the tests than would be expected by chance alone. A potential testing irregularity exists when the answer sheets being compared show evidence that the responses were not made independently (i.e., each student doing his or her own work).

Definition 3: Very Similar Test Responses

Very similar test responses are measured when greater similarity between the responses for two or more tests exists 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). Examples of testing irregularities that could result in very similar test responses are when students copy answers from each other, when teachers or test administrators erase and modify answers in blocks so the same set of answers appear across multiple answer sheets, or 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.

Multiple marks are detected and counted during the answer sheet processing. The number of multiple marks is statistically excessive when the probability of the number of observed multiple marks is extremely low. In order to compute this probability Caveon Data Forensics uses the statewide rate of multiple marks to estimate the number of multiple marks that are typical on a filled-in answer sheet.

Definition 4: Multiple Marks

Multiple marks occur when answers are changed on the scan sheet. An excessive number of changed answers from an incorrect to a correct response is a potential indicator of a testing irregularity. Examples of testing irregularities that could result in extreme numbers of 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. Multiple marks can also occur naturally during the test without being irregular. For example, students who originally mark their sheets in the wrong sequence and then discover the error will erase and generate multiple marks.

High gain scores are measured by evaluating the change in a student's test performance relative to other assessments. The presence of large numbers of high gain scores within a classroom or school may indicate the occurrence of a testing irregularity.

Definition 5: Unusual Gains

Unusual gains occur when an unusually large number of high gain scores 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 may indicate a testing irregularity has occurred, such as inappropriate coaching or inappropriate answer changing. Alternative explanations of unusual gains must always be considered and include laudable education efforts such as excellent teaching and improved instructional resources or access.

The gain score analysis fills a much needed gap in detecting testing irregularities that may not be detected by the other measures of statistical inconsistencies that are utilized in this analysis. Large numbers of students with high gain scores are very unusual and only two rational explanations seem appropriate: the instruction is extremely strong or test security has been breached. Thus, statistical inconsistencies that are detected because of unusual gains warrant further investigation on the part of a school district.

The fourth type of statistical inconsistency measured in this analysis is test aberrance.

Definition 6: 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 testing irregularities, other atypical behaviors can 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.

Statistical aberrance in the test response pattern may indicate a testing irregularity. For example, if the student gets help answering some questions and not others, the student's responses may reveal that the test was taken in more than one mode (i.e., the mode of being assisted as well as the mode of working according to one's ability). Under normal circumstances, a student takes the test in a single mode corresponding to his or her knowledge. The bimodal aberrance statistic used by Caveon measures when two test-taking modalities are present in the test responses. Simulations done at Caveon show this statistic is a very powerful detector of test-taking modalities.

Definition 7: 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.

Definition 8: High-Mode Aberrance (HMA)

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.

Technically, the bimodal aberrance statistic measures whether a student is answering some test items at a much higher knowledge or proficiency level than the knowledge or proficiency level at which the other items are answered. Conceptually, if a student misses easy items but gets difficult items correct, the student is demonstrating two modalities.

When the student misses many easy items on the test it appears as if the student has no or little knowledge in the tested subject matter. When the student answers the difficult items correctly it appears as if the student has great knowledge in the tested subject matter. By measuring this statistical inconsistency, the bimodal aberrance statistic may be used to detect potential testing irregularities.

Appendices A and B provide specific illustrations of the Caveon Data Forensics similarity and aberrance statistical indicators.

Performing and Interpreting the Analyses

A very conservative statistical approach is used in performing these analyses. The conservative approach ensures that while not every potential instance of a statistical inconsistency is identified, those that are identified will be so anomalous that reasonable explanations of these inconsistencies by referring to normal circumstances become improbable. This approach strengthens the inference that a testing irregularity may be a likely explanation for such a result. However, a conclusion that a testing irregularity has occurred should not be presumed purely on the basis of the statistical results. The statistics should aid and assist but not guide or replace human judgment or follow-on investigation. Other forms of evidence that confirm or explain the statistical observations should be searched for and obtained, if at all possible.

Although experience has shown that these statistical indicators are potent in discovering patterns of testing irregularity, the reader must remember that alternative factors can also produce similar anomalous observations. Some of these factors might arise from classroom demographic differences such as placing many students with a particular learning disability in the same class or classrooms with students having a distinct difference in cultural orientation towards testing. Environmental factors such as a disruption during the test or perhaps emotional crises such as the death of a close friend or family member can induce aberrant test taking. Factors that contribute to tests with very similar test responses can be quite subtle. These might include intense pre-test review or classrooms that form highly collaborative study groups. Or, commonly held misunderstandings by a group of students may cause many of them to select the same particularly attractive incorrect answers.

An example of the extreme nature of the anomalous data presented in this Report is illustrated by Case II in the section “Selected Case Descriptions of Anomalous Data.” This data illustrates a Math grade 3 class having 15 students. The answer sheets for all 15 students had an excessive number of multiple marks on the answer sheets. The percentage of answer sheets statewide with an excessive number of multiple marks is 1.63% (4,474 of 274,481 answer sheets). In this context it is extremely unusual for 15 out of 15 answer sheets to have excessive multiple marks, the associated probability of this observation is less than 1.0e-25 (this value in scientific notation is less than 1 chance in 10,000,000, …, 000 (25 zeros)). However, when the answer sheets for these 15 students were inspected, it appeared to be a case where students had been taught to take the test by identifying incorrect answers with a “mark” for incorrect responses, gridding the correct answers as they answered the questions and then erasing the marks for the eliminated answers. Thus, this anomalous data appears to be explained by a plausible test-taking strategy followed by the grade 3 students.

In the discussion that follows, classroom and school observed counts are compared using statewide rates. No school or classroom is detected as anomalous unless the probability of the observation is less than 2 in one million for Math and Reading/ELA (exact thresholds are based on sample size; for Math the school probability threshold is 1.413e-6), and less than 6 in one million for Science and Social Studies (for Science the school probability

threshold is 6.173e-6). These probability thresholds are selected so that the probability of observing one or more anomalous observations by chance alone in each subject area is 1 chance in 100 (i.e. the Type I error probability is .01). This conservative approach provides substantial foundation for inferring that the detected statistical inconsistencies are not chance events.

Establishing the Baseline for Assessing Anomalous Observations

This analysis requires the estimation of typical values for the four types of statistical inconsistencies being measured. If every test is taken independently, and if each test can be assessed as excessive or high in each of the four areas (i.e., aberrance, highly similar test responses, excessive multiple marks, and unusual gains) then the counts of observed tests that are excessive in any of the four categories will follow binomial distributions and the binomial proportion can be reasonably estimated using the statewide rate.

The statewide rates for each of the four types of inconsistencies (i.e., aberrance, very similar responses, excessive multiple marks and unusual gains[3]) have been computed by grade level for Math, Reading/ELA, Science and Social Studies (See Appendix C). The rates are computed by accumulating the number of tests which are excessive (i.e., exceed the pre-selected statistical threshold) for each of the four measures in the state at the selected grade level across all classrooms, all schools and all districts, and then dividing by the total number of tests at that grade level. The observed number of tests in each of the four categories within the school or classroom is statistically evaluated using the statewide baseline rates to determine whether the observed number is excessive.

Using the statewide rates in this manner is similar to calibrating a gauge to ensure that it reads properly. Once the statewide rates have been computed the detection of anomalous data follows standard outlier detection methodology where the counts of tests which exceed the threshold of the statistical indicator are assumed to be binomially distributed.

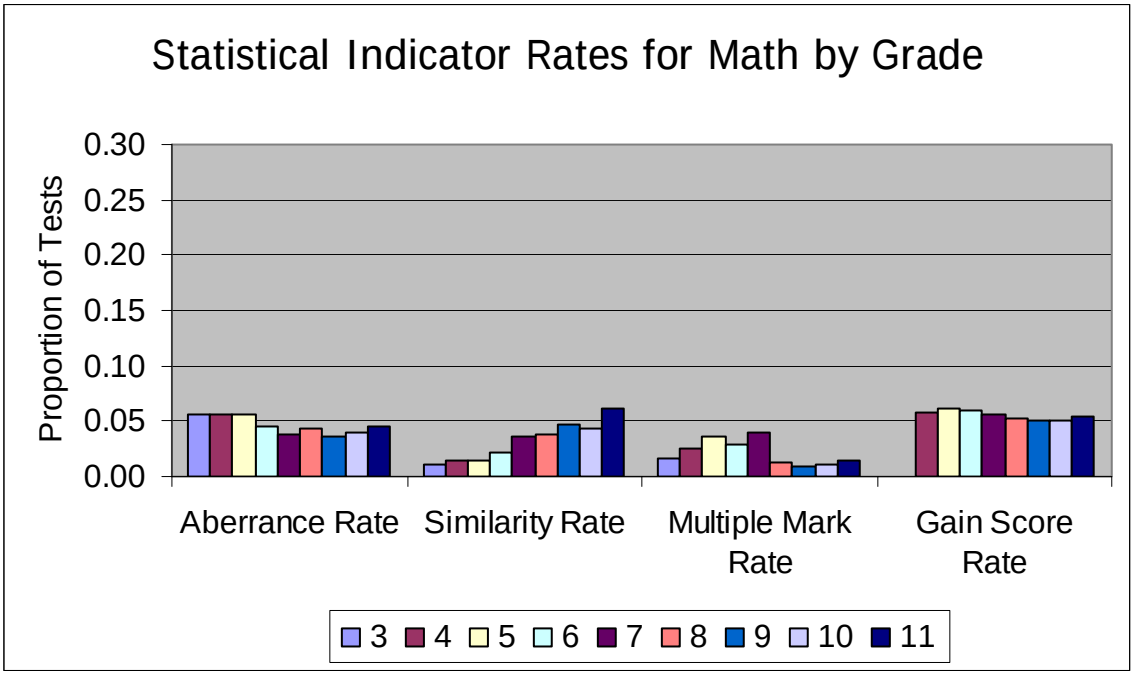
Figure 1 provides the statewide rates of the statistical indicators for the Math test across the 9 grades. Figure 2 provides the statewide rates of the statistical indicators for the Reading/ELA tests across the 9 grades. Figure 3 provides the statewide rates of the statistical indicators for the four grade 11 subjects: Math, Reading, Science, and Social Studies.

Definition 9: Statistical Indicator

Statistical indicator, as used in this report, refers to one of the four statistics (i.e., very similar test responses, high gain scores, high multiple mark counts, and high-mode aberrance) that has been analyzed.

[3] Gain scores are computed for each student using that student's scores for prior years. Gain score rates were not computed for Math and Reading in grade 3 since prior year data is not available. Gain scores for grade 4 are based upon grade 3 scores. Gain scores for Science and Social Studies in grade 11 were based upon grade 10 scores. Gain scores for all other grades and subjects were computed using the prior 2 years of scores, if they were available.

