

# Recent Innovations in Data Forensics

August 24, 2006

Dennis Maynes

Chief Scientist

Caveon Test Security

## ***Overview***

This document describes some of the innovations that have been introduced in Caveon Data Forensics since October 1, 2005. Several technical innovations have been made relating to the models and algorithms, but there are two innovations are especially apparent to customers.

## ***Technical Innovations***

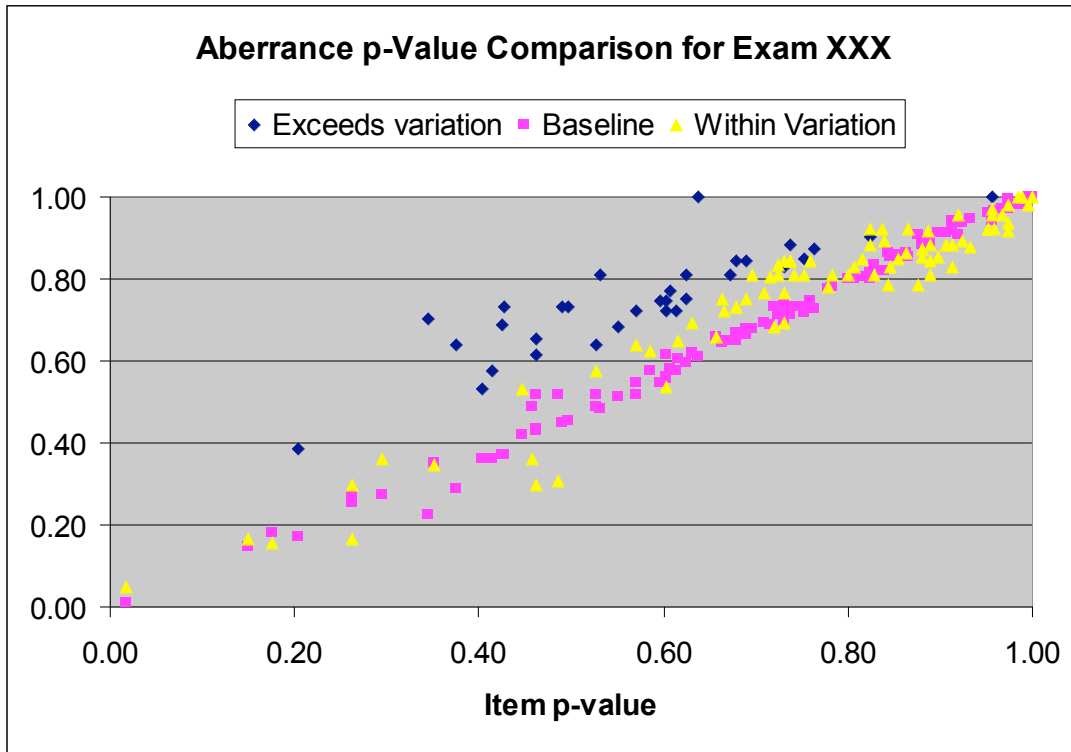
Caveon has enhanced its algorithms for estimating IRT models beyond the typical multiple choice format. The models are capable of handling hundreds of different responses to a question, through statistical pooling of response models. Typically, this means that as open-ended items (such as gridded-in responses) may be processed, using appropriate probability estimation techniques.

Caveon has developed techniques for estimating item latency aberrance in extremely speeded test situations. A speeded test inherently introduces aberrance in the responses and new techniques allow for detecting outliers in this constrained environment.

