July 2016 Overview of Caveon Data Forensics for DC OSSE

Introduction

Caveon Data Forensics™ provides DC OSSE with the ability to measure and monitor test security threats in order to ensure the integrity of test scores and test result data. Caveon Data Forensics uses statistical analyses of test-response data to detect anomalous[1] test response data for schools, classrooms and students. Patterns of highly unusual statistical situations identify test-taking patterns indicative of *possible* test fraud (e.g., cheating and/or piracy) or security breaches[2].

The results of these analyses provide critical information regarding where and when suspect activity occurred, by whom, and its effects on test results.

Data forensics analyses can detect potential test security breaches. Additionally, the data forensics analyses can be used to evaluate security risks associated with groups of students, such as all individuals who were taught by the same instructor. The analysis of groups identifies program-specific elements (e.g., schools, grade levels, classrooms) where *validity of test results* was questionable because statistical patterns yielded highly improbable results that could be the result of scores impacted by irregularities or test security violations.

This document provides a brief overview of Caveon Data Forensics services, specifically the statistics used in the analysis of test-response data, and provides examples of how students or schools may be flagged by Caveon Data Forensics.

Description of Statistical Methods

Whenever one of the security statistics provides an anomalous result, we say that we have found an extreme statistical value which *may* be associated with test security violations, such as answer copying or lax test security. The presence of anomalies which may be associated with test security violations does not necessarily mean that test security violations occurred. Because anomalies are unusual and rare, we do not expect to find many of them in these analyses. When they are found, an underlying cause is present. These statistics are designed to detect unusual situations where test security violations may be the underlying cause.

Caveon Data Forensics results are communicated in terms of an "index" value. Index values are a convenient way of representing very low probabilities. The value represents the probability using a power of 10. For example, an index value of 6 means that the probability is one in 1,000,000 or 10^{-6}.

---

[1] An observation is statistically anomalous when the measured attributes are seen to be extremely different than the expected values for those attributes. A common euphemism to describe anomalous observations is "outlier." Statistical practice for outlier detection or declaring an observation to be anomalous is usually based upon statistical tests, where the probability value of the test statistic is extremely small.

[2] A test security breach occurs when test security is violated and there is evidence that suggests the integrity of the test administration was compromised.

© 2008-2016, Caveon, LLC. All Rights Reserved. Permission required for distributing copies.

Higher index values correspond to lower probabilities. The mathematical relationship between the index value and the probability is: *probability* = 10^{-index}

Probability thresholds guide "flagging" decisions for detecting and defining anomalies related to individual test instances or groups of test instances (e.g., all the tests administered at a school).

It is important to note that anomalous test results are not conclusive evidence of misbehavior or even invalid test scores. Even so, they need to be understood and vetted. This may be particularly true in situations where a school has more than one grade/subject examination flagged. Schools with these multiple flags may be prioritized for additional review. DC OSSE was required by the Test Integrity Act of 2013 to implement a process by which test integrity reviews may be conducted. The use of flagging through statistical analysis of test data to initiate reviews is outlined in Section 4.B.iii of the DC OSSE's 2016 Test Integrity Guidelines.

Detection Statistics Overview

Caveon Data Forensics uses several statistics to detect potential testing irregularities and security violations. Three of the most important statistics include:

1. Similarity
2. Answer Changes
3. Gains/Losses

1. Similarity

Tests are flagged by the Similarity Analysis when the number of identical responses two students (or more) share is significantly greater than the number that would be expected, assuming the students completed the exams independently of one another and did their own work. Caveon's Similarity Analysis identifies when two or more test takers may have inappropriately shared information or collaborated during test taking rather than worked independently. The analysis takes into account the fact that answer choices depend upon student competence, which is measured using the student's overall exam score. Given the student's observed level of performance, the Similarity Analysis uses a statistical model (i.e., item response model) to compute the probability that the student would have correctly and incorrectly answered each exam item.

For each item, the probability is calculated that two students would have responded to that item in the same way (i.e., both students responded to the item with the same correct or incorrect answer). After computing the match probability for each item, they are summed up for all the items on the exam. The resulting sum is the expected number of identical responses between the two exams.

The observed number of identical responses between the two exams is compared statistically to the expected number of identical responses between the two exams under the assumption that the two students worked independently. If the observed number of identical responses is significantly greater

© 2008-2016, Caveon, LLC. All Rights Reserved. Permission required for distributing copies.

than the expected number, this is evidence that one or both of the students did not work independently and potentially received assistance on his/her exam.

Common questions about the Caveon Similarity Analysis:

Shouldn't we expect strong performing students to have more identical correct responses? The short answer is "yes." The best example of this occurs when two students answer every question correctly. They will have identical correct responses for every single question. Caveon's model takes this into account. As detailed above, we start by calculating the likelihood of a student's response, and at high-performance levels students are more likely to answer an item correctly. Thus, identical incorrect responses usually provide more evidence of non-independent test taking than identical correct responses. The analysis adjusts the expected values of identical responses based upon student ability. It is also the case that very small probabilities for high-scoring students are not usually observed.