Figure 1: Statistical Indicator Rates[4] for Math by Grade

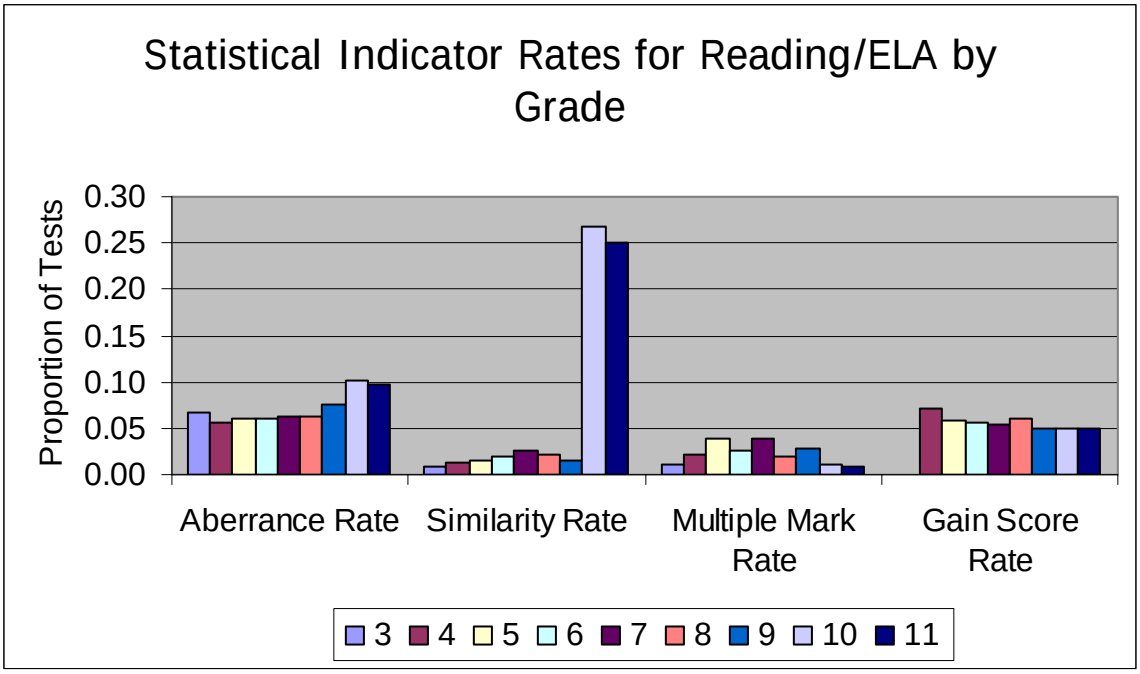


In Figure 1 above, the proportion of Math tests that exceeded the counting threshold for each of the four statistical indicators are presented. The aberrance rates are quite consistent at about 5%. The similarity rates steadily increase through the grades beginning with a rate for Grade 3 of 1% and ending with a rate for Grade 11 of about 6%. The multiple mark rate fluctuates between 1% and 4% across the nine grades. The gain score rate is steady at about 5% for all the grades. There is no gain score rate for Math Grade 3 since gain scores were not computed for this grade (i.e., there are no Grade 2 TAKS tests).

In Figure 2 below, the grade 10 and 11 similarity rates are much higher than would be expected using the pre-determined statistical threshold. These high rates are due to test structure. Technically, the test models assume that the test items are independent. By design and intent, the test items on the ELA tests for grades 10 and 11 are NOT independent because the test items are based upon shared reading passages. In addition, the ELA tests include short-answer and extended constructed response questions. Therefore, this rate is inflated above the nominal rate. This is not a problem for this analysis, since all classrooms and schools are evaluated using these higher-than-expected statewide rates.

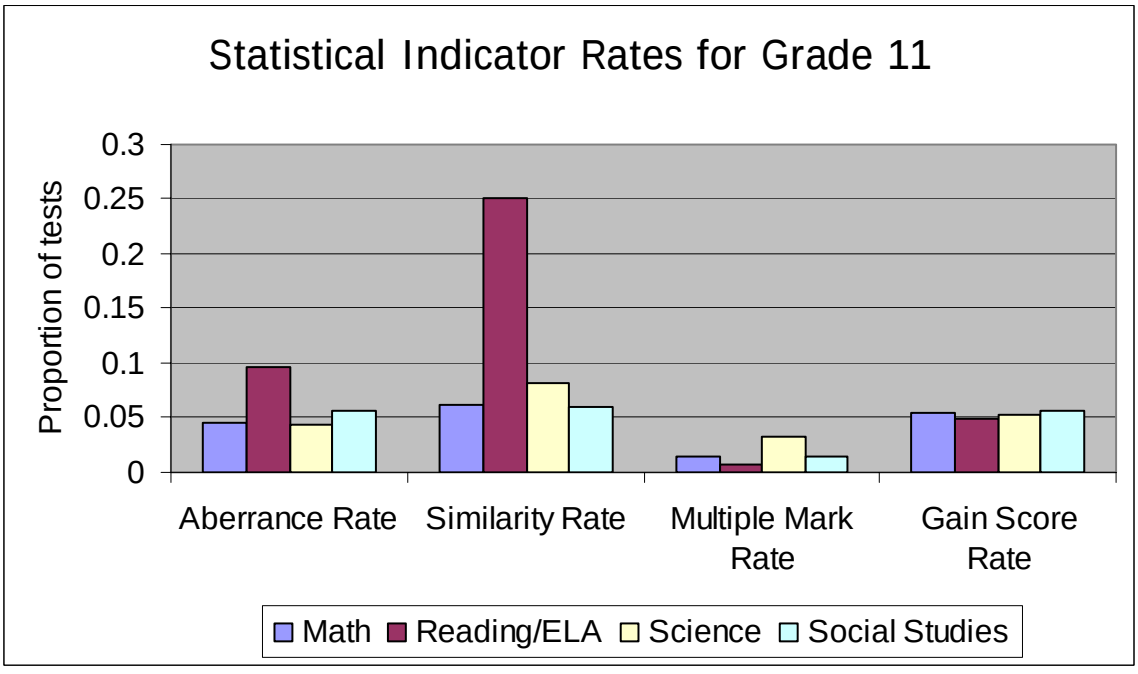
[4] Rates are computed as the proportion of excessive tests in the state for each statistical indicator. The rates are not averages of classroom, school or district rates.

Figure 2: Statistical Indicator Rates for Reading/ELA by Grade



Since Science and Social Studies are only analyzed for grade 11 and not the other grades, the statistical indicator rates for grade 11 are shown in Figure 3 below (The Math and Reading/ELA rates have been repeated).

Figure 3: Statistical Indicator Rates for Grade 11



The underlying statistics of the four statistical indicators for each test were evaluated using the upper 95% probability level. Therefore, if the TAKS data perfectly conformed to the assumed distributions the expected rates of these statistical indicators would be 5%. As can be seen from the figures, the aberrance and high gain score rates are approximately 5%. The multiple mark rate is approximately 2% to 3%. This rate is probably lower than 5% due to tighter variances in the actual distribution than in the assumed distribution. The similarity rates are typically between 1% and 2% for Reading/ELA grades 3 through 9. As previously mentioned, these rates for ELA in grades 10 and 11 are approximately 25%. Similarity rates for Math increase by grade level from 1% to 6%. The similarity rate is 8% for Science, and 6% for Social Studies.

Review of Schools and Classrooms

The goal of this part of the analysis was to identify schools[5] and classrooms[6] that have an unusually high number of statistically anomalous results.

About 1% of the classrooms (702 of 73,793) and 8.6% of the schools (609 of 7,112) have test administrations that are statistically anomalous. Because the tests of hypotheses used in the analysis of schools and classrooms are very conservative it is possible that testing irregularities in a few schools and classrooms have not been identified in this report. This failure to identify testing irregularities (i.e., this is known as Type II error) is a result using a very conservative approach (i.e., maintaining a very low Type I error rate).

A statistical index for each classroom and school in the state has been computed for each of the tests or subjects that were administered in the classroom or school. The statistical index is a combination of the four statistical indicators and provides the standard for determining whether the data from a classroom or school are anomalous (See Appendix D for a technical discussion dealing with creating the statistical index for the schools and classrooms). If the value of the statistical index is extreme, then an exception is generated. Table 2 provides the number and percent of classrooms where the results are anomalous or where an exception was detected.

Definition 10: Exception

An exception is generated whenever the statistical index (which is a combination of the four statistical indicators) is so large that the rate is deemed to be statistically greater than the statewide rate (using a probability value less than 1% and the extreme value distribution of the statistic). Larger values of the statistical index correspond to more anomalous observations.

Table 2: Number and Percent of Classrooms with Exceptions

	Math	Reading /ELA	Science	Social Studies
Percent of Classrooms with Exceptions	0.7	0.3	4.8	4.2
Number of Classrooms with Exceptions	456	252	125	110
Total Number of Classrooms in the State	69661	73793	2613	2624

Table 3 provides the number and percent of schools where an exception was detected (i.e., the data for the school are anomalous).

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

[6] The test administration personnel at each school group the answer sheets into batches and assign each batch of answer sheets a code. Actual classrooms were not identified in the data. The batch header codes are used as surrogate identifiers for classrooms. Batches of answer sheets may be organized by teacher, grade, test administration location, or other groupings. Consequently the term “classrooms” must be interpreted with this understanding.

Table 3: Number and Percent of Schools with Exceptions

	Math	Reading /ELA	Science	Social Studies
Percent of Schools with Exceptions	6.2	4.4	6.9	6.0
Number of Schools with Exceptions	441	314	113	98
Total Number of Schools in the State	7095	7112	1628	1633

Since the statistical index is developed independently for each tested subject area (that is, reading, math, etc.) it is possible for exceptions to be detected in each subject area for a classroom or school. Tables 4 and 5 provide the breakdown of exceptions by subject combination for the classrooms and the schools.

Table 4: Categorization of Classrooms with Exceptions

Subject Combination	Classrooms	Percent of Classrooms
Math only	301	42.9
Reading/ELA only	158	22.5
Math, Reading/ELA	83	11.8
Science only	41	5.8
Math, Science	8	1.1
Reading/ELA, Science	1	0.1
Math, Reading/ELA, Science	0	0.0
Social Studies	23	3.3
Math, Social Studies	8	1.1
Reading/ELA, Social Studies	3	0.4
Math, Reading/ELA, Social Studies	1	0.1
Science, Social Studies	19	2.7
Math, Science, Social Studies	50	7.1
Reading/ELA, Science, Social Studies	1	0.1
Math, Reading/ELA, Science, Social Studies	5	0.7
Total Number of Classrooms with Exceptions	702	

As an example in using the above table, the number 50 under the Classrooms column and for the “Math, Science, Social Studies” combination indicates that 50 classrooms had exceptions in all three of the listed subject areas.

