Caveon Data Forensics™ - Legal Defensibility of Scoring Decisions

by Robert Hunt

VP Client Legal Support & General Counsel, Caveon

Statistical evidence of test fraud is generally expressed as a probability that an examinee's test response patterns agree with one or more statistical models of cheating and in some cases, test theft. The question addressed in this paper is whether such evidence can, or should be, be regarded as sufficient evidence of cheating to justify the cancellation of test scores and other actions.

As an initial matter, it's worthwhile to remember that all types of evidence can, when appropriately qualified for relevance, error, subjectivity and veracity, be regarded as probability statements about events. Evidence of cheating supplied by a test proctor, for example, is susceptible to errors of observation and memory.

The probability that an examinee copied the work of another based on statistical similarity of test responses, accordingly, can (and has) been viewed by certain courts as credible evidence of "misconduct"—upholding the efforts of large testing programs such as the SAT® and ACT ®, to use statistical analysis as a means of policing test fraud.

Why is data forensics so appealing to these large-scale testing programs? The answer is that data forensics can reveal test-taking behaviors that are "invisible" to other types of surveillance including wireless communication between examinees, or prior access to secure test information on the Internet. It can also be comprehensively and systematically implemented with little additional cost and effort.

Certain types of data forensics may also provide better evidence of cheating than other "first-hand" sources because the strength of the finding can be quantified and expressed with some precision (e.g., the chances that the similarity between two examinees test results could have occurred by chance are "one in a million"). Due to subjectivity and the potential for error by contrast, "first-hand evidence" of test fraud is rarely so specific, or emphatic.

Belief in these capabilities and results however, involves the further belief that the behavioral models underlying the statistical analyses, accurately describe (and can therefore identify) cheating and other forms of misconduct. Copying analyses, for example, hinge on the assumption that above a certain threshold, the similarity between two examinee's test records could only have occurred as a result of cheating.

That the opportunity created by adjacent seats resulted in test copying however, is perhaps the firmest, most intuitive behavioral model available to data forensics experts, others such as gain-score analyses, involve more complex models of examinee behavior and less intuitive assumptions (e.g., that legitimate test preparation methods are incapable of producing score changes of certain magnitudes). Still others involve models that seek to explain departures from normative test-response behaviors; behaviors which are intrinsically very difficult to study.

Within this context of possibilities, methods and assumptions, the purpose of the following discussion is to identify the conditions under which testing programs may reasonably rely on

statistical findings as a "trigger" for actions need to preserve the integrity and security a program's tests.

When Will Statistical Evidence Suffice?

Currently, few testing programs, including those which routinely conduct statistical analyses of test results, are willing to rely on statistical findings as exclusive evidence of test fraud. Explanations for this likely range from misunderstanding and discomfort to inertia, all of which testing programs can be partially forgiven.

Anglo-American legal tradition has a longstanding preference for "first-hand" evidence. Only since the introduction of the Federal Rules of Evidence in 1975 have federal courts welcomed scientific, technical, or other "specialized knowledge" in deciding issues of fact, without the necessity of peer review and other indicia that the underlying science is generally accepted. "This relaxation of the usual requirement of first-hand knowledge" the Supreme Court stated in *Dauber v. Merrill Dow* (1993), "is premised on an assumption that the expert's opinion will have a reliable basis in the knowledge and experience of his discipline."

Adding to the confusion for testing programs, the few courts that have examined the specific issue of the admissibility of statistical evidence, have done so under labels other than cheating (i.e., the "validity" of a test result), and under different lines of analysis: contractual and constitutional.

Contractual Analysis

In *Langston v. ACT* (11th Circuit, 1989) and *Murray v. ETS* (5th Circuit, 1999) two U.S. Circuit Courts of Appeals examined and upheld the use of gain-score and copying analyses in as a valid exercise of testing program rights under test-use agreements. Where such agreements exist, the courts concluded, testing programs have the right (if the test-use agreement so provides) to investigate possible test fraud, and if necessary, to cancel test scores. Crucially, the courts stated that the decisive issue in each case was not whether the examinee cheated, but only whether the testing program carried out its investigative rights in "good faith."

As a result of this focus on contract performance, neither court was much interested in a detailed evaluation of statistical methods and findings. In *Murray*, for example, the court ruled that ETS fulfilled its contractual obligation "by following established procedures for determining the validity of questionable scores."

> ETS provided the district court with substantial evidence regarding its reasons for questioning Murray's scores and the procedures it followed to determine whether Murray's score should be withheld. Moreover, ETS provided the district court with copies of its policies and procedures, as well as the testing agreement which every student must sign before taking the SAT I.