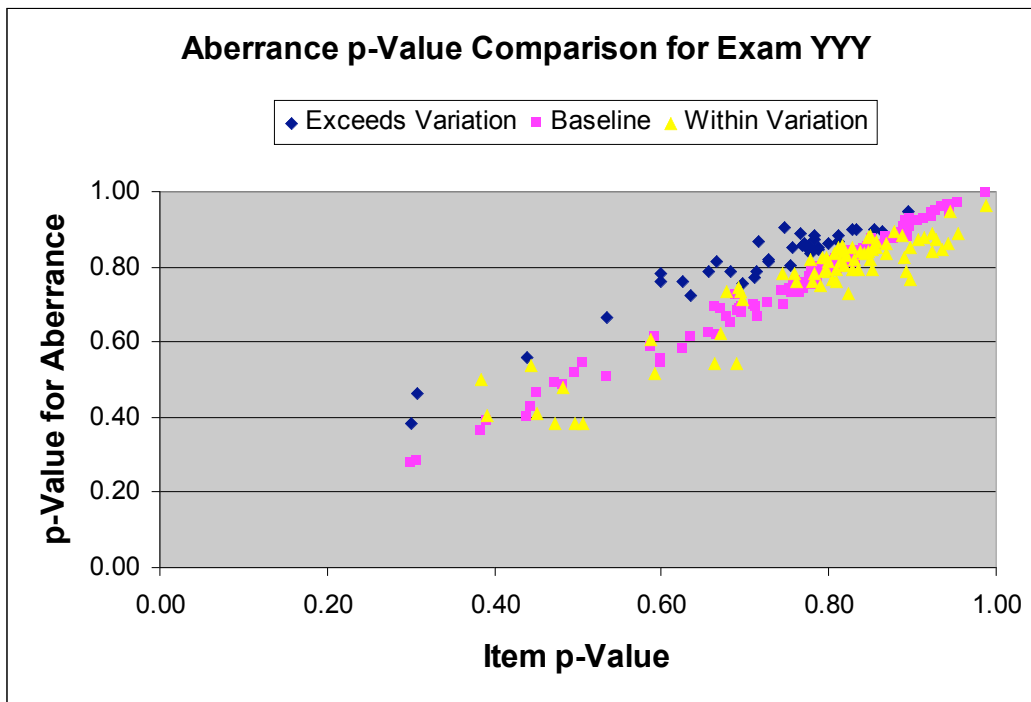
## ***Item Aberrance Analysis***

Caveon has enhanced the test analysis with Data Forensics to associate shifts in the p-values of individual items with aberrance. This allows for item compromise rates to be estimated. It also allows the creation of lists of items that are most likely to be compromised. There are two components to this new analysis. The most important component is a global, overall view of aberrance effects on items. This analysis is shown using a scatter plot similar to Figure 1 and Figure 2.