Table 5: Categorization of Schools with Exceptions

Subject Combination	Schools	Percent of Schools
Math only	210	34.5
Reading/ELA only	122	20.0
Math, Reading/ELA	138	22.7
Science only	19	3.1
Math, Science	9	1.5
Reading/ELA, Science	1	0.2
Math, Reading/ELA, Science	12	2.0
Social Studies	14	2.3
Math, Social Studies	7	1.1
Reading/ELA, Social Studies	2	0.3
Math, Reading/ELA, Social Studies	3	0.5
Science, Social Studies	10	1.6
Math, Science, Social Studies	26	4.3
Reading/ELA, Science, Social Studies	0	0.0
Math, Reading/ELA, Science, Social Studies	36	5.9
Total Number of Schools with Exceptions	609	

As an example in using the above table, the number 122 under the Schools column and for the "Reading/ELA only" combination indicates that 122 schools with exceptions only had an exception in the Reading/ELA subject area, but not in any of the other subject areas.

A comparison of Tables 4 and 5 reveals an 11% increase in exceptions where Math and Reading/ELA are both present for schools rather than classrooms. This is reasonable since many teachers, especially those at the upper grades, do not teach both Math and Reading/ELA.

When a classroom or school has an exception there is cause for concern, but that concern may be heightened when there is an exception in more than one subject area. This does not mean that those classrooms or schools are more anomalous than classrooms and schools that have an exception in only one subject area.

The statistical index is created by combining and equally weighting the individual indicators of statistical inconsistency (i.e., aberrant results, very similar test responses, multiple marks, and unusually high score gains). A more in-depth examination of the results reveals the underlying patterns of the statistical index. These patterns provide guidance for formulating the action plan for verifying the anomalous results and determining whether they are due to testing irregularities or other factors. The results are so anomalous that the identified schools and classrooms are very different from other schools and classrooms in the state, but as has been stated previously in this Report, by themselves the statistical results cannot confirm that testing irregularities are present in these classrooms and schools. The results must be tempered with judgment and skill in determining the correct course of action for verification of causes and resolution.

Table 6 provides the breakdown of the exceptions by the combinations of the four indicators of statistical inconsistency for the classrooms having anomalous data. The reader should note that the evidence from a large value on one statistical indicator may be lessened by very small values on the other indicators. However if an indicator value is extreme, an exception will still be detected. On the other hand, if none of the indicator values are extreme, but they are all high, an exception may be detected because their collective evidence is very strong. The classrooms where none of the statistical indicators are extreme, by themselves, are counted in the "None" category in Table 6.

Table 6: Breakdown of Classroom Exceptions by Statistical Indicator Combinations

Statistical Indicator Combination	Math	Reading /ELA	Science	Social Studies
None	25	14	3	3
Gain Score	133	69	18	20
Multiple Marks	66	61	9	1
Multiple Marks, Gain Score	3		1	
Similarity	196	85	77	74
Similarity, Gain Score	6	3	9	4
Similarity, Multiple Marks	10	1		2
Similarity, Multiple Marks, Gain Score	1			
Aberrance	9	5		1
Aberrance, Gain Score	1	10		1
Aberrance, Multiple Marks	1	3		
Aberrance, Multiple Marks, Gain Score				
Aberrance, Similarity	1		4	3
Aberrance, Similarity, Gain Score	4		4	1
Aberrance, Similarity, Multiple Marks		1		
Aberrance, Similarity, Multiple Marks, Gain Score				
Total Number of Classrooms with Exceptions	456	252	125	110

Table 7 (below) provides the breakdown of the exceptions by the combinations of the four statistical indicators for the schools having anomalous data. The interpretation for the categories of Table 7 is the same as the interpretation of the categories for Table 6. The reader is reminded that schools where none of the indicators are extreme, by themselves, are counted in the "None" category in Table 7. A close inspection of the data shows which of the four indicators are the strongest as an aid in evaluating the schools in this category.

Table 7: Breakdown of School Exceptions by Statistical Indicator Combinations

Statistical Indicator Combination	Math	Reading /ELA	Science	Social Studies
None	27	15	2	2
Gain Score	127	78	20	21
Multiple Marks	66	68	9	3
Multiple Marks, Gain Score	2	1	1	
Similarity	168	112	64	63
Similarity, Gain Score	4	2	4	1
Similarity, Multiple Marks	23	18	4	2
Similarity, Multiple Marks, Gain Score				
Aberrance	2	6		1
Aberrance, Gain Score	8	9	1	1
Aberrance, Multiple Marks	4	1		
Aberrance, Multiple Marks, Gain Score				
Aberrance, Similarity	3	2	4	2
Aberrance, Similarity, Gain Score	2	1	4	2
Aberrance, Similarity, Multiple Marks	4	1		
Aberrance, Similarity, Multiple Marks, Gain Score	1			
Total Number of Schools with Exceptions	441	314	113	98

These tables aid in understanding the relationship between the different statistical indicators. Stronger evidence that a testing irregularity may exist is provided when the values of multiple statistical indicators are high.

The data presented in Tables 6 and 7 indicate that aberrance is rarely seen by itself (e.g., by the counts of exceptions classified as “Aberrance”). These data also indicate that similar responses, gain scores, and multiple marks, by themselves account for larger proportions of the detected anomalies than combinations of them. The concepts of excessive multiple marks and unusually high numbers of gain scores as measures of testing irregularity are easily understood. Very similar responding on tests is more difficult to explain, but this measure is indicating more evidence of testing irregularities than any other. It is especially prevalent in the Math tests.

Since more statistical inconsistencies are attributed to the similarity statistical indicator, it is important to understand how similarities in test responses occur. Similar test responses can arise in several ways (not all of which are a result of testing irregularities). These are now listed and explained.

1. Strong teaching – Excellent teaching can cause large numbers of students in a classroom to answer test questions in similar ways. An essential aspect of detecting similarities in test responses is the presence of identical incorrect answers. For example, perfect answer sheets where every question is answered correctly do not suggest a testing irregularity.

2. Collaborative learning – When students study together they learn the same material at approximately the same level. They will also tend to share some of the same misconceptions and have a higher than expected probability of providing identical incorrect answers.
3. Test taking and guessing strategies – When students run out of time or when they don't know the answer, they usually guess. Students that use the same guessing strategy on the same set of questions will score higher on a statistic that measures very similar responding. If test taking strategies are taught to the students so that the same guessing strategies are used by the students in the classroom (or the school), it is likely that similarity rates will be elevated. As an example, if the rule "select C if you run out of time or have no idea" were followed by groups in a class or school, the tests could be classified as extremely similar.
4. Disclosure of answers – When the teacher discloses the answers to the test content (either intentionally or inadvertently), then similar responses may be measured, especially if students are confused and answer several of the questions incorrectly in the same way. If the students do not really know the answers to the questions that have been disclosed, aberrance may also be detected.
5. Test coaching – When specific portions of the tests are coached so that a large number of the students do better than expected on blocks or sections of the test, then very similar responding may be detected.
6. Answer sheet modification – If the answer sheet is tampered with (i.e., by using an eraser) so that the same answers commonly appear on multiple answer sheets then similarity in the test responses will be detected. The multiple marks analysis may also detect that an unusual number of answer sheets within the classroom have been changed inappropriately. Such tampering often produces aberrance since inconsistencies in the test response patterns will be introduced.
7. Test design – The way the test is constructed may inadvertently induce similarity effects, especially if the test items are not statistically independent. However, because the similarity statistic was determined by taking an appropriate sample, common test design effects will show increased statewide similarity rates. This means that the similarity detection incorporates test design effects that induce similar test responses.
8. Answer copying and text messaging – When students are able to copy from each other, or when they are able to work together on answering the test questions, very similar test responses may be detected. If students successfully use text messaging to share answers then similarity will likely be detected.
9. Crib sheets – If students work from a common source that provides test answers then similarity may be detected. If the test forms are made available and accessed before the test, students or educators could create shared crib sheets.
10. Exam exposure – When the test content becomes well-known to the educators throughout the state, statistical similarity will likely increase as a result of "teaching the test." Teaching the objectives that are measured on the test is expected and desired. But, if the test content itself is taught so that students are able to score well on the exam without having mastered the material, then this could aptly be named "teaching the test." Normal security efforts should be taken to ensure that all test booklets are accounted for and never copied by unauthorized

personnel. This is especially important if the test forms are to be reused in a later test administration.

In summary, when unusual instances of similar tests are found, it is important to determine the cause. Extensive analysis of other data sets indicates that such groups of very similar tests are closely associated with inappropriate increases in test scores and pass rates.

Ranking the Statistical Indicators for Detecting Testing Irregularities

The different statistical indicators vary in their ability to detect testing irregularities. All are important and each brings a unique contribution to the Data Forensics analysis, however once the anomalous data has been detected an understanding of the quality of the detection guides the prioritization of verification and investigation.

The overall statistical index that guides this analysis is an excellent method for ranking the anomalous results since this index measures overall statistical extremeness or inconsistency. Each statistical indicator contributes equally to the statistical index. However, for the purposes of verification it may be desirable to take into account the quality of the statistical indicators.

Table 8 provides a priority schedule which weights the statistical indicators by their detection quality. When more than one statistical indicator has detected an anomaly, confidence increases that a testing irregularity occurred. Table 8 uses a simple weighting scheme[7] to rank order the combinations of statistical indicators to derive a suggested priority schedule for verifying and determining causes of the anomalous data.

Table 8: Suggested Priority Schedule in Descending Priority Order for Verification of Anomalies

Combination	Priority
Aberrance, Similarity, Multiple Marks, Gain Score	14
Similarity, Multiple Marks, Gain Score	12
Aberrance, Multiple Marks, Gain Score	11
Aberrance, Similarity, Gain Score	10
Multiple Marks, Gain Score	10
Aberrance, Similarity, Multiple Marks	9
Similarity, Gain Score	8
Similarity, Multiple Marks	7
Aberrance, Gain Score	7
Aberrance, Multiple Marks	6
Gain Score	5
Aberrance, Similarity	5
Multiple Marks	4
Similarity	3
Aberrance	2
None	0

The remainder of this section discusses the relative strengths and weaknesses of the statistical indicators. This discussion provides guidance for making decisions about using the statistical evidence from the indicators in verification and investigative efforts.

[7] The weights that are used in Table 8 are: 2 for Aberrance, 3 for Similarity, 4 for Multiple Marks and 5 for Gain Scores. As an example the weight for the "Aberrance, Multiple Marks" combination is 2 + 4 = 6.

Each statistical indicator provides information about a potential testing irregularity. The indicators vary by directly or indirectly measuring testing irregularities. Direct measurements are usually higher quality than indirect measurements. The statistical indicators vary in the complexity of the statistical distributions that are required. Those indicators with complex distributions are usually of lower quality than statistical indicators having simpler probability foundations.

Unusual Gains

The unusual gains statistical indicator evaluates whether the observed number of students with high gains is significantly greater than the expected number using the statewide baseline. The purpose of education is to increase student knowledge and proficiency; therefore high test score gains do not constitute a testing irregularity. However, large numbers of high test score gains indicate that a testing irregularity may have occurred.