The *Langston* court similarly indicated that the question of whether ACT fulfilled the "letter of its contractual promise" depended on the overall fairness of the score cancellation process, paying only marginally more attention to the type or quality of ACT's investigative efforts:

> ACT's investigation of plaintiff's scores was extensive. ACT considered the dramatic increase in plaintiff's scores, the alarming similarity between plaintiff's answers and those of test number 413619, plaintiff's self-reported grades, and the letters that plaintiff forwarded to ACT.

As in *Murray*, the *Langston* court did not evaluate the methods or assumptions connected with a "dramatic increase" in scores, though it clearly credited ACT for having conducted the analyses, and otherwise encouraged its efforts to police test fraud with "second-hand" evidence:

> To demand that ACT prove by eyewitness testimony that an individual cheated before invalidating a score would undermine ACT's primary function of providing colleges with scores that are highly reliable. ACT could not possibly catch every student who cheats on its exams if it had to produce an eyewitness to confirm every instance of misconduct.

As we have seen, under the Rules of Evidence, federal courts have the responsibility of assessing the relevance and reliability of "second-hand" evidence developed by experts. In both *Langston* and *Murray,* the courts tacitly acknowledged that the testing behaviors identified by gain-score and copying analyses are somehow indicative of test fraud, and therefor relevant and reliable evidence of a testing program's good-faith effort to investigate its tests for the presence of test fraud.

This contractual line of judicial analysis suggests several important conclusions. First, testing programs which are able to form binding test-use agreements can establish rights to police its tests using statistical analyses of test fraud, if in the agreement also confers the right to cancel test scores when the program is unsure of the test result.

Second, courts will only (if at all) examine the results of a testing program's statistical analyses to confirm the persuasiveness of its findings—not to ascertain the accuracy of the method or the finding. Finally, based on these other conclusions, it is reasonable to believe that federal courts would be receptive to more complex types of data forensics, such as response and latency aberrance; especially if the underlying behavioral models have the intuitive appeal of a copying or gain-score analysis.

Constitutional Analysis

The other line of analysis alluded to earlier, involves a different set of labels and legal theory, although the outcome for testing programs is similar. The legal theory derives from the due process protections embedded in the U.S. and state constitutions which require "state actors" to exercise care when their actions affect private interests. A state licensing authority, such as a bar association, is good example of a state actor.

The form of care required is described within the legal doctrine of "due process" which has two well established aspects; substantive and procedural. Two cases provide specific guidance in the application of these requirements to statistical evidence of test fraud; *Murray* and *Scott v. ETS* (1991)

Scott involved the cancellation of an examinee's score after her third attempt to pass the National Teachers Examination. On the last attempt, ETS detected an improbable score gain as well as similarity between the Scott's test-responses and those of another examinee (who was not seated near her). The probability that the similarities between the test records could have occurred by chance, ETS concluded, was less than 4 in 10 million, and on that basis, ETS cancelled Scott's score but offered her an opportunity to retest.

On appeal, Scott alleged that ETS (acting on behalf of the State of New Jersey) had denied her due process rights by failing to hold an evidentiary hearing, and by failing to prove that she had actually cheated. The New Jersey Superior Court rejected both arguments, ruling that ETS only needed to have formed a "substantial question" regarding the validity of Scott's score in order to invalidate her test result.

In the fashion typical of decision-making on constitutional issues, the court examined the important interests effected by ETS's decision before concluding that "they express a common concern that ETS test scores be reliable."

> The relevant interests here are several. Plaintiff has a legitimate interest in assuring that she is not stripped of a valid test score. ETS has an interest in assuring the accuracy of the test results it reports and the predictions it thereby makes. The other test-takers are entitled to assurance that no examinee enjoys an unfair advantage in scoring. The school officials to whom test results are certified need to be assured that all reported test results are reliable. Finally, the public at large has an interest in assuring that all persons certified as teachers have in fact fulfilled the requirements of that certification.

Although *Murray* was premised on a contractual analysis, the court engaged in a similar evaluation and prioritization of competing interests in an examination of ETS's right to investigate test scores. Ultimately, the court concluded that "to the extent that [ETS] can accurately predict the aptitude of a candidate by means of its test results it performs a highly valuable service not only to the [schools] but to the public as well."

By making the trustworthiness of test scores the preeminent interest, in both cases the courts spared ETS from having to prove test fraud—allowing it to satisfy a much lower standard of proof.

> … her argument that a finding of unreliability must be tied to proof of wrongdoing is faulty. Proof of wrongdoing is one way of establishing unreliability; but if unreliability is otherwise shown, an absence of proof as to how it came about is of no matter. …. The fact that ETS had no proof of actual wrongdoing did not in any way undermine that showing of unreliability.