**Figure 1: Sample plot of Aberrance Impacts on Items – Exam XXX**



**Figure 2: Sample plot of Aberrance Impacts on Items – Exam YYY**

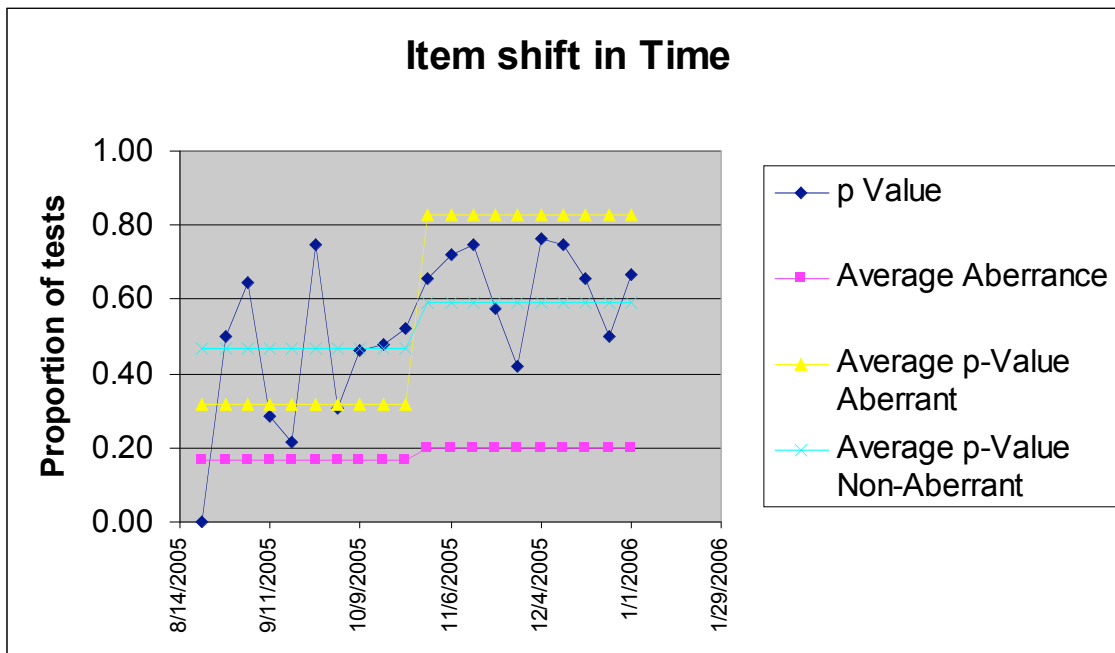


The pink squares in Figures 1 and 2 show item p-values for non-aberrant test takers. The blue diamonds and yellow triangles show the p-values for the same items that are associated with aberrant test takers. The blue diamonds represent items where the p-values for aberrant test takers are significantly higher than the p-values for non-aberrant test takers and, consequently, may be compromised.

The estimated compromise rate is derived from the number of items that are plotted using the blue diamonds. The plot for Figure 2 shows a moderately strong aberrance relationship. The plots tend to be sinusoid shaped, with the blue diamonds concentrated at the lower p-value ranges (i.e., the more difficult items) and the yellow triangles concentrated at the upper p-value ranges (i.e., the easier items). This is consistent with the adage that “Aberrance is getting the hard items right, while missing the easy items.”

A secondary component of the item aberrance analysis is the availability of item trend data. As a rule, these data are made available in the spreadsheets but are not plotted. However, it is easy to plot these data. Generally, we do not see temporal effects that are associated with the items, but it does happen. An example of this is shown in Figure 3.

**Figure 3: Example of an Item Shift due to Aberrance**



The blue diamonds plot the actual p-values that are calculated on weekly numbers. There was one change detected at end of October 2005 where the overall p-value increased for the item and aberrance appears to be associated with this increase. The light-blue x's represent the p-value of the item for non-aberrant tests. The yellow triangles represent the p-value of the item for aberrant tests. At the end of October, the data indicate that this item may have been subjected to a security breach. The aberrant test takers are now doing much better on the item than non-aberrant test takers.

## Refinements in Collusion Analyses

Caveon has introduced a refinement in its collusion analysis. This is a reporting refinement that is intended to help understand the nature of collusion and why the tests are similar. This refinement is not offered as a standard component of Caveon Data Forensics. Instead it is used during interpretation of the data to aid with displaying some of the more egregious cases of collusion. However, displays of this nature can be produced for specific test patterns, as requested by clients. Such displays could be useful in adjudication proceedings when score invalidations are considered.

A triplet of tests is shown in Table 1.

