Detecting Cheating in Computer Adaptive Tests Using Data Forensics

James C. Impara, Caveon, Caveon, LLC

Gage Kingsbury, NWEA

Dennis Maynes and Cyndy Fitzgerald, Caveon, LLC

Key Words: Test security, detecting cheating

Paper presented at the 2005 Annual Meeting of the National Council on Measurement in Education and the National Association of Test Directors, Montreal, Canada

Detecting Cheating in Computer Adaptive Tests Using Data Forensics

Overview

Cheating is on the rise in both high school and college settings. In a series of surveys and a review of research on trends in cheating in college, McCabe (2005) of Rutgers University and his colleagues found that in 1961, only 26% of students admitted to copying from another student during a test. That percentage rose to 52% thirty years later in 1991. In a 1999 study, 75% of students admitted to some form of cheating. Cizek (1999) wrote, "…one conclusion from the trend studies is clear: All agree that the proportion (of cheaters) is high and not going down." (p. 35)

Over the past 15 years, there has been a strong movement in credentialing testing to move from paper and pencil to computer-based testing (CBT). This trend has been slower to occur in the educational community, especially for high-stakes testing, but there is movement in that direction in both local districts and for state assessment programs. Although providing many advantages for the testing program, students, examinees and users of the test results, this trend has also produced at least one major security problem: an enhanced ability to capture and share test information. Davey and Nering (2002) warn that "…at least some of what has been learned over the years about securing conventional high-stakes tests must be updated to meet the new problems posed by CBT administration." (p. 166). They add, "The danger is not that question pools will be disclosed. As stated, that much is a given—they will be. The danger is they will be disclosed so quickly that economics, logistics and pretest requirements make it impossible…to keep up." (p. 188) Note that in most educational settings, even when CBT is employed, it is not done in an "on-demand" context. Testing windows are fixed, rather than continuous. This fixed window administration strategy does not necessarily prevent the security risks alluded to by Davey and Nering, but it may help to reduce such risks.

Cheating

Cheating can occur using a variety of methods such as using inappropriate materials (e.g., PDAs, text messaging, cheat sheets), not stopping when time is called, copying/collusion; by teachers: correcting student errors en masse (erasing wrong answers and inserting correct answers), watching over shoulders and assisting individual students directly, by putting answers on the board, or by obtaining some (or all) test questions in advance of the testing window and using these as "practice" tests. Some of these strategies are made more difficult in a CBT testing mode, but not all. Using computer adaptive tests (CATs) also makes some of these strategies more difficult, but it does not preclude some of these security risks.

Table 1 illustrates how conditions of testing can also influence the type of cheating.

Table 1
Potential cheating methods across different test delivery modes

Cheating/Test Delivery mode	Paper and Pencil	Computer based – linear	Computerized– adaptive
Examinee: text messaging and other forms of two-way communication (e.g., two-way radios)	X	X	X
Examinee: Using unauthorized materials (e.g., calculator	X	X	X
Examinee: Collusion with another examinee (e.g., copying)	X	X	
Examinee: Proxy testing (having another person take the test)	X	X	X
Examinee: Using braindumps		X	X

Various approaches can be used to detect cheating. One of the most common approaches is to receive reports of observations by others (someone "rats out" the miscreant); score change anomalies (volatile changes in scores across years – either students, teachers or schools), and data forensics (looking for unusual response patterns, latent response times, erasure analysis, computing collusion indices).

The most common statistical approaches to detecting cheating on CBTs include using latent response times, employing "cheat" programs, and looking at score differences across time (e.g., year-to-year classroom/school/district score changes within a grade or for a student cohort, or item drift). Caveon's approach employs some of these same strategies, but enhanced in several ways. These strategies have not previously been used with CATs.

Purpose

The purpose of the present paper is twofold. The first section of the paper discusses data forensic methods for detecting test fraud using indicators of aberrance (unusual response patterns and unusual patterns in item response times) and the second half provides results of the application of these measures to an educational assessment program that uses computerized adaptive testing (CAT).

NWEA and Caveon conducted a study to investigate the potential and power that data forensics methods, founded in measures of aberrance[1], collusion[2], and score volatility[3] have for detecting exam fraud in an educational CAT environment. The primary study goals were to assess detection rates in live data and to assess the impacts of aberrant test-taking upon the test results. Secondary study goals were to evaluate the security strength of CAT as a means of test delivery and to determine whether measurement of aberrance in a CAT (which is inherently adapted to the examinee's ability and theoretically a near-optimal test) is possible and what it might mean if it were discovered.

In order to assess detection rates of test fraud in live data, the data were seeded by NWEA staff (and unknown to Caveon) with known instances of anomalous test results[4]. The data forensics analyses were performed by Caveon to present the impact of aberrant and collusive test-taking on the test results.

The results are presented in two stages. In stage one the overall results are described based on a particular, and new, approach to analyzing data to look for data anomalies. The second stage

[1] Aberrance is observed when a subject answers the test questions in a manner that is inconsistent with demonstrated knowledge and behavior. Examples are inconsistencies in the amount of time taken to respond to test items, and answer selections that are inconsistent with a student's demonstrated ability on other test items.

[2] Collusion occurs when examinees share answers for the test items either during or before the test. It also manifests itself when an educator provides the same answers to test items to multiple students. Statistically, collusion indicates that the tests are not being taken independently.

[3] Score volatility is measured when a student retakes the test and demonstrates an extreme score change. When the change is so extreme the practitioner disbelieves that the result is due to chance and may believe that the result is due to cheating.

[4] Note that the seeding (usually changing a wrong response to a correct response) did not include adapting the routing of the test. Thus, responses to subsequent items were not affected as would have been the case had the "cheating" been done by the student. This may have had an impact on the ability of the data forensics to detect the cheating that was seeded by NWEA.

represents how accurate this approach was to detecting “known” cheating situations. Before discussion of the results, some terms are defined and the nature of the data is described.

Aberrance

Aberrance refers to a test result that does not conform to the test response model. Because there are many types of non-conformance this term lacks a precise definition. Some types of non-conformance include wild guessing (as opposed to educated guessing), poor test preparation, mis-keyed test questions, and cheating (or pre-knowledge of some or all of the test content).

Identifying response aberrance begins with an examination of the individual responses in context of all the responses. A single response to a single test question cannot be construed as aberrant or not, except in the sense that the response may be so improbable that it causes the test administrator extreme surprise. This concept of surprise (due to observing extremely improbable events) is an essential aspect of aberrance. IRT (Item Response Theory) models provide the statistical framework for objectively measuring the probability of a set of responses and from probability to “surprise” when the item responses do not conform to the testing model.

Different kinds of non-conformance, or aberrance, are in reality different patterns of unusual or improbable responses. A method that attempts to evaluate patterns of responses must be capable of differentiating between the different patterns as they relate to the observed responses.