What level of evidence is needed to establish "unreliability" (i.e., the credibility of a test result)? Here *Scott* provides the only guidance. Without scrutinizing ETS's methods or findings, the court concluded that ETS's copying analysis provided "substantial evidence" of unreliability.

> ETS questioned plaintiff's scores on the basis of a statistical analysis showing hardly more than a 4 in 10 million chance that they were fairly earned. That gave ETS, and would give any other observer, substantial grounds for doubting the reliability of the scores.

Finding the Right Mix

The outcome of each line of judicial analysis described above (contract and constitutional) should buoy the interest of testing programs in the use data forensics. Because there are different types testing programs and statistical analyses, however, testing programs need to prepare differently.

A secondary purpose of the earlier discussion was to illustrate the different requirements of the two lines of legal analysis. In the *Langston* and *Murray* cases, both of which examined statistical evidence within the context of valid test-use agreements, the courts sought to answer a single question: did the testing programs act in good-faith in investigating and canceling test results? In both cases, the answer confirmed the use of data forensics to police and investigate examinee behavior.

In *Langston*, the court noted that a provision in the test registration booklet reserved to ACT "the right to cancel any test score if it finds reason to believe that the score is invalid due to testing irregularities or student misconduct." Applying state contract law (Alabama) the court continued that "the outcome of plaintiff's case does not turn on whether or not plaintiff cheated on his exam, but only on whether or not ACT carried out its contractual obligations in good-faith."

In its determination that ACT had acted in good-faith, the court cited three important findings: that ACT had 1) "engaged in an extensive investigation; 2) offered the examinee a retest, and 3) offered the examinee the opportunity to arbitrate." Importantly, the "extensive investigation" noted by the court consisted of a statistical gain-score analysis followed by a copying analysis, the results of which identified many identical responses shared by Langston and a examinee seated near him during the test.

In *Murray*, the court offered an even narrower interpretation of good-faith: "[t]he only contractual duty ETS owed to Murray was to investigate the validity of Murray's scores in good-faith." Without commenting on the validity of ETS's gain-score, copying and erasure analyses, the court concluded that "ETS dutifully fulfilled its contract with Murray by following established procedures for determining the validity of questionable scores."

The dictum that emerges from these examples for non-governmental testing programs is that test-use agreements invest testing programs with broad power of investigation and that data forensics can satisfy the program's responsibility to exercise that power in "good-faith." The only apparent requirements of those analyses is that they reflect an objective effort to identify signs of test fraud and other sources of uncertainty about the reliability (in the common, rather than the psychometric sense) of a test result.

The latter results suggests that within the context of binding test-use agreements, courts may be receptive to some of the emerging data forensics methods employed by Caveon such as aberrance analysis. With the marriage of sophisticated behavioral models of test fraud, sophisticated statistics and computing power, these methods can provide even greater investigative capability, and can be considered an even more concerted investigative effort.

Aberrance models and methods, on the other hand, lack the "accessibility" of copying and gain-score analyses (i.e., easily understood behavioral models) and may, under due process analysis, require qualification in the form of expert testimony, etc. to satisfy the "substantial evidence of unreliability" standard. That analysis would logically involve: 1) the behavioral models of the analysis; and 2) the method used to identify those patterns.

The *Langston-Murray* line of judicial analysis, to summarize, suggests that courts will support withholding/cancellation decisions premised on the results of statistical detection methods. This will be especially true where detection efforts are supplemented by measures such as offered retests, which provide test-takers with an opportunity to refute the finding.

Shifting to testing programs whose actions are scrutinized against constitutional standards, including for example the National Teacher Certification program at issue in *Scott*, courts have similarly supported the right of testing programs to use data forensics, but have stipulated that the results of those analyses must provide "substantial evidence" of the unreliability of the test result in order to justifiably serve as the basis for score cancellations and other actions.

While this requirement is clearly more rigorous than the "good-faith" standard applied under contractual analysis, demonstrating the "unreliability" of a test result is also much less rigorous than proving that the test-taker in fact cheated. Noting that "state actors must

afford due process to persons seeking admission to a profession or occupation," the *Scott* court arrived at the following compromise:

> We are satisfied that the relevant public and private interests are fairly accommodated by a procedure which permits ETS to cancel scores upon an adequate showing of substantial question as to their validity, without any necessity for a showing of actual cheating or other misconduct.

References

Daubert v. Merrell Dow Pharmaceuticals, Inc. 509 U.S. 579 (S.Ct., 1993)

Johnson v. Educational Testing Service, 754 F.2d 20 (1st Cir., 1985)

Langston v. Act, 890 F.2d 380 (11th Cir., 1989)

Murray v. Educational Testing Service, 170 F.3d 514, (5th Cir., 1999)

Scott v. Educational Testing Service, 600 A.2d 500 (N.J. Super., 1991)