**Table 1: Example of Collusive Tests**

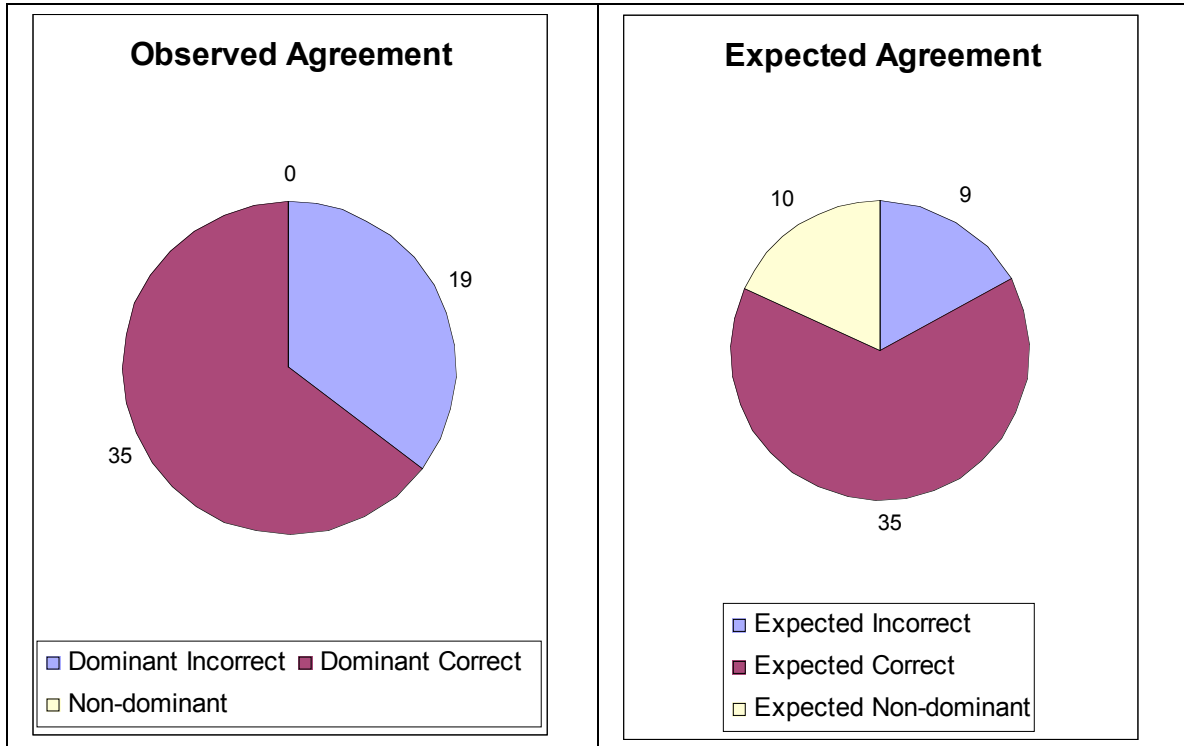
Item ID	Dominant Response	Dominant Score	Dominant Correct Prob	Dominant Incorrect Prob	376261	376262	376263
1	DE	0	0.646	0.086	DE	DE	DE
2	C	1	0.983	0.010	C	C	C
3	B	1	0.841	0.076	B	B	B
4	A	1	0.942	0.029	A	A	A
5	AD	1	0.787	0.101	AD	AD	AD
6	B	1	0.913	0.053	B	B	B
7	C	0	0.890	0.063	C	C	C
8	D	1	0.861	0.054	D	D	D
9	C	1	0.882	0.066	C	C	C
10	A	1	0.784	0.086	A	A	A
11	A	1	0.737	0.112	A	A	D
12	B	0	0.177	0.517	B	B	D
13	C	0	0.966	0.011	C	C	C
14	B	1	0.616	0.277	B	B	B
15	ABD	1	0.708	0.137	ABD	ABD	ABD
16	BD	0	0.213	0.316	BD	BD	BE
17	B	1	0.668	0.232	B	B	B
18	BC	1	0.848	0.049	BC	BC	BC
19	B	0	0.163	0.566	B	B	B
20	AC	1	0.147	0.250	AC	AC	AC
21	C	1	0.891	0.042	C	C	C
22	B	1	0.996	0.002	B	B	B
23	AD	0	0.063	0.528	AD	AD	AD
24	ACE	1	0.230	0.289	ACE	ACE	ACE
25	AD	1	0.711	0.108	AD	AD	AD
26	D	1	0.978	0.008	D	D	D
27	A	0	0.574	0.193	A	A	A
28	B	1	0.943	0.030	B	B	B
29	AB	1	0.543	0.163	AB	AB	AB
30	C	0	0.818	0.100	C	C	C

31	BC	1	0.388	0.293	BC	BC	BC
32	D	1	0.993	0.002	D	D	D
33	C	0	0.681	0.159	C	C	C
34	A	1	0.557	0.222	A	A	A
35	ACD	1	0.789	0.049	ACD	ACD	ACD
36	ACE	1	0.233	0.228	ACE	ACE	ACE
37	A	0	0.423	0.318	A	A	A
38	B	1	0.892	0.052	B	B	B
39	B	0	0.349	0.363	B	B	B
40	C	1	0.847	0.094	C	C	A
41	A	1	0.765	0.093	A	A	A
42	AB	1	0.518	0.165	AB	AB	AB
43	C	1	0.450	0.389	C	C	C
44	B	0	0.465	0.256	B	D	B
45	E	0	0.284	0.342	E	E	E
46	A	1	0.605	0.161	A	A	A
47	AE	0	0.478	0.118	AE	AE	AE
48	B	0	0.421	0.312	B	A	B
49	AC	0	0.744	0.117	AC	AC	AC
50	B	0	0.670	0.164	B	B	B
51	CD	0	0.496	0.373	CD	CD	CD
52	CD	1	0.417	0.313	CD	CD	CD
53	A	1	0.910	0.039	A	A	A
54	C	1	0.936	0.029	C	C	A