Another aspect of aberrance is that not all responses will necessarily be improbable. We are then left with the situation that each observed response has a different probability and the numbers of improbable responses directly correlate with our notion of surprise or non-conformance. A single improbable response should rarely convince us (or provide sufficient evidence) that the test is being taken inappropriately. An exception might be the case where a person does extremely well answering nearly all questions correctly but then answers an easy question with a very improbable incorrect response (which could also be termed a blunder). Similarly, when pilot testing of new items is done by embedding them in an operational test (as compared to stand-alone pilot testing), test behavior on the pilot test items that is not consistent with performance on the operational items can be indicative of anomalous performance.

The aggregated evidence of conformance versus non-conformance needs to be evaluated in order to convince us that a test was taken inappropriately. It is only by considering all the responses in context and then ordering those responses by the degree of surprise (or improbability of occurrence) that a particular set of responses can be viewed as “aberrant.” In other words, aberrance is a property derived from the individual responses, based on all the responses.

For the current purposes, aberrance is defined as the number (or percent) of improbable responses and the degree of improbability associated with those responses in the observed test responses. It is well known that the sample mean and sample standard deviation are not resistant to outliers and influential observations. In the same manner, the estimation of theta in the IRT model will be heavily influenced by outliers or aberrant responses. This influence has the potential to mask the true aberrant responses, making the detection of aberrance (or responses that generate “unusual surprise”) difficult.

If a test has psychometric integrity, little or no aberrance will be seen in the test responses of the individual who responds to the test fairly and honestly. The cheater can be viewed as a test-taker who has an unfair prior knowledge (or knowledge gained during the exam) of the test content. If the cheater has gained access to the entire test content, then it is unlikely that response aberrance models will detect this behavior. Instead item response latency aberrance models will be required. On the other hand, if the cheater has gained access to less than 100% of the content, then this individual can be viewed as responding differently to the questions, depending on prior knowledge. The individual will respond to the questions with prior knowledge at a higher

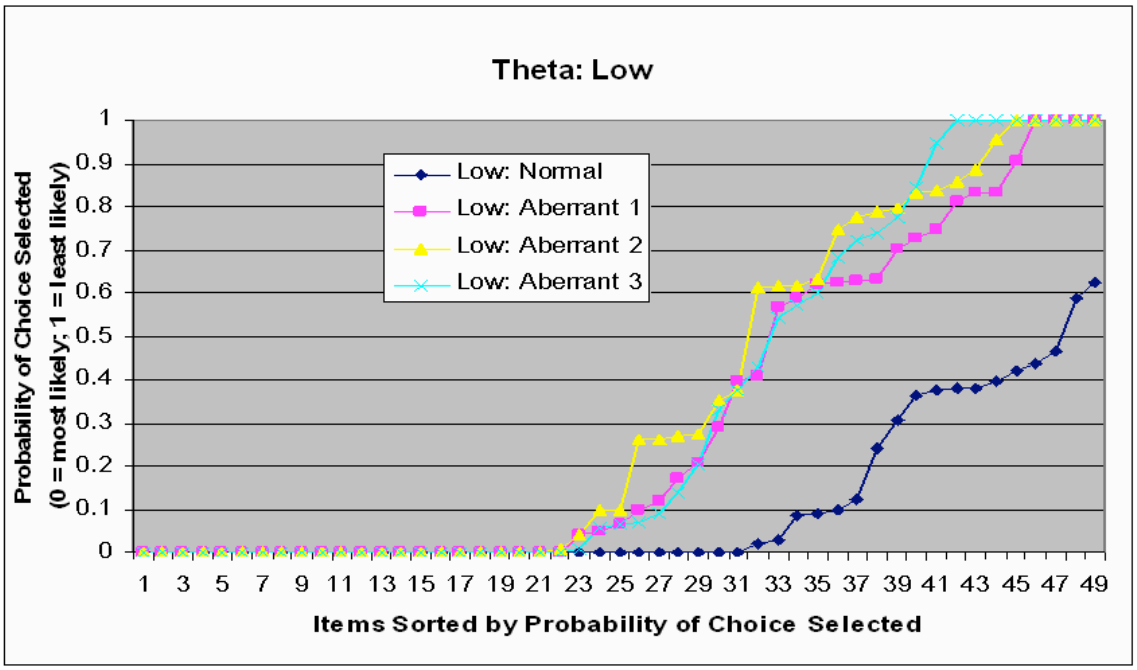
level of theta than to the questions without prior knowledge. The cheating model is then a model where two levels of theta are presumed to be exhibited (i.e., the response pattern will be bi-modal in terms of estimating theta).

With sufficiently large data sets, even unlikely patterns will show up from time-to-time. An example is the lucky guesser who is able to guess a significant number of correct answers. And another example is the individual who makes a lot of "stupid" mistakes (e.g., who may have accidentally got off line on the response sheet). In both of these cases the actual responses will not reflect the test taker's actual knowledge. These are such low probability occurrences that they do not merit separate models, but they will be present in large data sets by chance alone.

Caveon believes that aberrance in response patterns and response latencies (when available) are one of the better indications of cheating and item theft. For computer-based testing, six different measures of aberrance are used in the analysis, three that look at the answers examinees select, and three that look at how long it takes the examinee to answer the question. By combining all six into a single aberrance value, we can get useful information on the rates of security problems.

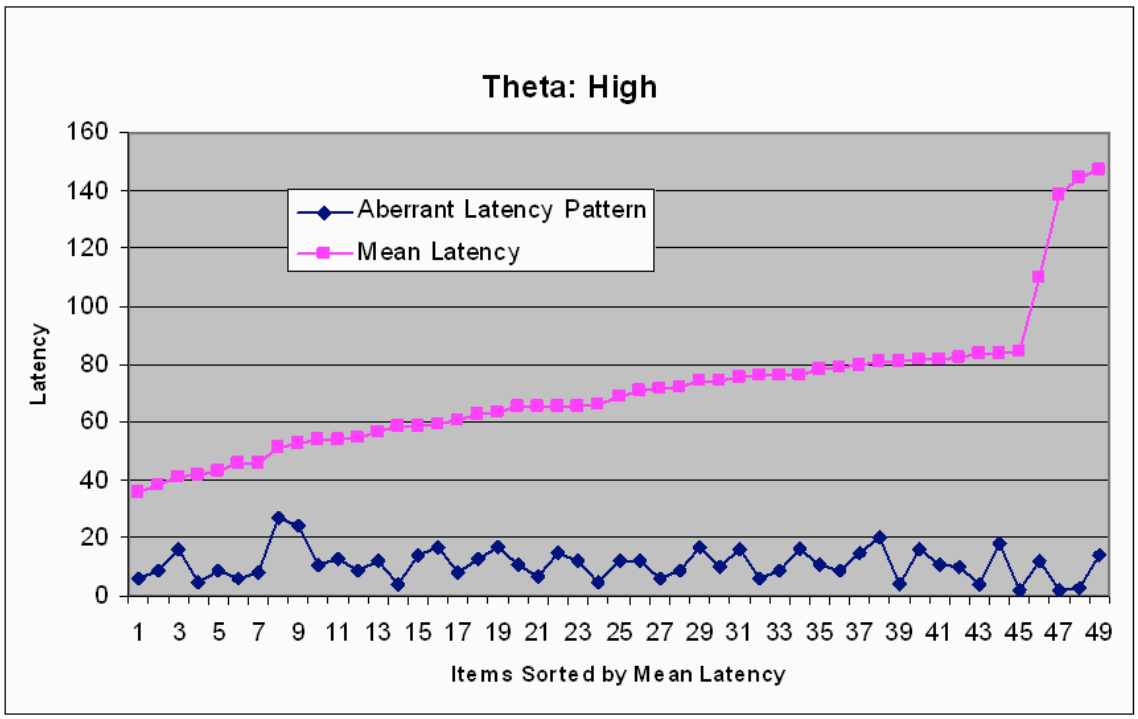
But let's first show two examples of Caveon's aberrance measures. These examples come from the certification and licensure arena. Figure 1 shows the response patterns for 4 individuals with the same low score. It indicates three aberrant tests and a normal one. Although the statistics are far from simple, unusual or improbable selections of answers to items make it easy to separate the aberrant tests from the normal ones. One can see such differences at every ability level except the very highest.