What if an item is "flawed" in that a particular distractor is unusually attractive to students? Because ALL students' results are analyzed, the analysis is sympathetic when the group is predisposed to select a particular response. In essence, a student selecting the attractive distracter is recognized as "usual" test taking, and so the probability of any student selecting that answer is increased. The net result of the approach means that the analysis will be less likely to flag test instances as similar when some distractors are especially attractive.

An Example of Extreme Similarity

In general, we observe relatively few identical responses between pairs of test takers. By design, the similarity index between any pair of test takers has an average value that is approximately 0.3.

In this example, based upon how well these two students performed, we would expect to see 19 identical-correct responses; 3 identical-incorrect responses; and 39 different responses—one student answers a question "A," the other student answers the same question "B," etc. The observed responses were in line with those expectations, resulting in a very low similarity index. It is clear to see that there are only a few identical incorrect responses (red) and there are many items with different responses (green)

© 2008-2016, Caveon, LLC. All Rights Reserved. Permission required for distributing copies.

Figure 1 – Independent Test Taking Between a Pair of Students

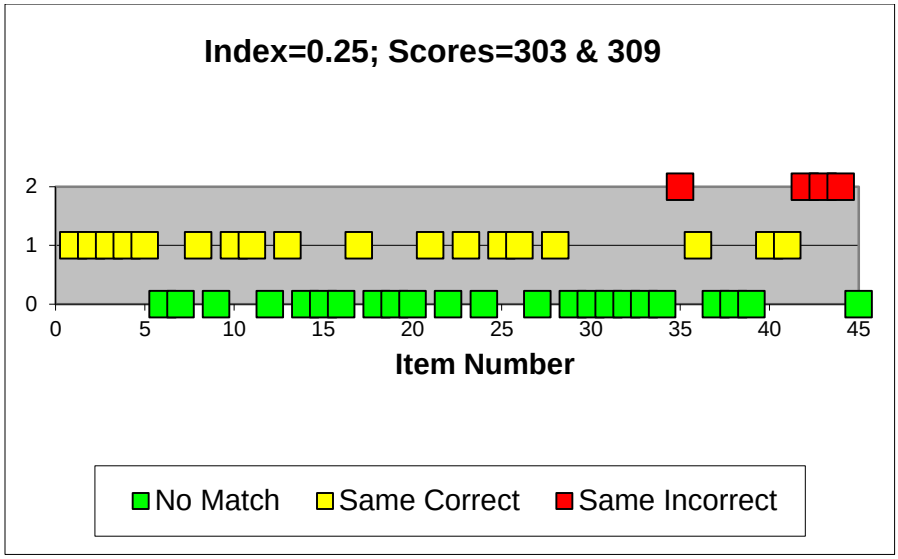
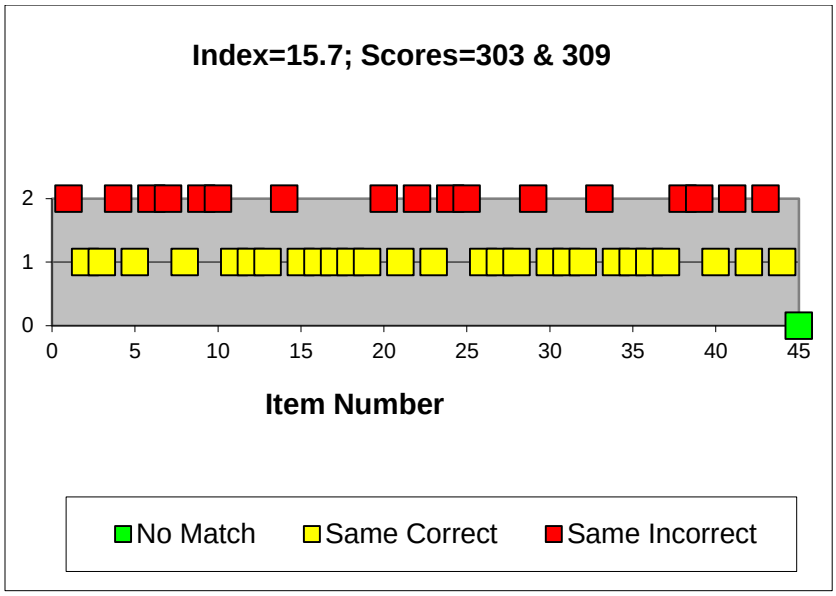


Figure 2—Possible Non-Independent Test Taking Between a Pair of Students



In Figure 2, this pair of students answered all but one live test question identically (this chart does not include 16 field test items).