Since actual scores are measured and compared, this statistical indicator is based upon direct observations of test results. As a result, it is a high quality indicator that is supported by the expected rate of high gain scores observed throughout the state.

Excessive Multiple Marks

The excessive multiple marks statistical indicator evaluates marks on scan sheets that indicate an answer has been changed. More evidence that an irregularity may have occurred is given when the marking analysis indicates the answer was changed from wrong to right. The number of other multiple marks on the answer sheet is used as a counter-balance so that a student who makes a lot of multiple marks (through erasing or smudging) is not necessarily counted as having an excessive number of multiple marks. The presence of multiple marks on the answer sheet does not constitute a testing irregularity, since occasionally students inadvertently misalign the answers on the scan sheet and upon discovery of the error erase the filled-in answer marks and make corrections if time allows. Some students are instructed to place a check mark next to answers where they are unsure, so they can review their answers if they have time. If a substantial number of students in the classroom or school make multiple marks in this manner, then the classroom or school would be observed as anomalous. An inspection of the answer sheets will be instrumental in determining causes of the anomalous data.

This is a moderately high quality statistical indicator in detecting testing irregularities. It is known that changing answers from wrong to right raises test scores, therefore this statistical indicator is directly associated with testing irregularities where the answer sheets have been inappropriately modified. Because of assumptions concerning the statistical distributions of the multiple marks, the implementation of this statistical indicator for the TAKS analysis is statistically conservative.

Similarity

The similarity statistical indicator assesses the similarity of the student-selected item responses between two tests. This indicator tests the assumption that the tests were taken independently (or, that each student's work is independent). Similarity is computed by comparing the responses from every answer sheet with every other answer sheet in the

school. The number of identical correct answers and identical incorrect answers are evaluated using an appropriate statistical distribution. This distribution relies upon the standard psychometric assumption that item responses on the test are statistically independent. However, the indicator is robust to the violation of this assumption because statewide similarity rates are used as the baseline.

Simulations show that the similarity statistical indicator is very powerful in detecting non-independent test taking. A finding of non-independent test taking must be carefully reviewed since many factors contribute to test response similarities. For a listing of some of those factors see the above discussion in the "Review of Schools and Classrooms" section.

This is a moderately high quality statistical indicator in detecting testing irregularities.

Aberrance

The aberrance statistical indicator postulates that bimodal test taking is responsible for inconsistencies in the test responses. There are many explanations of test-taking inconsistency or test response aberrance. And many sources of test response aberrance, such as guessing, poor preparation and fatigue, are always present. If these sources of test-taking inconsistency are relatively constant throughout the state, the statewide aberrance rate will incorporate these sources. The aberrance statistical indicator measures excessive amounts of inconsistent test taking. The data show that this statistical indicator is working quite well in this regard; aberrance is strongly related to increased pass rates.

This is a moderate quality statistical indicator because it is based upon a postulated statistical model that approximates an ideal model of cheating (i.e., having pre-knowledge of test items or answers). Its importance increases when other statistical indicators are extreme. Simulations show that this indicator is substantially more powerful than popular person-fit indices when used to detect bimodal test taking.

Relationship Between Statistical Inconsistencies and Test Pass Rates

The statistics employed by Caveon Data Forensics are specifically designed to detect statistical inconsistencies that may be associated with testing irregularities. In the analyses presented in this section, pass rates in classrooms where statistical inconsistencies were detected ("extreme classrooms") are compared with pass rates for classrooms where statistical inconsistencies were not detected ("non-extreme classrooms"). For the purposes of this analysis, the pass rate is defined to be the proportion of students at or above the TAKS "met standard"level in the subject area for the test that was taken based on 2005 test results. These analyses do not consider the TAKS "advanced" level, although such analyses could be carried out.

Definition 11: Extreme Classroom

A classroom is extreme for a particular statistical indicator (e.g., aberrance, similarity, multiple marks, high gain scores) when the number of tests detected by the statistical indicator is extremely high as compared to the statewide rate for that statistical indicator. The number is extremely high if the probability of the observed data is less than the experiment-wide alpha-controlled threshold (which is sample size dependent; see Appendix D).

The data reveal complex associations between the test score distribution and the statistical indicators. The grouping of classrooms into "extreme" and "non-extreme" categories allows a more insightful analysis of the effect on the pass rate by potential testing irregularities. The data discussed in this section are tabulated in Appendices F and G. The following summaries of these data seem to be appropriate findings:

1. Aberrance has a large effect in Math (20% increase in pass rates for extreme classrooms), moderate effects in Reading/ELA and Social Studies (5% increase in pass rates for extreme classrooms) and a negative effect in Science (6% decrease in pass rates for extreme classrooms, but within extreme classrooms there is a 12% increase in pass rates when tests with and without aberrance are compared).
2. Similarity is more prevalent among lower-performing classrooms in Math and Reading/ELA. The effects within extreme classrooms are large for Math (12% increase in observed pass rates) and Science (17% increase in observed pass rates). The effect within extreme classrooms is moderate for Social Studies (7% increase in observed pass rates). It is slight for Reading/ELA (1% increase, but examination of Figure F-6 (in Appendix F) shows that in the non-extreme group the pass rate for very similar tests is 6% lower than the pass rate of tests that are not very similar. This reversal of differences is an interaction and the total pass rate swing is 7% = 1% + 6%).
3. Multiple marks appear to be more prevalent in lower-performing classrooms. The pass rates of classrooms where multiple marks are detected are 3% to 6% lower than classrooms where they are not detected (for Math, Reading/ELA and Social Studies). There is a slight positive effect in Science (2% increase in pass rates for extreme classrooms). Generally, multiple marks appear to be infrequently

associated with potential testing irregularities. However, a few statistical inconsistencies in the multiple marks data provide indications that some testing irregularities may have occurred.

4. Gain Scores have an extremely large effect on the pass rate for all subjects (e.g. Math, Reading/ELA, Science and Social Studies). In nearly all classrooms where large numbers of high gains are detected the pass rates approach 100%. This would be expected given the definition of gain scores.

The Math and Science subject areas show more evidence of increased pass rates associated with the statistical inconsistencies than the Reading/ELA and Social Studies subject areas. When present, unusual gain scores have a strong association with increased pass rates in all subjects. Test response similarities exhibit larger pass rate effects in Math and Science than in Reading/ELA and Social Studies. This is probably due to the nature of the curriculum and the way in which these subjects are taught.

Multiple marks and similar tests appear to be more prevalent among lower performing classrooms. The data do not suggest why multiple mark and similarity rates appear to be higher in lower performing classrooms. Only by examining what happens in the classrooms where testing irregularities have potentially occurred can one obtain an accurate assessment of the effects of these types of potential testing irregularities.

The complex relationship between the statistical inconsistencies and pass rates merits in-depth analysis. This analysis is provided for the Math tests in Appendix E. The figures for all four subject areas are found in Appendices F and G.

Distribution of Exceptions within the Schools

The analysis of school data in this Report focuses on school-wide rates in determining whether the measurements for a particular school are anomalous. In this regard it is a parallel analysis to the classroom analysis and it disregards the classroom exceptions. The analyses of schools and classrooms are not independent because statistical inconsistencies that exist within classrooms may be detected also when the schools are analyzed. An example of this occurs when all the answer sheets within a school are processed as one batch, resulting in one "classroom" even though many teachers are present in the school. In this case the school and classroom data will consist of the same set of tests.

Some statistical inconsistencies may result from campus-wide activity. In these cases, classroom exceptions will not be reported. These campus-wide statistical inconsistencies may result from different sources than classroom-specific statistical inconsistencies and they need to be measured and assessed. Table 9 categorizes the school data by exceptions and subject areas.

Table 9: Schools by Number of Classrooms with Exceptions

School Category	Number of Classroom Exceptions	Math	Reading /ELA	Science	Social Studies
No school exception	None	6586	6743	1513	1534
	1	66	55	2	1
	2	2	0	0	0
	3 or 4	0	0	0	0
	5 or more	0	0	0	0
School exception	None	199	174	10	11
	1	154	101	95	76
	2	57	30	5	7
	3 or 4	26	6	1	3
	5 or more	5	3	2	1

In the table above the school counts by subject area are not equal for Math and Reading/ELA because the number of schools in the data having Math and Reading/ELA tests differed. Each cell in the table provides the number of schools at the particular combination. For example, 101 schools that administered the Reading/ELA exam had an exception and they also had one classroom with an exception. The data highlight the phenomenon that classroom exceptions may exist within a school and the school does not have an exception. For example, 66 schools administered the Math test without a campus-wide exception but one classroom in the school was detected with an exception.

The salient observations to be drawn from the data in Table 9 are:

1. Nearly always classroom exceptions result in campus-wide exceptions (this is especially true with multiple classroom exceptions),

2. Exceptions for multiple classrooms are rarely seen in Science and Social Studies (this is probably due to batch processing Science and Social Studies tests from schools in a single batch),
3. Exception rates are higher for Math than Reading/ELA, and the numbers of multiple classrooms with exceptions in Math are substantially higher than the numbers of multiple classrooms with exceptions in Reading/ELA, and
4. Only about half the campus-wide exceptions in Math and Reading/ELA are attributable to classroom exceptions.

These data suggest that ordering the schools by the number of classrooms where exceptions were detected will provide a useful guide for further verifying the exceptions and statistical inconsistencies that are reported in this analysis. That is, those schools where exceptions were detected in multiple classrooms might be investigated first.

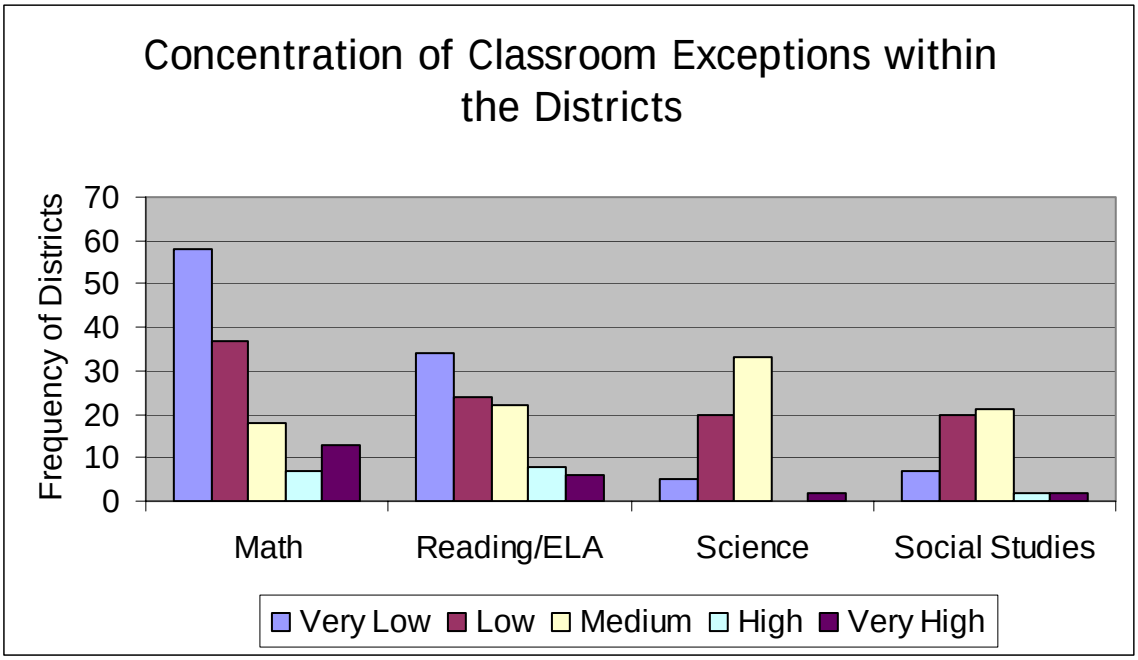
Distribution of Exceptions within the Districts

The primary focus of this analysis is the evidence of potential testing irregularities on the TAKS tests provided by detected statistical inconsistencies within the analyzed classroom and school data. However, this analysis would be incomplete without describing the distribution of classrooms and schools within the districts where anomalous results have been found.