Figure 1



With computerized tests we are able to add in item response time measures to the aberrance analysis. Figure 2 shows a test record for an individual who scored very high on the test even though the amount of time to respond to each question was too brief to allow him or her to answer the questions in a normal way. The amount of time taken was uncorrelated (tau = -.04) with the amount of time the test takers as a whole took to answer the questions.

Figure 2



At Caveon we have found aberrance to be a useful tool in tracking differences across the entire testing population. The statistical models used generate a 5% aberrance rate if there is no actual aberrance in the test results. Using this as a baseline comparison, we can evaluate and compare across localities (e.g., regions, districts), and even specific testing locations (e.g., schools, classrooms, testing centers). There are a few terms that need to be defined in order to be able to interpret the data presented below. They are presented in Appendix A.

Table 2 provides an illustration of statewide rates for a sample of 20,661 educational tests. What is most alarming is that a full 17% of the tests show as aberrant in this case. (Only the tests with the highest aberrance levels are shown.)

Table 2
District Summary

District	Tests	Mean Percentile	Pass Rate %	Pass Rate Index	Aberrance Rate %	Aberrant High Score Tests	High Score Tests	High-Score Aberrance Rate	Cheating Index	Aberrant Low Score Tests	Low Score Tests	Low-Score Aberrance Rate	Piracy Index	Comments
Overall	20661	0.50	49	0.0	17	1753	10221	17	0.0	1815	10440	17	0.0	
107	1614	0.46	45.1	-3.6	19.3	141	728	19	1.3	171	886	19	1.2	
263	1463	0.52	52	1.4	20	149	761	20	1.5	144	702	21	1.9	
444	1386	0.53	55.1	4.9	23.7	184	764	24	7.2	145	622	23	4.5	The elevated high-score and low-score aberrance rates make this anomalous. These rates are indicative of cheating.
912	1028	0.43	38.9	-11.4	13.7	63	400	16	0.1	78	628	12	0.0	
412	820	0.42	37.6	-11.5	17.6	57	308	19	0.6	87	512	17	0.2	
719	807	0.55	56.1	3.9	18.5	81	453	18	0.5	68	354	19	0.7	
122	670	0.51	50.3	0.2	26	79	337	23	3.0	95	333	29	7.6	The elevated high-score and low-score aberrance rates make this anomalous. These rates are indicative of cheating.
988	614	0.50	46.6	-0.8	17.3	44	286	15	0.1	62	328	19	0.6	

We can see similar results when looking at individual schools in Table 3. The school at the top, labeled 4379, has a very high overall aberrance rate of 36%, mostly for tests with high scores. The Caveon Cheating Index of 12.4 indicates the probability of this result occurring by chance as less than .00000001!

Table 3
School Detail

School	Tests	Mean Percentile	Pass Rate %	Pass Rate Index	Aberrance Rate %	Aberrant High Score Tests	High Score Tests	High-Score Aberrance Rate	Cheating Index	Aberrant Low Score Tests	Low Score Tests	Low-Score Aberrance Rate	Piracy Index	District	Comments
8500	807	0.55	56	3.9	18	81	453	18	0.5	68	354	19	0.7	719	
9456	789	0.55	59	7.1	14	71	464	15	0.1	43	325	13	0.0	777	This is a high pass rate.
7001	545	0.50	50	0.1	13	35	273	13	0.0	34	272	13	0.0	289	
5514	501	0.38	33	-12.9	13	24	166	14	0.1	39	335	12	0.0	912	
8052	442	0.52	49	-0.1	24	50	217	23	2.0	55	225	24	2.6	851	
4379	383	0.56	61	5.4	36	81	234	35	12.4	56	149	38	10.5	444	This pass rate is in the presence of high aberrance for high and low scores. This may indicate test coaching at the school.
8849	372	0.53	52	0.6	11	22	195	11	0.0	18	177	10	0.0	777	
9621	359	0.42	38	-5.0	26	32	136	24	1.6	60	223	27	4.1	528	A high amount of low score aberrance with low pass rates. Most likely the students are unprepared to take the exam.
7047	345	0.53	56	1.8	22	45	193	23	2.0	30	152	20	0.7	444	
9453	324	0.50	50	0.1	18	24	163	15	0.1	35	161	22	1.1	777	
9161	317	0.63	68	11.0	17	30	217	14	0.0	25	100	25	1.7	875	Very high pass rate.
5117	306	0.44	41	-2.5	11	17	126	13	0.1	18	180	10	0.0	940	
6669	295	0.41	35	-6.3	21	27	103	26	2.1	34	192	18	0.3	107	

Using statistical models, it is also possible to discover and verify proxy testing, copying and other forms of collusion. As an example, if a proxy testing service is operating (a fairly frequent occurrence in some foreign certification and admissions testing programs), then scores for different examinees should appear too similar and have other suspicious patterns (this is an unlikely scenario in an educational setting because the teachers know who the examnees are!). Similarly we should be able to identify individuals who are taking tests collectively as a group, with or without the help of an instructor. Or if a group is using similar crib/cheat sheets, unauthorized Web resources, or some other means of working together other than copying from each other's test papers.

Collusion Analyses

Table 4 shows a cluster of seven tests with different examinee IDs. These data are from a sample of examinees who took a certification test. Also shown are their scores and the date and time of the test. Caveon's collusion statistic identified the tests as matching too closely to have occurred by chance. And notice the pattern of testing. All tests occurred on the same day. Furthermore, mostly each test started at about 20-30 minute intervals. This is likely a situation where a proxy test taker is operating. As noted above this type of proxy test taking is not something that may be of high concern for most educational tests. However, as more and more high stakes tests come on line this may become more of a problem, especially when proctors are not the students' teachers..