The analysis uses the concept of a dominant response. A response is dominant if more than half the test takers provided the response. In Table 1, correct dominant responses are highlighted using tan (or beige) and incorrect dominant responses are highlighted using gold. There are no non-dominant responses shown in Table 1. However, probability analysis indicates that at least 10 non-dominant responses were expected if the tests were answered independently.

Figure 4 provides a side-by-side illustration of the observed and expected agreement between these three tests.

**Figure 4: Observed and Expected Agreement**

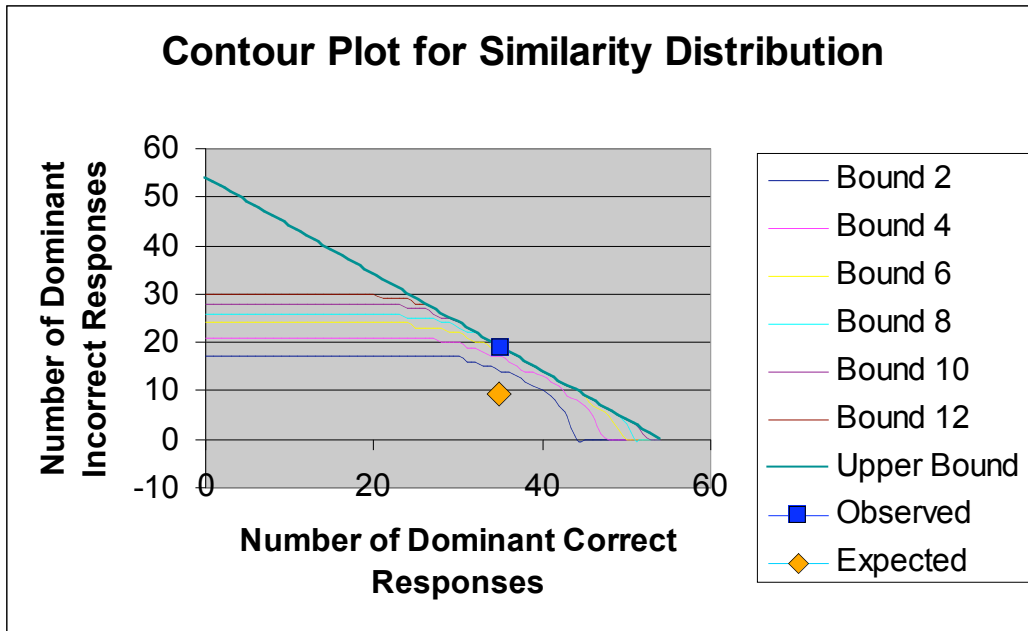


The left panel in Figure 4 provides the number of observed responses, whether they are dominant correct, dominant incorrect, or non-dominant. The right panel in Figure 4 provides the expected numbers of responses in each category under assumptions that the tests are independent.

The probability of this level of agreement between the tests is 1 in  $10^6$  or 1 million.

The probability space for the distribution of these tests is shown in Figure 5.

Figure 5: Contour Plot for the Similarity Distribution



The blue square in Figure 5 represents the observed level of agreement. The orange diamond represents the expected agreement. The contour lines are at increasing powers of 100. The first line, Bound 2, represents a probability level of .01. The second line, Bound 4, represents a probability level of .0001. And so forth, until the last line at Bound 12 represents a probability level of 1 in  $10^{12}$ . The upper bound is the absolute limit of the counts and represents 54 which is the number of items on this exam.