The *expected* number of identical responses is very similar to the previous pair. Indeed, we anticipate only 3 identical-incorrect responses, and the vast majority of responses between the two test takers will

© 2008-2016, Caveon, LLC. All Rights Reserved. Permission required for distributing copies.

involve different answers. However, the observed results are very different. Note the very large number of identical responses that were actually observed: 27 identical-correct responses and 17 identical-incorrect responses. The probability of such similar results *under the assumption of independent test taking* a probability smaller than 1 in 10^{-15}.

2. Unusual Numbers of Wrong-to-Right Answer Changes

When answer-change data are available, these data may be analyzed to determine whether unusual wrong-to-right (WTR answer-changing behavior was potentially present within a group of test takers. Caveon compares the number of students in a grade/subject group with elevated numbers of wrong-to-right answer changes against the baseline for the population (e.g., all students in the state, including those who tested online and using paper and pencil). Because the definition of "elevated number of WTR answer changes" is statistically set at approximately 5%, there is not supposition that a particular student with an elevated number of WTR answer changes was associated with a test security issue. Instead, the goal of the analysis is to detect concentrations (or "pockets") of students with an elevated number of WTR answer changes. Here is an example:

- In spring 2015, across the state, **10%** of the 4,159 students who sat for the PBA ELA 10 exam were noted for elevated wrong-to-right answer changes.
- One school was flagged where **40%** of the 84 students (33 students) who sat for the exam were noted for elevated WTR answer changes. The frequency of elevated answer changes was **four times** the statewide rate. Thus, this school was identified with a concentration or anomalous number of students with WTR answer changes.
- Using the statewide average or baseline as the definition of normal testing, the probability of seeing such a volume of elevated answer changing in one school is calculated at 1 in 10^{-13}, or approximately **1 in a trillion**, assuming that elevated answer changing is a random, statistical event associated with students taking tests.

Wrong-to-Right minus Right-to-Wrong Answer Changes

It is relevant and useful to determine whether the concentration of students with elevated WTR answer changes was associated with a significant improvement in test scores. To help understand this dynamic, Caveon calculates a WTR minus RTW answer-change statistic. This analysis helps compute net-improvement on a test taker's score by subtracting the number of RTW answer changes from the total number of WTR answer changes. For instance, if a student has 5 WTR answer changes on his exam and 3 RTW answer changes, his score is only improved by 2 correct items. This statistic helps gauge the actual impact of answer changes on students' scores. In the example listed above, the average score of those students with elevated answer changes was 10 percentage points greater (.41) than those students without elevated answer changes (.31).

3. Gains/Losses

Large gains/losses are an important statistical analysis and have been used as a trigger for investigating schools in many states and educational systems. Large gains are relied upon nearly universally as a

© 2008-2016, Caveon, LLC. All Rights Reserved. Permission required for distributing copies.

trigger or a supporting factor in an investigation. Conversely, large score losses which are observed after the implementation of rigorous test security measures can also initiate an investigation (for instance, if scores drop during test administrations that are carefully monitored).

Because 2015 was the first year of PARCC testing in DC, Caveon was unable to analyze gains or losses on PARCC data at this time, but will use this analysis after 2016 results are available.

A fundamental aspect of a security investigation is to seek for reasons that explain the anomalies or allegations. In the context of large gains, this means attempting to understand why the gain was present. The following are some reasons why gains might be observed:

- The student population changed,
- The students' language proficiency level increased,
- The teaching and educational resources improved,
- The teaching staff changed,
- The teaching was excellent,
- The teaching was focused on the tested concepts, and
- The teaching was informed by live test questions.

Large gains that are truly anomalous deserve attention. If the teaching is excellent and meritorious it should be recognized and replicated.

Statistical data may help understand the presence of the large gain if sensitive statistical measures are available. When evaluating gains in schools, it is best to remove extraneous sources of variability. For example, cohort gains analyses (i.e., year-to-year changes for individual students) are more powerful than cross-sectional gains analyses (i.e., year-to-year changes for grades within a school) because the effect of a student population changing has been removed. It may be difficult or even impossible to adjust the statistical results for many of the factors (e.g, new teachers, new resources, improved language proficiency) because they may be unknown or not measured.

Going Forward

The Test Integrity Act of 2013 requires DC OSSE leadership to investigate possible testing improprieties that are flagged through statistical analysis of test response data. In May 2016, Caveon analyzed test data from spring 2015 PARCC test administrations. Caveon Data Forensics analyses of spring 2016 PARCC test data are planned for August – October 2016. Results from these analyses will be compared against the 2015 results, and will provide additional data regarding schools and classrooms that can be useful in planning and implementing various test security measures, including supplemental training; scheduling live site monitoring; and prioritizing investigation efforts.

© 2008-2016, Caveon, LLC. All Rights Reserved. Permission required for distributing copies.