Table 4
Proxy Testing Illustration

	A	B	C	D	E	F	G	H	I	J	K
1	Examinee	Test	Site	Country	Date	Time	Score	Prob			
2	283	101	Site12	US	1/15/2003	1:04:18 PM	0.77	246			
3	405	101	Site 4B	US	5/24/2003	3:17:44 PM	0.87	143	Looks like a proxy test taker.		
4	351	101	Site 4B	US	5/24/2003	3:47:03 PM	0.88	258			
5	860	101	Site 4B	US	5/24/2003	4:12:16 PM	0.88	5029			
6	446	101	Site 4B	US	5/24/2003	4:48:52 PM	0.85	88			
7	440	101	Site 4B	US	5/24/2003	5:16:50 PM	0.83	5029			
8	123	101	Site 4B	US	5/24/2003	4:09:46 PM	0.88	85			
9	559	101	Site 4B	US	5/24/2003	4:46:14 PM	0.82	85			
10	756	101	Site 17	US	1/24/2003	2:34:00 PM	0.85	2134			
11	659	101	Site 17	US	4/11/2003	9:42:30 AM	0.85	2134			

For paper-and-pencil tests, the collusion analysis is equally effective, identifying traditional copying or instances where a teacher may be systematically erasing and changing the answers of students in his or her class.

Retake Analyses

Although not typically a problem in many educational settings, most certification testing programs have policies that permit, but impose conditions on retakes. It's important to track violations of these policies as well as looking at large gains or losses in test scores as tests are retaken. Both the violation of retake policies and retake gains and losses – what we refer to as volatile retakes – may be indicators of attempts to cheat or steal questions. Table 5 shows a list of volatile retakes from 9 examinees, retakes where the scores have changed, up or down, more than they should. Two of these examples are discussed. First, notice that Examinee 8879 at the bottom scored 85% after a score of 0% on the previous exam. Examinee 6343 retook the test after passing with a very high score of 92%. His/her score on the retake was only 5%. The first example may be an examinee who was either trying to familiarize him or her self with the test on the initial test, or who was trying to steal items. The second example suggests the examinee was trying to memorize items during the second attempt with no intention of getting a high score (after all, they had already passed). The relevance of examining the results of retakes and volatile retakes becomes relevant in the following educational contexts. First, when a test, like a graduation test, is given several times a year, students who retake may be retaking in violation of policy (policy may restrict retakes once the test is passed). Students in this situation may also demonstrate volatile retakes that should be flagged because they may be trying to gain knowledge of the test content to share with their friends who did not do well on their first attempt or they may have gained knowledge from their friends who shared information from their initial testing experience. Second, when the year-to-year scores within particular classrooms, schools, or districts demonstrate volatile retakes. Large year-to-year changes (in either direction) should raise a flag that suggests a variety of explanations (e.g., change in student population, change in administration/teaching emphases, cheating).

Table 5
Examinee Report

	A	B	C	D	E	F	G	H
1	Examinee ID	Test Site	Country	Passed	Score	Previous Score	Difference Z-score	
2	2022	1111	MEX	1	0.92	0.50	3.30	
3	8271	2222	JPN	1	0.90	0.43	3.83	
4	5723	2222	JPN	1	0.88	0.50	3.00	
5	6183	3333	USA	1	0.95	0.57	3.42	
6	6273	3333	USA	0	0.17	0.40	-3.65	
7	5778	3333	USA	0	0.00	0.67	-9.40	
8	6343	6666	AUS	0	0.05	0.92	-7.74	
9	8879	7777	SGP	1	0.85	0.00	4.23	