The presentation in this section provides guidance for identifying districts with higher than expected numbers of classrooms or schools having anomalous test administrations. The disparity in district sizes precludes using straight percentages as a measure of degree or numbers of exceptions to determine districts having high numbers of anomalous results. Therefore, district sizes must be used, in addition to the percent of classroom or school exceptions within the district, in determining where high concentrations of exceptions have been detected.

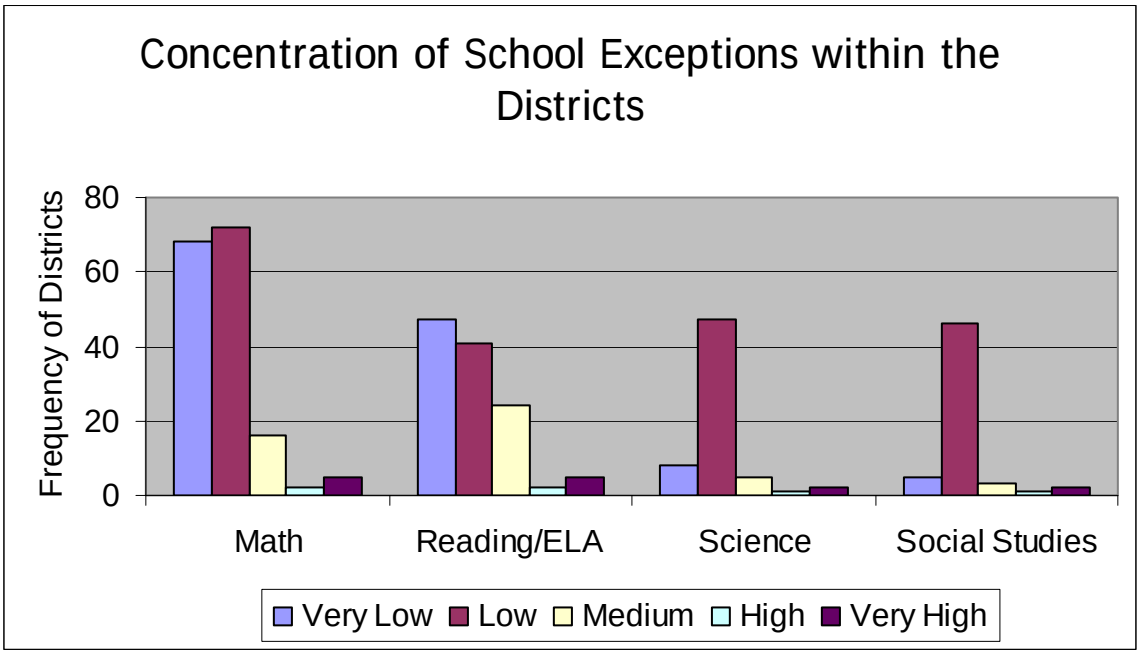
Figure 12 provides a histogram of districts where concentrations of classroom exceptions are present. The districts are classified into five concentration categories: very low, low, medium, high, and very high. These categories were determined using probability computations described in Appendix H. The computations compare the observed number of exceptions (based upon the district size) against the expected number of exceptions within the district.

Figure 12: Concentration of Classroom Exceptions within the Districts



Most of the districts with classroom exceptions are listed in the very low to low concentration categories. For verification purposes of the report findings, Caveon recommends that districts in the medium, high and very high categories be reviewed first.

Figure 13: Concentration of School Exceptions within the Districts



These data should be used as an aid in determining where the school and classroom exceptions are concentrated or clustered. The data have not been assembled with the purpose of categorizing districts by prevalence of testing irregularities. Such an analysis is beyond the scope of the current investigation.

Data Forensics Findings and Recommendations

Potential Testing Irregularities exist in a Small Percent of Schools and Classrooms

Caveon's analyses have detected statistical inconsistencies (which may be due to testing irregularities) in a very small percent (less than 1%) of the classrooms (702 of 73,793). Statistical inconsistencies have been detected about 8.6% of the schools (609 of 7112). This higher detection rate for the schools is an artifact of three elements. First, anomalous test administrations in any classroom usually result in the detection of anomalous test administrations in the school. Second, the larger sample sizes that are present in the schools increase the power of statistically detecting test administration anomalies, if they exist. Third, large sample sizes result in higher rejection rates of statistical hypotheses tests, when the tested value is statistically significant, but not practically significant.

Statistical evidence alone is insufficient to conclude definitively that testing irregularities have occurred. However, the conservative statistical approach used in this analysis has reported .7% of the Math (456 of 69,661), .3% of the Reading/ELA (252 of 73,793), 4.8% of the Science (125 of 2613), and 4.2% of the Social Studies (110 of 2624) classrooms in the state as having statistical inconsistencies that might be associated with testing irregularities.

The lists of classrooms and schools from this analysis provide a starting point for further investigations. The lists should not be used to estimate the number of testing irregularities that might have occurred in the state. Besides the presence statistical inconsistencies there is no evidence that testing irregularities have occurred in the listed classrooms and schools. If testing irregularities have occurred they are more likely to be found within these lists than by sampling the general population of classrooms and schools.

Based on this analysis, testing irregularities within the state appear to be isolated. They may have occurred during the TAKS 2005 Spring administration in a relatively small number of schools and classrooms.

Caveon recommends that TEA work with test districts across the state to establish an investigative process for the anomalous data findings for the classrooms and schools identified in this Report to the extent that its mandate and resources allow. This process should provide districts with guidelines for following up results based on the prioritization suggestions outlined in this Report.

In addition, additional security training for the personnel who handle and administer the TAKS tests may be warranted, as outlined in the security report recently completed for TEA by Dr. Greg Cizek.

Test Similarity is the Predominant Statistical Inconsistency

Statistical similarity was detected in nearly half of the identified anomalies (e.g., exceptions for schools and classrooms). Similarity of answer sheets results can be caused by collusive behavior on the part of students (e.g., answer copying). It could also be

caused by some portions of the exams becoming known prior to administration or by unauthorized answer sheet modification (see Case VII for a possible example) or other forms of testing irregularities.

Caveon recommends that the test form security be reviewed to ensure that the exam security is not breached and that copies of the exams are not made available to students before the test administration. For some schools and districts, more stringent monitoring of the test administration and control of test materials before during and after testing may prevent unauthorized copying of test forms or memorization of content.

Since test similarity is the predominant statistical inconsistency, Caveon also recommends asking districts where test similarity exceptions were detected to specifically address potential sources of answer-sharing and answer-copying in the schools and classrooms in follow-up investigative work. Placing additional attention on these security threats may help to deter them in the future. In addition, detailed analysis of the answer sheets where similarities were detected may yield clues as to the source of the test similarities (e.g., by examining the handwriting and manner in which the answer bubbles are filled).

When Gain Score Exceptions are Detected the Rate of Students not meeting the TAKS Standard Decreases Dramatically

While high year-to-year gain scores are not evidence of testing irregularities, the detection of statistical inconsistencies due to unusual gains, in at least a few cases, useful in identifying classrooms where additional explanations for the anomalous data should be sought. There appear to be only two explanations for unusual gains: excellent instruction or testing irregularities.

One finding of note in this analysis was the association of extreme gain scores with higher pass rates in all subject areas. The numbers of classrooms where statistical inconsistencies due to high gains have been detected are approximately .3% (three-tenths of one percent; 219 of 69,661) of the Math classrooms and .1% (one-tenth of one percent; 109 of 73,793) of the Reading classrooms. These percentages indicate that such testing irregularities, though potentially potent, are not widely observed throughout the state.

Caveon recommends that the TEA encourage districts to carefully review all detected statistical inconsistencies that are due to unusually high gain scores.

Math and Science are Displaying More Statistical Inconsistencies than Reading/ELA and Social Studies

The Math and Science subject areas show more evidence of increased pass rates associated with the statistical inconsistencies than the Reading/ELA and Social Studies subject areas. When present, unusual gain scores have a strong association with increased pass rates in all subjects. Test response similarities exhibit larger pass rate effects in Math and Science than in Reading/ELA and Social Studies. This may be due to the nature of the curriculum and the way in which these subjects are taught.

Caveon recommends that School Districts review the Math and Science curriculum emphasis in those schools where statistical inconsistencies were found to ensure that deliberate or inadvertent teaching of the test content is not occurring.

Using the Results and Findings of This Report

Caveon has provided several analyses in this Report, all of which are useful in identifying both classrooms and schools where testing irregularities may have occurred. This section suggests several ways for TEA to use the information in this Report to decrease the potential of testing irregularities.

1. Finding confirmatory evidence or alternative explanations through investigation. This Report provides direction and information that can strengthen test security. Some actions that TEA or the school districts could take are:
    a. Use the exception lists for prioritizing the order of investigations.
    b. Create summaries of the exceptions for each classroom and school to guide further investigations.
    c. Use the similarity and gain score data as indicators when it may be useful to collect supporting documents during the investigations. Supporting documents include test administration and test form tracking logs, answer sheets (for multiple mark verifications), classroom materials that focus on test preparation (e.g., sample classroom assessments, handouts the teachers have given to the students), and seating charts from the Spring 2005 test administrations.
2. Taking preventative or remedial action. This Report indicates the nature of testing irregularities that may exist for the TAKS tests and can be useful in guiding actions and implementing plans that strengthen test security. Some actions that TEA or the school districts could take are:
    a. Review current security training and improve the security training for teachers and other school, district and state employees. The training should result in decreased potential for testing irregularities.
    b. Inform the public and school personnel that data is being collected and used in order to detect potential testing irregularities. Knowing that intentional testing irregularities are discoverable may deter those individuals who are concerned about being caught or those who were unsure if their actions were illegal, unethical, or contrary to existing legal agreements.
    c. Assign additional monitors during the test administrations to those classrooms and districts where testing irregularities appear to be more prevalent. This may involve simply a restructuring of the existing monitoring program (if there is one) to switch resources from where they are less needed to where they are seriously needed.
    d. In cases where test scores appear to have been inflated inappropriately, it may be appropriate to retest the involved students so that current teachers have more reliable information about student performance.
    e. In some cases where exceptional performance was commended the decision for commendation may need to be reconsidered, if the exceptional performance was attributable to testing irregularities.

f. In extreme cases, and after careful verifications, more stringent disciplinary measures may be warranted. Such disciplinary actions need substantial evidence beyond what is available in this Report.