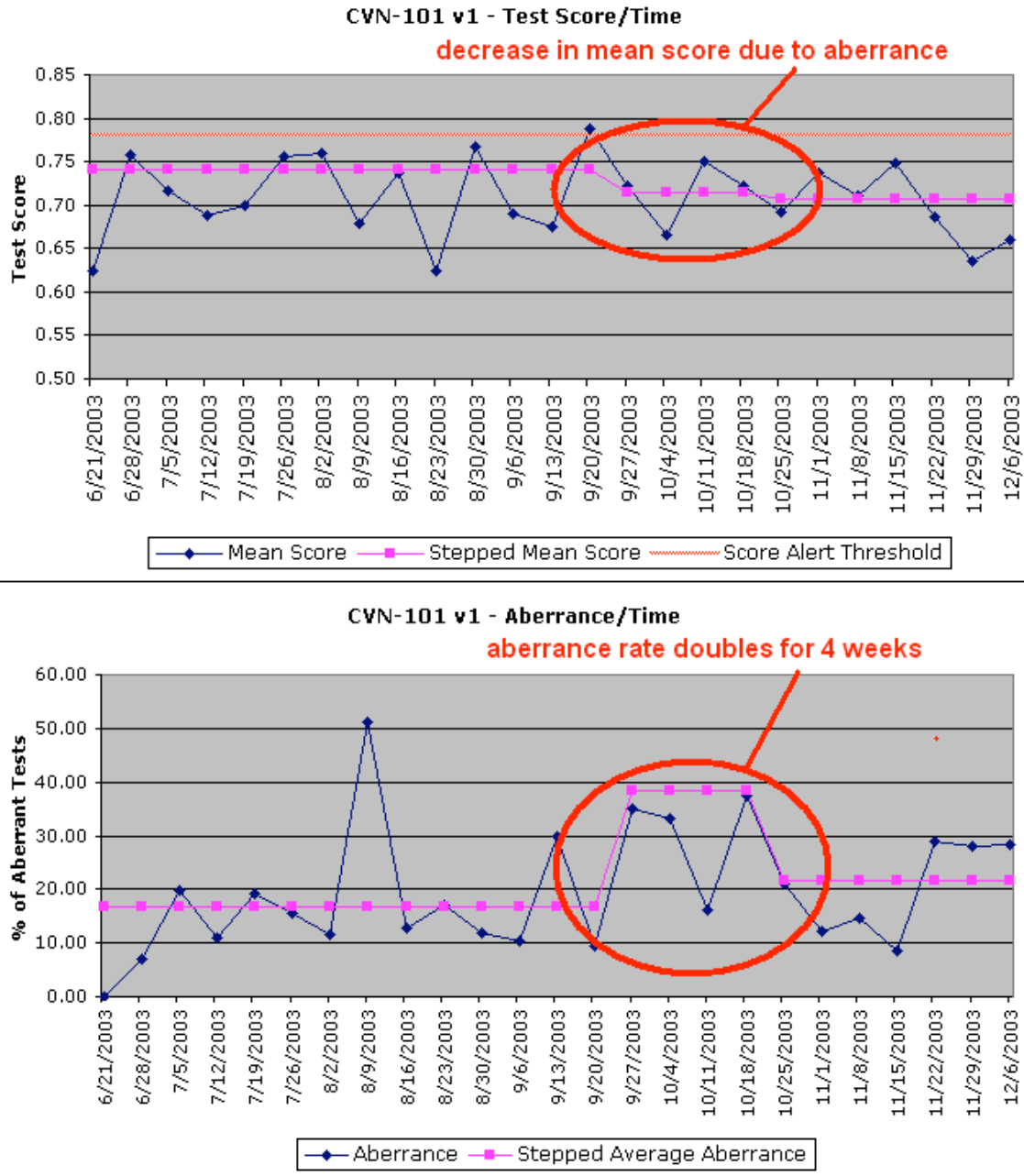
Effects on Test Performance

Caveon's data forensics can determine the effects of aberrance, collusion, and volatile retakes on the performance of the exam by also looking at item drift. Item drift is probably a misnomer given how item exposure occurs today. Drift implies a gradual degradation of an item's effectiveness. What is more often observed is not gradual; in some cases, it's immediate. "Rapid Deterioration" may be a more appropriate term. Instead of casually planning item replacement schedules; testing programs may need to detect and replace items (or entire test forms) when the compromise to item integrity is discovered. This can be problematic in a setting with a relatively wide administration window and a relative small item pool. One state that has recently instituted a computer-based testing program in schools, for example, expects its testing window to be about three weeks and there are fewer than 150 items in the pool. There is substantial risk that virtually all items will have been exposed by the end of the window. This has serious validity implications and it has implications for year-to-year equating and other important aspects of the program.

So, what effect does all of this have on the performance of the test? Is the test doing what it was designed to do? Is it performing as it did when it was originally published?

Caveon has been research these questions. Figure 3 shows an example where a strong relationship exists between aberrance and changes in test scores. In this case, an increase in aberrance in a continuously administered certification test over a 4-week period led to a real decrease in test scores over that same time. This suggests that the aberrance was indicative of a group of examinees who were likely memorizing test items.

Figure 3



It is a short step from there to also discover the real performance changes in items. Hambleton (2004) presented a method based on observed changes in typical item statistics. These efforts should be viewed as more than looking at item drift and test performance at particular points in time, but rather, using new methods to continuously monitor the exam and item performance.

The above is a general overview, using certification and licensure data along with the NWEA educational data, of how various measures of aberrance can be used to detect test fraud. The remainder of the paper provides results of the application of these measures to an ongoing educational assessment program that uses computer adaptive testing.

Four tests using test results from 20,661 test administrations were used to check for evidence of test security compromise, including cheating and piracy. The reporting period of the analysis was from March 1, 2004 to June 30, 2004.

Identification of modified records

In order to test the approaches used to identify cheating, 2 school districts, 5 schools, and 3283 students were identified as "cheaters". The response records for these individuals were modified to measure the data forensics detection rates.

Ten percent of the items from the CAT item bank were marked as exposed. Whenever a test record from the modified set contained one of the exposed items, that item response was changed so that it was correct (unless it was already correct, then no change was made). As noted in a footnote above, only that item was changed so the conditions did not mimic the actual testing situation in which the examinee might have been routed to a different next item that might have resulted in an aberrant response pattern that would have been more readily detectable. The number of items given to the students on the CAT tests was at least 50, which was also the modal test length. Therefore, on average 5 test questions would have been modified on each test. Review of the CAT data shows that most generally the probability of responding correctly to a question is 50%. Therefore, the responses for 2.5 questions on average would have been changed from incorrect to correct. This is a very small amount of change and cheating in such low incidence situations is very difficult to detect.

The summary of the results of Caveon's data forensics analysis are:

- One of the two school districts as having a high cheating rate was identified.
- None of the 5 schools as having a high cheating rate was identified.
- Forty one of the 3283 students was identified as being potential cheaters.

In this analysis Caveon's Data Forensics examined four types of test security risk:

1) Collusion -- answer copying, collaboration and communication during testing such as text messaging, teacher coaching and proxy test taking,
2) Cheating -- having advanced knowledge of some or all of the exam content,
3) Piracy -- stealing test items by memorization or technology, and
4) Volatile Retakes -- extreme score changes between successive test administrations.

Table 6 lists the tests and some general test details:

Table 6: Overall Test Summary

Test	Security Assessment	Number of exams	Testing sites	Pass rate[5]
NWEA-394	Moderate/High	5056	16	49
NWEA-396	Low	5304	48	50
NWEA-704	Low	5138	76	49
NWEA-708	Low	5163	26	49

One test, NWEA-394, is deemed to have significant risk of test compromise.

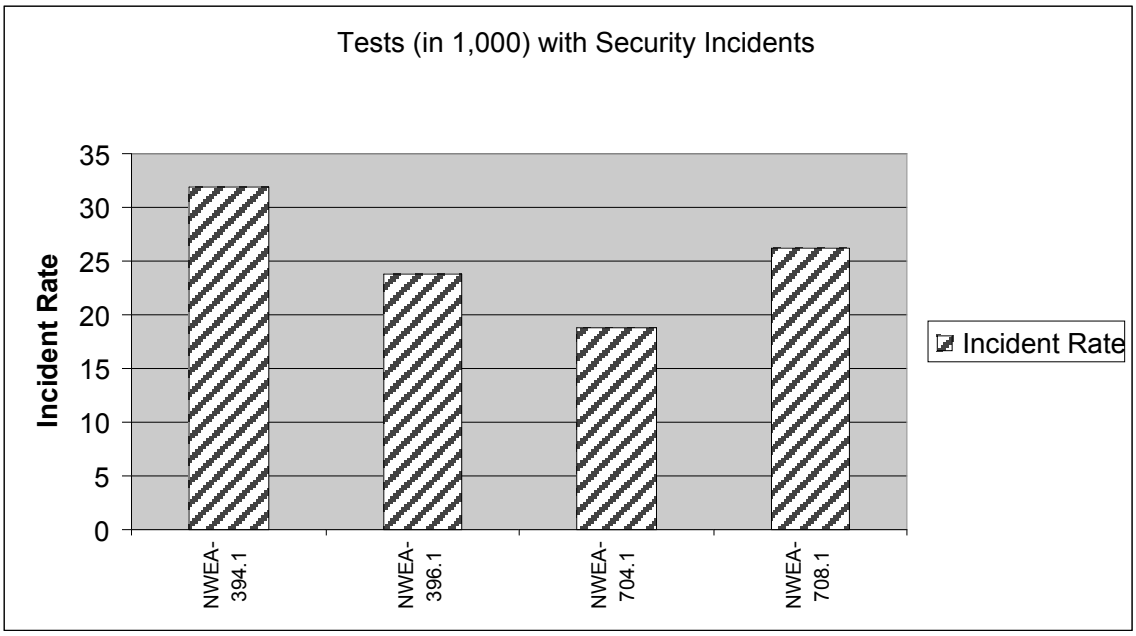
[5] Pass rates were not defined for this test; to demonstrate the analysis a passing score was set arbitrarily at the median score. Thus, the pass rates should be close to the 50th percentile, subject to the granularity of the score distribution.