3. Evaluating the effectiveness of test security decisions from year to year. One of the more valuable uses of the information in this Report would be to evaluate the effects of any test security actions taken, including professional development for teachers, additional monitoring of test administrations, or any of the other actions possible. The Data Forensics analyses are sensitive to change or interventions designed to improve the security situation. That an action works can be easily validated by comparing these analyses on a yearly basis.

More In-Depth Investigations

Several additional investigations (data mining the data used to generate this Report or collecting and analyzing other relevant data) by TEA could enhance the findings contained in this Report by providing additional perspective on the security and integrity of the TAKS tests.

1. Perform subset analyses to investigate the effects of factors such as grade levels, classroom sizes, school demographics, etc. Each of these factors could have a bearing on test security and integrity may help identify potential risk factors for testing irregularities.
2. Evaluate the numbers of students by district, school and classroom who take the TAKS tests against the number of students who take alternative grade-level assessments or who take no grade-level assessments at all. Such an analysis would uncover any attempts to raise test scores by assigning students to take other assessments.
3. Develop clusters of similar tests to measure the potential prevalence of student- or teacher- organized cheat rings. If desired, Caveon can assist in this investigation by using already developed algorithms that generate clusters tests with similar responses.

Selected Case Descriptions of Anomalous Data

In this section some of the more anomalous situations that are observed in the data are described. The reader is forewarned that these cases exhibit very extreme and anomalous results.

Case I – Multiple Statistical Inconsistencies in an 11th Grade Math Class

These data were selected because they are the most extreme data (i.e., its ranking is 1) for the Math test in the entire state based on the overall statistical index (90.9307 in this case). The classroom data are from an 11th grade Math test. There are 91 students in this classroom (or for this batch of answer sheets). Being the most extreme case, the data exhibit multiple statistical inconsistencies. The rates of the various statistical indicators for this classroom are compared with the statewide rates in Table 10.

Table 10: Anomalous Data for Case I

	Grade 11 Math Baseline	Classroom	Annotation
N	225984	91	
Pass Rate	0.8096	1	The normal pass rate is 81%. This class has 100%
Mean Score	2203	2261	The average score for this class is 58 points above the state
Statistical index		90.9307	Very unusual: p < 1.0e-90
Mean w/o Incidents	2191	No Data	All tests had incidents in this class.
Mean w/ Incidents	2269	2261	Every test in this class is anomalous in some way.
Aberrance Rate	0.0449	0.5495	55% of the tests are aberrant, only 4% should be.
Similarity Rate	0.0611	0.978	98% of the tests are very similar, only 6% should be.
Multiple Mark Rate	0.015	0.033	The number of multiple marks is within expectation.
Gain Score Rate	0.0546	0.4894	49% of the tests where gains scores were computed were very high gains, only 5% should be.

In the above data, 98% of the tests are very similar to at least one other test in the school. This is very unusual and an in-depth look at the data shows some tests as "highly similar" to as many as 80 other tests in the class. 82 of the tests are "highly similar" to at least 10 other tests in the class.

When the gain scores are viewed, it is seen that the lower performing students are experiencing very unusual gains. Only 47 gain scores were computable, but of those one-half or 23 had gains greater than 1.645 standard deviations (upper 95% point for the normal distribution). There should have only been 3 on average. The average gain for this group of 24 students is 307 scale score points, from 1972 to 2279. The average gain for

the remainder of the class is 41 scale score points, from 2215 to 2256 (where gain scores can be computed).

In this group of tests there are 7 clusters of identical answer sheets. These identical answer sheets are driving the extreme similarity that is seen in the data. The scale scores on the identical answer sheets are: 2365 (2 tests), 2344 (7 tests), 2324 (3 tests), 2306 (3 tests), 2289 (2 tests), 2289 (2 tests), 2289 (2 tests). There are a total of 21 tests that have identical answers with at least one other test in the class. The probability value[8] that these identical answer sheets occurred by chance is so small as to approach the realm of impossibility.

It is difficult to say what has caused these anomalous data. The presence of multiple statistical inconsistencies seems to indicate that normal alternative explanations for these data are unlikely. However, viable explanations, other than one or more testing irregularities, may exist for these data; and those explanations should be sought.

Case II – High Numbers of Multiple Marks in a Grade 3 Math Class

These data were selected because of the high numbers of multiple marks on the answer sheets and the smallness of the classroom. This case study illustrates that Caveon Data Forensics is powerful in detecting potential testing irregularities in both small and large samples. These data are anomalous because every test had an excessive number of multiple marks indicating large numbers of wrong-to-right answer changes. These classroom data are from Math grade 3. There are 15 students in this class. This classroom data ranks in 54th place when the data are ordered using the statistical index for the 456 anomalous Math classrooms. When the anomalous classrooms are ordered by the multiple marks statistical indicator these data are the most extreme for the entire state. Generally, across the state, we are not observing large pass rate changes because of excessive wrong-to-right answer changes; however these data show that sometimes multiple marks are strongly associated with increased scores. Table 11 gives the summary comparison with the statewide baseline.

[8] There are very few instances of identical answer sheets in the TAKS data. These are extremely unusual and in order to assess how unusual they are, ten million random samples of 91 tests were drawn from the entire set of Math grade 11 tests without replacement. The largest observed duplicated count was 8 duplicate tests, giving an approximate probability of 1 in 10,000,000. There were only two observed samples with 7 duplicated tests. These data had thirteen duplicates and the probability that these data occurred by chance alone is much smaller than one in ten million. Using a Poisson approximation, the desired probability value is less than 1 in 10 to the 21st power ($p < 1.0e\text{-}21$).

Table 11: Anomalous Data for Case II

	Grade 3 Math Baseline	Classroom	Annotation
N	274481	15	
Pass Rate	0.8193	1	The pass rate is 100% and it should be about 82%. The probability of a 100% pass rate for this size group would be .05, which is not excessive.
Mean Score	2246	2375	The mean score is 130 points above the state average.
Statistical index		25.5393	This is a very small p value < 1.0e-25
Mean w/o Incidents	2248	No Data	All tests had incidents in this class.
Mean w/ Incidents	2223	2375	Every test was flagged with an incident in these data.
Aberrance Rate	0.0565	0.2667	This rate is high, but not extremely high.
Similarity Rate	0.0115	0	There were no tests that were highly similar in these data.
Multiple Mark Rate	0.0163	1	Every test had excessive multiple marks. This is extreme.
Gain Score Rate	N/A	N/A	There are no gain scores in the third grade.

This is such a small data set and it is so intriguing that it is worth illustrating with individual student results. The essential aspects of the data are shown in Table 12.

Table 12: 15 Tests with Excessive Multiple Marks

Scale Score	Raw Score	Wrong to Right	Other	Multiple Mark Indicator[9]
2400	0.93	32	4	50.0
2337	0.90	25	6	50.0
2261	0.85	23	3	47.3
2337	0.90	21	1	42.0
2714	1.00	21	2	42.0
2565	0.98	20	0	39.5
2565	0.98	19	1	37.0
2337	0.90	17	4	32.1
2202	0.80	15	4	27.4
2337	0.90	15	3	27.4
2230	0.83	12	1	20.7
2261	0.85	12	2	20.7
2714	1.00	10	1	16.4
2261	0.85	6	2	8.6

The normal rate of multiple marks observed on the Math Grade 3 answer sheets is 29 wrong-to-right answer changes in 1,000 item responses and 18 other answer changes in

[9] The multiple mark indicator values of 50 correspond to such improbable data that the probability computations returned a value of 0, meaning the precision used in the computation could not represent such a small value.

1,000 item responses. If students occasionally find answers to change, then we would expect on the 40-question test to observe only .12 wrong-to-right answer changes and .07 other answer changes. So typically we would expect to see one answer being changed for every 5 answer sheets.

In the above data we are seeing truly anomalous numbers of multiple marks. How would a student know that 32 of the selected wrong answers needed to be changed? The probability of observing this many wrong-to-right answer changes is less than one in 10,000, …, 000 (49 zeros).

An inspection of these answer sheets was performed and the marks and erasures on the sheets were consistent with students taking the test who had been instructed to eliminate incorrect answers by "marking" them, to grid in the correct answer, and then to erase the "marks" that were made to indicate the eliminated incorrect answers. This strategy has been observed on student answer sheets in the past. Thus, visual inspection indicates that this unusual test taking strategy is a plausible explanation for these observed data. The statistical analysis is unable to conclusively determine that a testing irregularity did, or did not, occur.

Case III – Multiple Marks and Aberrance Abound in a Grade 5 Reading Class.

These data are in 20th place for anomalous results on the Reading/ELA tests. They were selected because of the presence of high aberrance and multiple mark rates. After careful review 15 of 27 answer sheets were found to be identical to another answer sheet within the classroom. The detection of the data was somewhat fortuitous since 13 of the identical answer sheets had only one incorrect, but highly improbable, answer, which triggered the high aberrance rate. If a different incorrect response had been selected for this question or a more difficult question had been missed, then the data may have appeared less anomalous and escaped detection.

The classroom data are from the Reading Grade 5 test. These data are significant because the statistics indicate that 19 other classrooms are even more anomalous in some way or another, illustrating the extreme nature of these cases. The summary of the data is provided in Table 13.

Table 13: Anomalous Data for Case III

	Grade 5 Reading Baseline	Classroom	Annotation
N	276261	27	
Pass Rate	0.7511	0.7778	The pass rate is about the same as the statewide rate.
Mean Score	2218	2338	The mean score is about 100 points above the state average.
Statistical index		26.8076	This p-value is < 1.0e-20
Mean w/o Incidents	2203	1990	There were 8 tests that did not have an incident.
Mean w/ Incidents	2300	2438	The disparity between tests with incidents and without incidents is huge for this data.
Aberrance Rate	0.061	0.6667	This rate is quite elevated. The probability value is less than ten to the -15th power.
Similarity Rate	0.0146	0.1481	The similarity rate is somewhat high, but not anomalous
Multiple Mark Rate	0.0388	0.3704	This multiple mark rate is very high and anomalous.
Gain Score Rate	0.0578	0.3913	This gain score rate is very high but it did not exceed the conservative threshold for gain score rates. The p-value is less than .00001

Table 14: 27 Reading Tests with Aberrance and High Gains

Test Number	Raw Score	Score 2005	Score 2004	Score 2003	Wrong to Right Changes	Other Changes	Aberrance
453745	0.98	2556	1996	1909	1	0	5.23
453746	0.98	2556	2039	1995	0	0	5.23
453756	0.88	2282	2400	2029	0	0	3.42
453771	0.45	1901			0	0	-0.14
453772	0.98	2556	2125	2182	1	0	5.23
453773	0.55	1953	2400	2342	0	0	0.87
453779	0.98	2556	2425	2182	3	0	5.23
453786	0.35	1809	2205	2287	0	0	1.69
453788	0.90	2322	2425	2400	1	0	4.23
453791	0.98	2556	2183	2160	0	0	5.23
453792	0.63	2005	2425	2287	0	0	0.41
453800	0.85	2247	2183	2400	0	0	4.61
453802	0.98	2556	2322	2182	0	0	5.23
453806	0.55	1988	2256	2299	0	0	0.35
453808	0.95	2442	2425	2400	0	0	2.75
453809	0.88	2282	2233	2029	0	0	2.97
453815	0.98	2556	2256	2233	0	0	5.23
453816	0.98	2556		2084	1	0	5.23
453824	0.43	1865	2183	1944	0	0	0.39
453828	0.98	2556	2400	2174	0	0	5.23

453833	0.98	2556	2400	2263	0	0	5.23
453837	0.85	2247	2322	2263	2	0	3.51
453838	0.98	2556	2521	2342	4	0	5.23
453839	0.88	2282	2125	2494	1	0	4.20
453841	0.98	2556			2	0	5.23
453846	0.98	2556	2205		3	0	5.23
453847	0.88	2282	2205	2046	0	0	4.20

Notes: Test numbers are unique to this analysis and cannot be traced back to actual students. Tests 453839 and 453847 were found to be "highly similar." Students having tests 453745 and 453846 repeated grade 4. Students having tests 453706 and 453809 repeated grade 3. 2005 was the first year in this school for students who had repeated grades 3 or 4.

Legend and Explanation for Table 14

Score 2005	A high gain score. Gains that were higher than the 95% threshold level are highlighted in Sea Green.
Wrong-to-Right Changes	A high number of wrong-to-right answer changes. Values with that are less probable than 1 in 20 are highlighted with Tan. Multiple marks indicating answer changes from wrong-to-right are quite rare at this grade level. The tests are especially anomalous when the fact is considered that answer changes on seven of the tests made the answer sheet identical with 6 other answer sheets, making 13 identical answer sheets in all.
Aberrant	An aberrant test response. These 13 tests having identical answer sheets are highlighted using Gold. These 13 tests had 39 of 40 correct answers, and an identical incorrect answer (which was very improbable given the performance on the other test questions).
Aberrant	An aberrant test response. These 2 tests having identical answer sheets are highlighted using Orange.
Aberrant	An aberrant test response. These answer sheets were not identical with other tests.

Based upon the detected statistical inconsistencies, these data are very anomalous and merit further investigation to determine the cause of the extreme values.

Case IV – Very Similar Tests in 6th Grade Reading

These data were specifically selected in order to demonstrate a group of extremely similar tests. This set of data is from the 6th grade Reading test. The data ranks in 28th position on the list of exceptions for classrooms in Reading. There are 18 students in the class. The similarity rate is 89% (i.e., 16 of 18 tests).

The data for this class are compared with the statewide baseline data in Table 15.

Table 15: Anomalous Data for Case IV

	Grade 6 Reading Baseline	Classroom	Annotation
N	287940	18	
Pass Rate	0.8516	1	Every student passed. This is not especially unusual in a class of 15 with an expected pass rate of 85%.
Mean Score	2296.695	2307.833	The mean score is about the same as the state average.
Statistical index		22.4954	This p-value is less than one in a billion times a billion. p < 1.0e-20.
Mean w/o Incidents	2280.21	2262	
Mean w/ Incidents	2393.983	2313.563	There is a 50 point difference between tests within the class for those with incidents and those without incidents.
Aberrance Rate	0.06	No Data	No tests were aberrant.
Similarity Rate	0.0194	0.8889	This similarity rate is extremely high. There is less than one chance in three that a class this size would have two or more similar tests; 16 similar tests is extreme; p < 1.0e-25
Multiple Mark Rate	0.0266	0.0556	Only one test had excessive multiple marks. This is within variation.
Gain Score Rate	0.0567	0.2308	13 gain scores were computed and 3 were unusually high. This value is not extreme.

Similarities are computed by comparing the student-selected answers from every test with the student answer of every other test in the school and then assessing the probability of the number of identical correct, the number of identical incorrect, and the number of different answers between the answer sheet pair. These computations have been done and are presented in Table 16.

Table 16: Pair-wise Similarity Indicator Values for 18 Tests

Answer Sheet Identifier	71	72	74	75	77	78	79	81	83	84	85	86	88	89	92	93	97
66	0.3	4.1	2.4	3.1	4.1	3.1	0.3	0.4	1.6	2.8	3.1	0.3	1.4	3.1	4.1	0.3	4.1
71		0.3	1.0	1.1	0.3	1.1	12.8	11.4	0.8	0.9	1.1	12.8	0.4	1.1	0.3	12.8	0.3
72			2.0	2.6	6.3	2.6	0.3	0.3	1.2	2.3	2.6	0.3	2.4	2.6	6.3	0.3	6.3
74				3.2	2.0	3.2	1.0	0.9	1.2	2.9	3.2	1.0	1.4	3.2	2.0	1.0	2.0
75					2.6	8.4	1.1	1.1	2.9	7.3	8.4	1.1	1.8	8.4	2.6	1.1	2.6
77						2.6	0.3	0.3	1.2	2.3	2.6	0.3	2.4	2.6	6.3	0.3	6.3
78							1.1	1.1	2.9	7.3	8.4	1.1	1.8	8.4	2.6	1.1	2.6
79								11.4	0.8	0.9	1.1	12.8	0.4	1.1	0.3	12.8	0.3
81									1.3	0.9	1.1	11.4	0.4	1.1	0.3	11.4	0.3
83										2.5	2.9	0.8	0.2	2.9	1.2	0.8	1.2
84											7.3	0.9	1.5	7.3	2.3	0.9	2.3
85												1.1	1.8	8.4	2.6	1.1	2.6
86													0.4	1.1	0.3	12.8	0.3
88														1.8	2.4	0.4	2.4
89															2.6	1.1	2.6
92																0.3	6.3
93																	0.3

Legend				
	Groups of Identical Tests	71, 79, 86, 93	72, 77, 92, 97	75, 78, 85, 89
	Pairs of Similar Tests			

The similarity indicator value is found in the table for a pair by ordering the pair with the smallest number used in the row position and the largest identifying number used in the column position. The value at the intersection of the row and the column is the indicator value. For example, the indicator value for the test pair (83, 75) is found looking across the row headed by 75 and down the column headed by 83. The value is 2.9.

There are three groups of identical answer sheets in this data.

- Group A consists of answer sheets 72, 77, 92 and 97. These answer sheets have 38 correct answers and 4 incorrect answers. Answer sheet 66 matched 41 of the answers with group A, replacing one of the incorrect answers with a different answer choice
- Group B consists of answer sheets 75, 78, 85 and 89. These answer sheets have 36 correct answers and 6 incorrect answers. Answer sheet 84 matched 41 of the answers, having one additional incorrect answer. Answer sheet 85 matched 41 of the answers, substituting one incorrect answer for one of the correct answers. Answer sheet 74 was most similar to group B, having 36 identical correct answers and 3 identical incorrect answers with group B.
- Group C consists of answer sheets 71, 79, 86 and 93. These answer sheets have 31 correct answers and 11 incorrect answers. Answer sheet 81 matched 41 of the answers, having one additional incorrect answer.
- Answer sheets 83 and 88 were not marked as similar by the statistical procedure, even though there are distinct similarities with other answers sheets.

The number of similar tests within this small class is difficult to explain as having occurred by chance. There are many improbable events in this grouping, not the least of which is 12 out of 18 answer sheets which are identical with at least one other answer sheet. It is very difficult for students who have actually studied together and who know the answers to the questions to remember the material and answer all of the questions in precisely the same way.

Case V—Aberrance in a Small Grade 11 Social Studies Class

These data were selected to illustrate aberrance when no other indicators are strong. This set of data ranks in 107th place of 110 exceptions reported for grade 11 Social Studies. The data are not as extreme as the other illustrations, but when one considers that 6 aberrant tests were found of 9 total tests and the average for the state is 1 in 20, the number of aberrant tests is extreme, even for such a small sample size. (When tests are performed with small sample sizes, statistical computations using small sample probability distributions are performed.)

The data for this class are compared with the statewide baseline data in Table 17.

Table 17: Anomalous Data for Case V

	Grade 11 Social Studies Baseline	Classroom	Annotation
N	229574	9	
Pass Rate	0.9452	1	The pass rate is high for this test and 100% passing is not unusual
Mean Score	2297	2377	The average score is 80 points above the state average.
Statistical index		5.5523	The overall statistical index represents a probability of 3 in 1 million. Most schools and classrooms that have exceptions are more extreme than this.
Mean w/o Incidents	2283	2133	These means are fairly typical of the statewide averages.
Mean w/ Incidents	2365	2447	300 point increase over those tests without incidents
Aberrance Rate	0.0555	0.6667	6 tests out of 9 are aberrant. This is very unusual for a class size this small. The probability is slightly higher than 2 in 1 million.
Similarity Rate	0.0597	No Data	No highly similar tests with
Multiple Mark Rate	0.0151	0.2222	2 tests have high multiple marks, but this is not extreme. The probability is between .007 and .008.
Gain Score Rate	0.0558	0.5	The gain score rate was not excessive given that only 2 gain scores were computable.

The degree of aberrance for these 9 tests is shown in the plots found in Appendix H.

These 9 exams show a strange pattern of different students answering incorrectly during different sections of the exam. The data are definitely unusual. They are not as extreme as some of the other demonstrated cases, but even in these data there are unusual aspects of the portions of the exams that were missed and the aberrance that is demonstrated. Several viable explanations for these anomalous data are reasonable.

Case VI – School Anomalies without Classroom Exceptions

These data were selected to demonstrate a school having an exception but no classrooms were reported with exceptions. There are 1431 tests from Math grades 9 (723), 10 (400) and 11 (308). The data are anomalous because of 3.9% of the tests had excessive multiple marks and 12.9% of the tests were similar with at least one other exam. Taken together these two indicator values have a probability of less than 1 in 10 to the 50^{th} power (The probability is very low due to the large sample size). The school is in 7^{th} place on the list of schools with exceptions in Math.