One test, NWEA-394, is deemed to have significant risk of test compromise. The reasons for this determination are discussed below.

Security Incidence Overview

Figure 4 compares the security incidence[6] rates for each of the tests. This figure shows the proportion of tests for which any security incident was identified. A security incident occurs when the scores demonstrate aberrance, collusion, piracy, or volatile retakes.

Figure 4: Overall Security Incident Rates

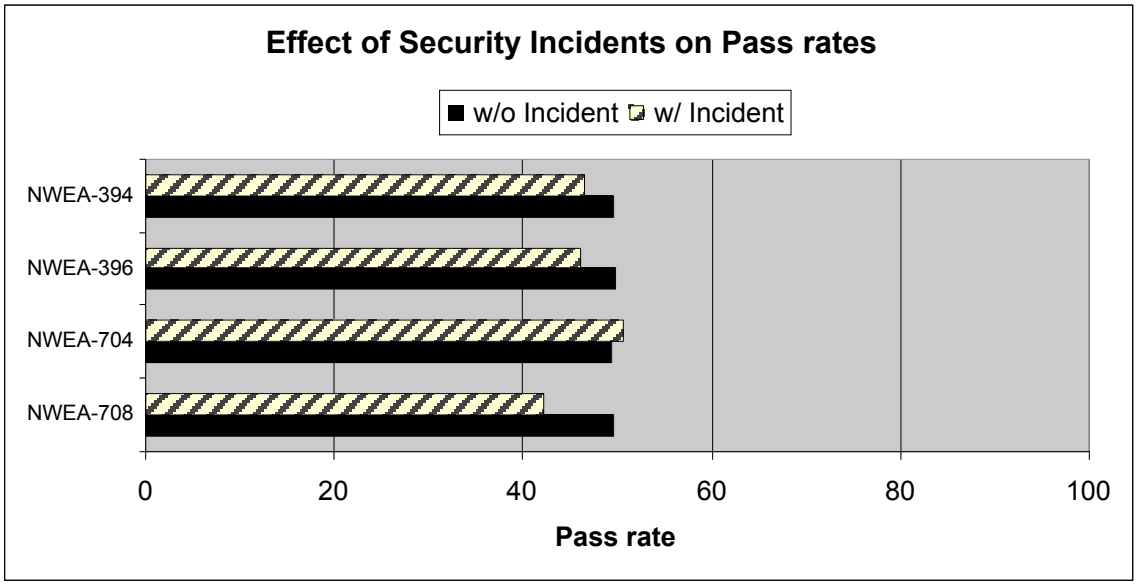


The proportions are based on tests per 1,000. For example test NWEA-394.v1 has 161 measured incidents in 5056 tests administrations, which is a rate of 32 per 1,000 or 3.2%.

The effect of security incidents on the test pass rate is shown in Figure 5. This figure shows the relationship between the pass rate (a passing score was set artificially for this test at the median, because the test is used for general accountability and there was no passing score set).

[6] An incident is measured and counted when the statistic being compared is extreme. As such, a security incident should not be construed as confirmation of an actual security compromise. It should be interpreted as an event indicating risk to the test's security.

Figure 5: Effect of Security Incidents on Pass Rates



The pass rate varies among the exams when comparing exams with and without security incidents across the different test forms. For example, for test NWEA-394, 46.6% of the exams that had a security incident resulted in a passing score, whereas the pass rate is higher at 49.5% for tests taken without a security incident. Normally, we have seen the opposite effect on pass rates. Having recently reviewed other educational data, two explanations for this effect are offered.

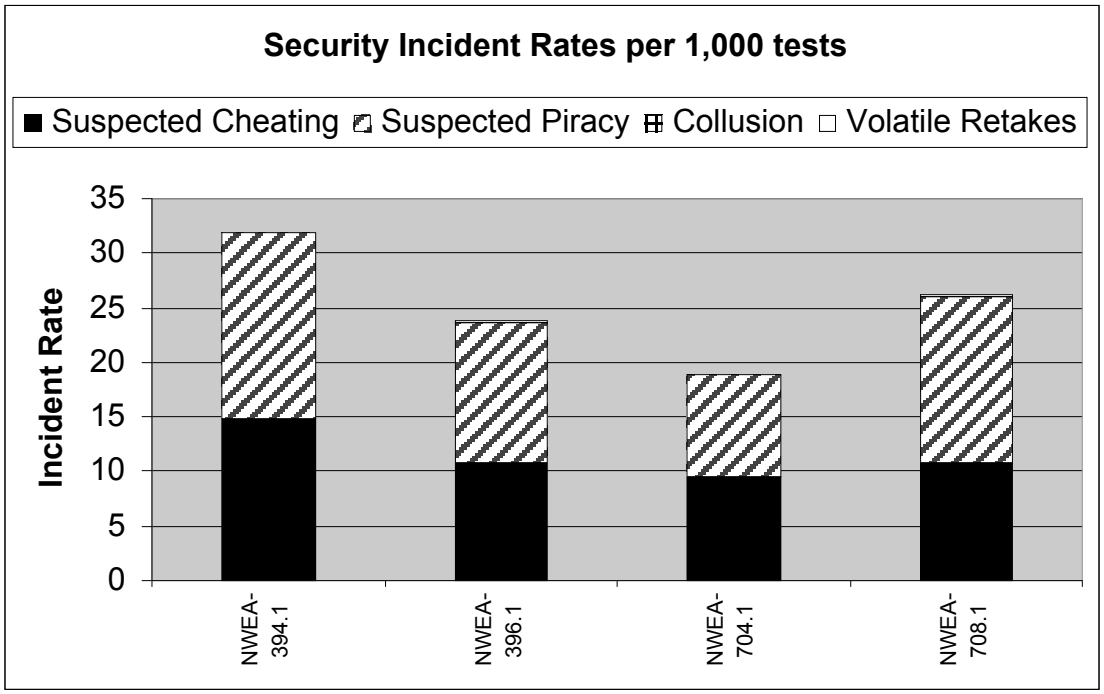
First, when students are not prepared to take the exam, high aberrance may be seen in the lower score range. It is likely that these students are receiving help to raise their scores, but without adequate preparation, the help is insufficient. A simple comparison of pass rates cannot detect this effect.

Second, the passing score selected for the analysis is not likely to be the actual passing score. The passing score for the analysis was set arbitrarily at the median. For some educational tests used for accountability purposes (e.g., NCLB) there may not be passing scores, per se. There may be cut scores associated with different proficiency levels, however, and these scores would be useful for this analysis. The passing threshold (or other classification cut scores) will nearly always be sensitive to the tail of the distribution (either low or high, depending on the targeted passing rate). Depending on the nature of the test compromise, aberrance rates will vary with the test scores. For example, if a group of "over-achieving" students have come together to "ace the test," then aberrance will be seen at the high score levels. If test coaching is concentrated on the students in the middle of the score range then aberrance will be observed at and above the middle of the score distribution. These kinds of behavior that compromise the test may not be directly measured as an effect on the pass rate when the passing score varies or when there are multiple cut scores for multiple classifications.