Table 18: Anomalous Data for Case VI

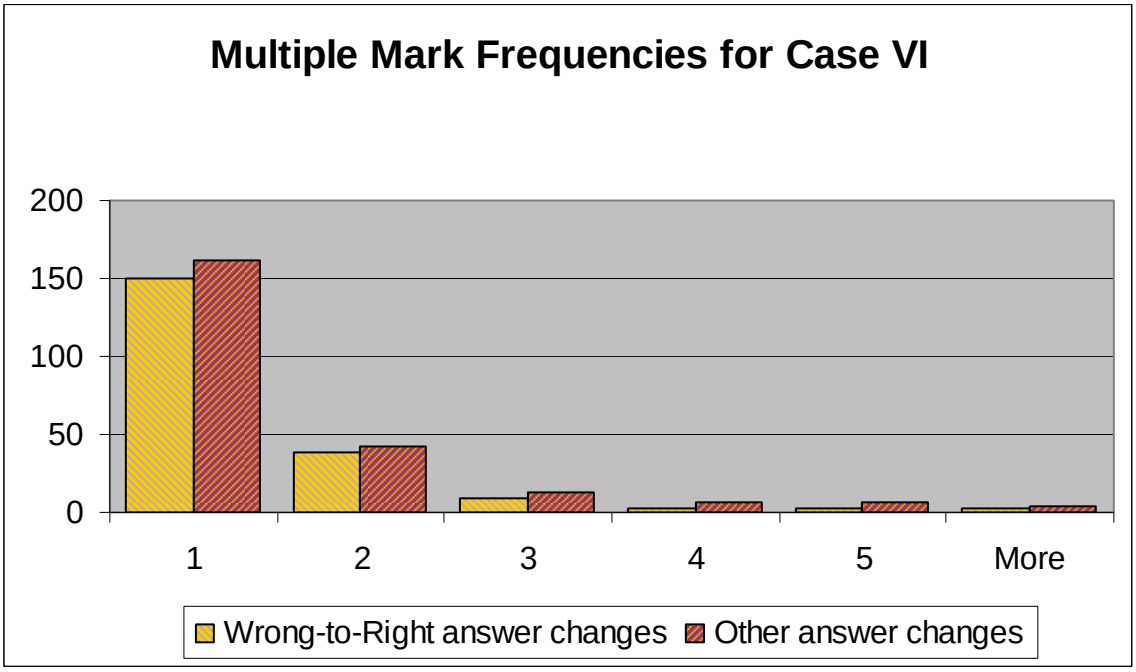
	Math Baseline for Grades 9, 10, 11	School (Grades 9, 10, 11)	Annotation
N		723, 400, 308	
Pass Rate	0.6220	0.3452	This school has a very low pass rate. Only slightly more than one-third of the students are meeting the TAKS standard.
Mean Score	2155	2024	
Statistical index		61.7029	The p-value is extremely small < 1.0e-60
Mean w/o Incidents	2141	2013	
Mean w/ Incidents	2247	2064	The score increase for incidents is only 50 points compared to a 100 point expected increase.
Pass Rate (Aberrance)	0.6674	0.4568	The pass rate increase for aberrance is 11.8% versus an expected 4.7% increase
Pass Rate (No Aberrance)	0.6201	0.3385	
Aberrance Rate	0.0388	0.0566	The aberrance rate is 1.8% higher than expected. It is elevated but not extreme.
Pass Rate (Similarity)	0.5868	0.3892	The pass rate increase for tests with similarity is 5% versus an expected 3.7% decrease. This represents a net change of 8.7% due to the similar test responses.
Pass Rate (No Similarity)	0.6235	0.3387	

Similarity Rate	0.0488	0.1293	The similarity rate is 2.6 times higher than expected. The probability of this many similar tests for this size of school was too small to compute; p < 1.0e-40.
Pass Rate (Multiple Marks)	0.5788	0.4286	The pass rate increase for excessive multiple marks is 8.7% versus an expected 4.6% decrease. This represents a net change of 13.2% due to excessive multiple marks.
Pass Rate (No Multiple Marks)	0.6224	0.3418	
Multiple Mark Rate	0.0109	0.0391	The multiple mark rate is 3.6 times higher than expected. The probability of observing this many tests with multiple marks is less than 1.0e-24
Pass Rate (Gain Scores)	0.9804	0.9630	The change in pass rates due to gain scores is within expected variation.
Pass Rate (No Gain Scores)	0.6359	0.3704	
Gain Score Rate	0.0516	0.0244	The gain score rate is less than expected.

It appears likely that there are instances of testing irregularities at this school. There are three clusters of identical answer sheets of sizes 4, 2 and 2 and from grades 11, 10 and 9 respectively. The score for the first cluster is 2289. For the second cluster the score is 2215, and for the third cluster the score is 1853. A unique element of the first cluster is that one student had 23 wrong-to-right answer changes and 5 other answer changes.

A frequency analysis of the multiple marks throughout the school indicates that the pattern of wrong-to-right and other answer changes is very similar. The data are shown in Figure 14. Only answer sheets having at least one multiple mark are shown.

Figure 14: Multiple Mark Frequencies for Case VI



There are 186 tests that were found to be very similar in these data. The proportions of the similar tests by grade are very close to the proportions of the tests by grade within the school (There is no statistical association between grade and numbers of highly similar tests). Caveon Data Forensics employs a clustering methodology in order to find structure and groupings in collusive tests. These data are presented by grade level in Appendix I.

Large groupings of data are difficult to visualize. There are 73 Math classrooms in this school. There may be groups of students copying from each other within the school, but the data does not suggest that inappropriate instructional methods are being employed to improve test scores.

Caveon Data Forensics is especially adept at detecting similar test clusters as a means of finding potential testing irregularities. These 186 tests consist of 76 similar test clusters: 38 in grade 9, 20 in grade 10, and 18 in grade 11. The expected number of tests at the nominal rate is 70. Using a 95% confidence interval there are between 90 and 140 excess similar tests. One form of test irregularity that could account for this data would be students collaborating with each other while taking the test at this school (e.g., through text messaging or answer copying). Even though this Report's focus is on the security of the TAKS tests in schools and classrooms, one must never forget that the stakes are high for the individual students, too.

Case VII – Schools that Have Many Classrooms with Exceptions

The data in this illustration were selected because of the high numbers of classrooms that have exceptions (7 of 15). This school is ranked second on the list of schools with exceptions. The data are extremely anomalous with a probability that is less than 1 in

1,000,000, …, 000 (72 zeros; p < 1.0e-72). These test data are from Reading grades 3 and 4. There are 262 tests that were given in the school. There are 21 perfect answer sheets for grade 3 tests and 8 perfect answer sheets for grade 4. In addition to these answer sheets with 100% scores, there were an additional 45 answer sheets that were identical with other answer sheets. Needless to say, this is extremely anomalous.

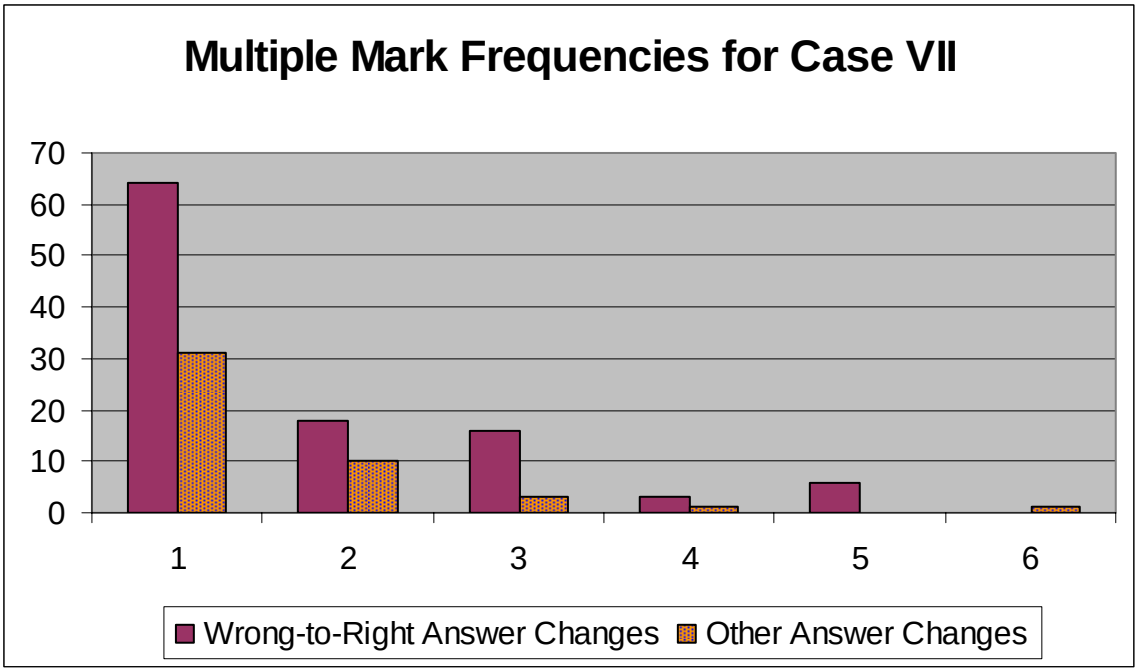
The data in Table 19 are mixed from the grades 3 and 4 for this school. The data for the grade 4 tests are much more anomalous than the data for the grade 3 tests.

Table 19: Anomalous Data for Case VII

	Reading Baseline for Grades 3 and 4	School	Annotation
N	116, 146	116, 146	There are 116 students in grade 3 and 146 students in grade 4.
Pass Rate	0.8364	0.958	The pass rate is 12% higher than expected and close to 100%; 251 students passed.
Mean Score	2267	2373	The class mean is quite a bit higher than the state average.
Statistical index		73.7888	The p-value is less than 1.0e-70.
Mean w/o Incidents	2258	2389	There were 141 tests with incidents.
Mean w/ Incidents	2312	2354	
Aberrance Rate	0.0609	0.2366	The Aberrance rate is extremely high. The p-value is less than 1.0e-30. 15 of the aberrant tests are from the 3rd grade and 47 of the aberrant tests are from the 4th grade.
Similarity Rate	0.0107	0.084	Given the low similarity rates in these grades in general, an 8% similarity rate is very high. The probability is less than 1 in 50 million. All the similar tests are from the 4th grade.
Multiple Mark Rate	0.0176	0.2939	The multiple mark rates are very high. There is typically not a lot of answer changing on these tests and nearly 3 tests in 10 have excessive multiple marks. The p-value was less than 1.0e-40. 64 of the tests with excessive multiple marks are from the 4th grade and 13 are from the 3rd grade.
Gain Score Rate	0.0718	0.0752	The gain score rate is within expected variation.

Figure 15 provides the histogram of multiple mark frequencies on the answer sheets. These data are much different than the multiple mark frequencies in Figure 14.

Figure 15: Multiple Mark Frequencies for Case VII



The number of wrong-to-right answer changes in this data is a lot higher than the other answer changes. Given that 77 of the answer sheets of 262 had excessive wrong-to-right answer changing, 62 of the answer sheets were detected with aberrance inconsistencies, and the pass rate is 12% higher than the statewide pass rate, it seems that answer changing has most likely resulted in higher scores for this school. A simple chi-square test of association shows that the multiple marks and the aberrance are statistically associated ($p < .00003$).

These data appear to have been the result of one or more testing irregularities. A review and verification of the circumstances will be required to determine if alternative explanations exist.