Figure 6 provides the rates[7] of security incidents per 1,000 tests for the tests. Stacked bars are used to show all the information.

[7] It is possible to observe more than one security incident on a test, thus the stacked bars in this chart are not precisely the same height as the bars in the Figure 1.

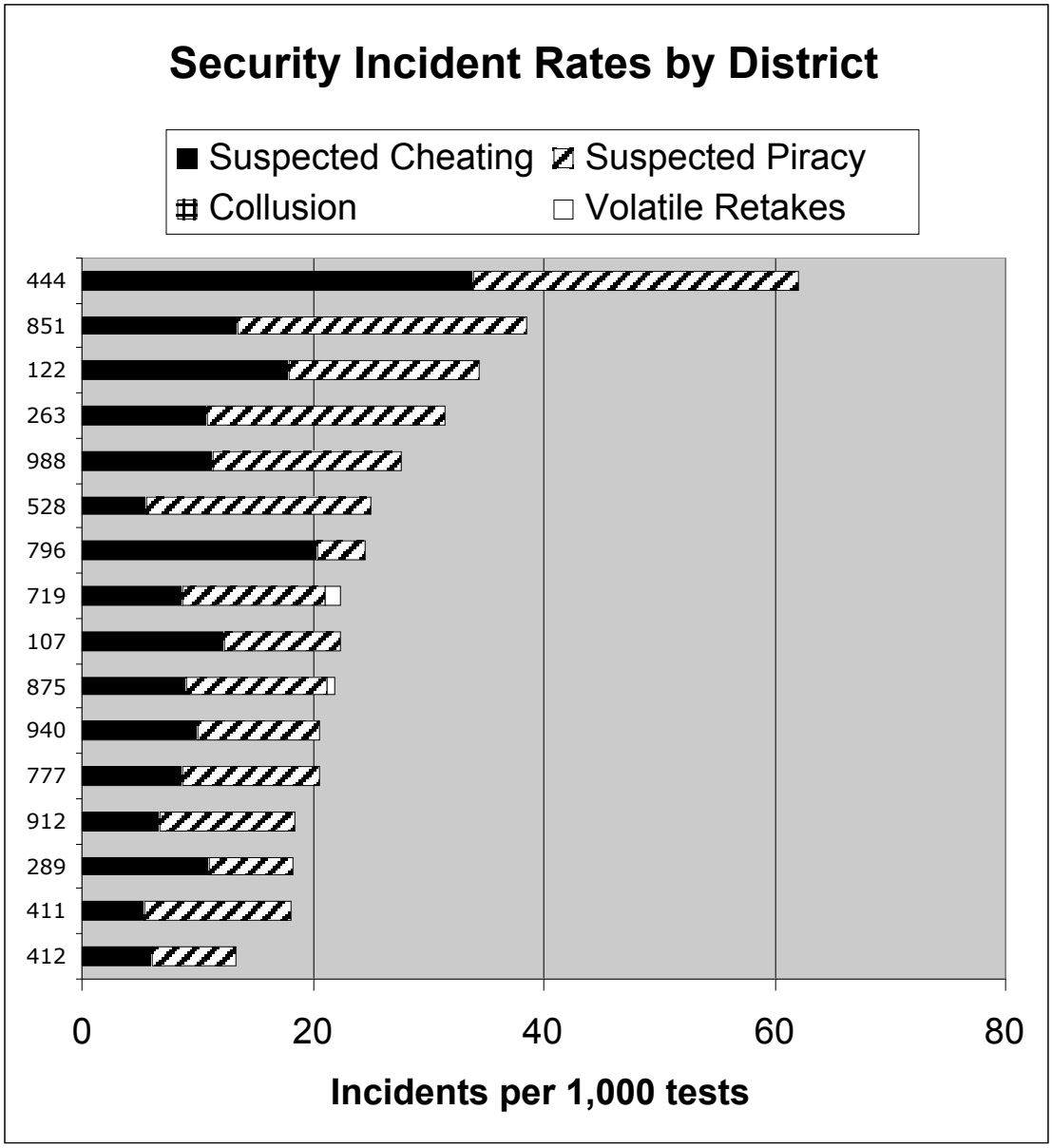
Figure 6: Security Incident Rates by Type of Incident



The comparison of the rates between the tests is illuminating. The greatest observed difference is between tests 394 and 704. The relative proportions between “suspected cheating” and “suspected piracy” for the exams appear to be relatively constant. This is probably a result of the CAT (Computerized Adaptive Testing) format for the exams which will discourage cheating and aberrance.

Figure 7 presents the breakdown of security incident rates by the districts in the study.

Figure 7: District Overview of Security Incidents



As was already noted, the two prevalent types of incidents are suspected cheating and suspected piracy. Two of the districts had relatively high rates of incidents that would raise a flag about the validity of their test results.

A total of 237 students were identified as having odd response patterns or odd latency. Forty-one of these were among the 3283 students with modified response patterns. Thus, slightly over one percent of the total number of examinees was identified as having anomalous results, and about the same percentage of the students with modified response patterns was identified. It is not known at this time if the 41 students identified from the modified data set were among those who had the greatest number of score changes (they would be the lower scoring students who had the

most items changed – 12 to 14 items). Some of the students with modified responses had as few as one or two items changed, making it virtually impossible to detect them unless their subsequent test performance would have made the anomalous responses more obvious.

Of the 237 students who were identified as having aberrant response patterns, 150 were high performing students (above the 75th percentile) whose response patterns had not been modified.

Discussion

Although it is possible that the statistical procedure was identifying cheating behavior in the data, it was not able to identify the data records that had been modified to reflect students who had inappropriate access to 10% of the item pool. It might be that this manipulation, although seemingly very large, was, in reality, too subtle to be identified. It may also be that the procedures designed to work with fixed-form examinations don't work well with adaptive tests in which few students see the same items. It may also be that detection procedures designed for fixed length, linear exams (either paper and pencil or CBE) do not work well in an educational setting where we expect more variation in performance from one school to another.

The statistics that were used in this analysis are z-scores. These statistics are means of independent random variables and the Central Limit Theorem can be used to assume the statistical distributions are approximately normal. The critical value was chosen so that only 5% of the time would the maximum z-score in the sample of size 20,661 exceed the critical value. This value was set at 4.7. The alpha-level of this value is .000001, or about 1 chance in a million. Even though the test statistic is not normally distributed, the normal approximation is sufficiently close that we can be assured that the alpha-level of the statistic is very small (<.001). The expected number of reported cheaters by chance alone in this study would be less than 10 (10,221 x .001).

Therefore we are left with a puzzle. Is the statistical procedure at fault, or is there a substantial amount of pre-knowledge already present in the unmodified data? Because the relative proportions of the detected cheaters in the known cheating set versus the unmodified data set are nearly identical: 1.25% versus 1.13%, we are left to conclude that the data forensics analysis shows no difference between the modified and unmodified data sets. Given the amount of known modification, we are left to conclude that cheating prevalence in the unmodified data is probably as large as the induced cheating prevalence in the modified data. The slightly lower rate in the modified data set could be due to the fact that items were changed only if they were answered incorrectly and no other responses were changed. In the unmodified data set, if the examinee answered correctly he or she was routed to a different item than if the item was answered incorrectly.

Simulation results indicate that the power of these test statistics is very low when students are armed with only 10% pre-knowledge. At an alpha-level of .001 on a 60-70 item test, the simulation results indicate power or detection rates for Caveon's best statistics are 6%. At an alpha level of .0001 in the same simulation, detection rates are approximately 2%. The simulation was not performed with extremely low alpha levels below .0001. The CAT algorithm, by virtue of adapting the test to the student's ability level, will in general lower the probability of a correct response, making pre-knowledge more difficult to detect than when the probability of a correct response is greater than 50%. This is because cheating detection relies upon finding item responses that are improbable (and usually incorrect). The improbability evidence is stronger when the probability of a correct response is higher.

Given the extremely conservative testing procedure used by the data forensics analysis and the low proportion of pre-knowledge on the exam, the above results are not surprising. Cheating detection is an extremely hard problem. The difficulty is compounded by the requirement that the procedures be conservative in order to minimize false positives.

Recommendations

Based on the analysis and results, some recommended actions that NWEA may want to consider are as follows:

General

- Ensure that test proctoring and administration procedures are being followed.
- Perform spot audits of the testing. Concentrate efforts in districts that are showing high incident rates and suspicious security related activity.

Exam 394

- Verify that this exam is functioning as designed and that the observed instability (as seen by a pass rate jump in tandem with high degrees of aberrance) was transitory:
- Monitor aberrance and pass rates for Exam 394 or review data from the exam administered during the 2004-2005 school year to ensure that the test has not been compromised. If it has been compromised revise the exam as schedule and resources allow.

Teachers

- Reduce and deter test “coaching” by teachers.
- Coaching appears to be occurring at a relatively low levels, but there are a few locations that indicate test coaching and other forms of inappropriate examinee assistance may be taking place. Reinforce exam administration procedures in training. Also, inform local district personnel of situations that should be monitored.

References

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G. J. (2001). An overview of issues concerning cheating on large-scale tests. Paper presented at the annual meeting of the National Council on Measurement in Education, April, 2001, Seattle, Washington.

Cohen, A. S. and Wollack, J. A. (in press). Test administration, scoring and reporting. To be published in Educational Measurement, Edition 4, 2005.

Tim Davey and Michael Nering. (2002) Controlling Item Exposure and Maintaining Item Security. In Computer-Based Testing: Building the Foundation for Future Assessments, Edited by Craig N. Milles, Maria T. Potenza, John J. Fremer, and William C. Ward. Lawrence Erlbaum Associates, Publishers, Mahwah, New Jersey.

Josephson, M. and Mertz, M. (2004). Changing Cheaters: Promoting Integrity and Preventing Academic Dishonesty. Josephson Institute of Ethics, Los Angeles, California.

McCabe, D. L. (2005). CAI Research. Center for Academic Integrity. http://www.academicintegrity.org/cai_research.asp.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., and Velasquez, R. (2004) Psychological Testing on the Internet: New Problems, Old Issues. *American Psychologist*, April, 2004, Vol 59, No. 3, 150-162

Appendix A
Glossary of Terms

Aberrance Threshold: The characterization of a test as aberrant hinges on the application of an "aberrance threshold"; a percentile on the aberrance score distribution for all tests above which differences in test-taking behavior (responses and response times) are deemed to be "significant" and indicative of test abuse.

Aberrance Rate %: This is the percentage of administered tests that were counted as aberrant (either the response latency aberrance statistic exceeded the threshold or the response aberrance statistic exceeded its corresponding threshold). An aberrant test exhibits response and response time values which significantly deviate from the test's normative response model. A test is characterized as "aberrant" if its aberrance score exceeds the aberrance threshold.

Aberrance Score: A statistic computed by comparing observed test response and response time patterns with a model of expected response and response time patterns. Deviations from the model (abnormal test response and response times) result in a positive aberrance score.

Alpha: The Type I error rate that is set for the statistical tests. Because multiple tests are performed (perhaps several thousand), the thresholds must be carefully adjusted to maintain the Type I error rate. Consequently, many results which would normally be reported as significant are not indicated as significant in order to avoid inflation of the Type I error rate.

Cheating Index: The statistical index that measures the test of significance for the high-score aberrance rate. The null hypothesis is that the high-score aberrance rate is the same as for all other geographical units in the world (excluding the unit being tested). The index is the absolute value of the logarithm (base 10) of the p-value of the test. This allows immediate interpretation of the index in odds language. An upper-tailed test is performed. High index values indicate aberrance rates for high-score tests above and beyond world levels. High values of this index indicate elevated levels of cheating.

High-Score Aberrance Rate: The percent of high-score (or passed) tests that are aberrant.

High-Score Threshold: A percentile of the distribution of all test scores above which a test is consider to be a "high-score test."

Latency Aberrance Threshold: A threshold normed against the standard normal distribution for counting whether a test is aberrant based upon the item response latency aberrance indices.

Low-Score Aberrance Rate: The percent of low-score (or failed) tests that are aberrant.

Mean Score : The average test score for all tests. The mean score of all test scores is typically the proportion of test items answered correctly, unless items scores are weighted.

Pass Rate Index: The statistical index that measures the test of significance for the pass rate. The null hypothesis is that the pass rate is the same as for all other geographical units in the world (excluding the unit being tested). The index is the absolute value of the logarithm (base 10) of the p-value of the test. This allows immediate interpretation of the index in odds language. A two-tailed test is performed. If the pass rate is lower than the expected value, then the index will be negative.

Pass Rate %: For tests where a passing standard is applied; the percentage of all tests which received a passing score.

Piracy Index: The statistical index that measures the test of significance for the low-score aberrance rate. The null hypothesis is that the low-score aberrance rate is the same as for all other geographical units in the world (excluding the unit being tested). The index is the absolute value of the logarithm (base 10) of the p-value of the test. This allows immediate interpretation of the index in odds language. An upper-tailed test is performed. High index values indicate aberrance rates for low-score tests above and beyond world levels. High values of this index indicate elevated levels of test piracy.

Response Aberrance Threshold: A threshold normed against the standard normal distribution for counting whether a test is aberrant based upon the response aberrance indices.

Report Section: Column in the report parameters page for the appropriate threshold level. For example, the threshold of 1.771 will be set for the world section when alpha is at .05.

Test Site: The location where a test was administered.

Tests: Count of number of tests.