federation forum
Spring 2011

Data Forensics

By John Fremer, President, Caveon Test Security

This article was developed from a presentation given at the Federation's 2010 annual meeting in Denver, Colorado.

Caveon Test Security didn't invent data forensics, which is a statistical methodology widely used in law enforcement and other investigations. However, in 2003, Caveon introduced it into the testing field, and now there are people working for other companies called data forensic analysts. There were no such things until Caveon came along in 2003.

Millions of people who have never met me hate me because of my past 35 years with various education testing services. I have had a hand in many major exams that they have endured. For instance, I led the team that revised the Scholastic Aptitude Test (SAT). We met all of our schedules and all of the criteria and we were all better friends when it was completed than we were when it began. It's probably one of my proudest accomplishments. I have been at Caveon, which is employee-owned and financed, for seven years. In addition to data forensics, we do security audits and investigations. Data forensics, though, is our best-known service.

Data forensics is a special class of data analysis looking at actual responses for individual test questions that we call items. As opposed to mean scores or summaries, we can obtain much better information from individual responses from every test taker to every question on every single occasion. From that information, we create models, which indicate normal response to an item or items.

It takes more time to answer a question if there is a lot of content in the screen but sometimes that's not what you find. Sometimes you find people answer questions so quickly that they could not have read the questions, and sometimes you find that they spend amazing amounts of time on what you would think would be very easy-to-read questions. Then we look at the patterns of distribution of time and a range for them, but there are many people outside that range.

For certain people taking a test, a lot of answers are answered very quickly but other questions take an enormous amount of time - and there is nothing in between. That's not what you see if you look at 99.9% of test-takers. Why is it like that? There can be various unusual reasons why something happens that isn't cheating, but when there are flags on multiple indicators, we sometimes get bizarrely unlikely outcomes. The results in a project for the Atlanta public schools, for instance, were one over 10 to the 52nd power. That would be similar to flipping two coins with the first one landing on its edge and the second landing on the edge of the first coin and staying there.

Perhaps the first indicator of possible cheating is extremely high agreement among peers or groups of test-takers. The group always says, "Well, we study together, we had the same book and we had the same teacher." But if you just look at *all* people who had the same book and same teacher, you don't get results like theirs. Not only did they choose to get the same questions right and the same questions wrong, but when they are wrong they chose the same answers. How could that be possible on question…after question… after question? If you ask people if they did something wrong, they always say "No," whether it's an individual, a program or a school, so that's not useful information. If they said "Yes, we did it, we were wrong and we are sorry," that would be meaningful.

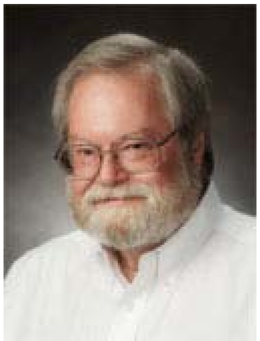
Another odd pattern is when there are multiple occasions of substantial gains or losses from one occasion to another. We only look for really amazing, extraordinary changes completely unlike what normal test-takers get and we still find them. Some test-takers' decisions are completely inappropriate.

FSBPT asks the toughest questions of any of our clients. It wants to know exactly what something means, why we reached a certain conclusion, how we do the analysis, if there are other interpretations and if we could run our data again. But once they are satisfied with a variety of different types of data, they want to take action. They use the data. They make it clear that there is no tolerance for cheating. That's not true for all of our clients. In some public education environments in which Caveon works, we bring unmistakable data and they don't want to take action. They don't want to deal with any of the interest groups that will protest. That means inappropriately behavior continues.

Other organizations, like FSBPT, use the information to good effect. The American Board of Internal Medicine used data forensics as part of an investigation and sanctioned 139 test-takers. That's not a world in which you want to be sanctioned. GMAT is another group that's very active in protecting its tests. The European association revoked 76 scores on its GMAT, which is its exam graduate management admission test. It's the door to going to accredited business schools in the U.S. They banned 58 testers, and notified 100 schools around the world of their actions. That's a good thing. We have to create the sense that cheating is not going to work and you are going to be sorry if you cheat.

Caveon searches the media worldwide and we put out something called 'Cheating in the News' every two weeks. At one level, it's depressing, but it's also informative because it explains how technology is being used to cheat. It helps point out why data forensics is so important.

Many high-stakes testing programs in licensing credentialing are now using data forensics. It's essential to act on the results. Don't just find problems; get to the root of them, take action and tell the public what you did. You must act on evidence of misbehavior to really applaud the fairness and validity of your exams. It's costly, but it's something you need to do to protect our programs.



John Fremer, PhD is the President of Caveon Test Security. He has 40 years of experience in the field of test publishing and test program development and revision, including management level positions at Educational Testing Service and The Psychological Corporation/Harcourt.

In his 35-year career at Educational Testing Service, Fremer led the ETS component of the team that carried out a major revision of the SAT. Fremer also served as Director of Exercise Development for the National Assessment of Educational Progress, and was Director of Test Development for School, Professional, and Higher Education Programs. During 2000-2003, Fremer designed and delivered measurement training programs to international audiences for the ETS Global Institute.

Fremer is a Past President of the National Council on Measurement in Education (NCME) and a former editor of the NCME journal ***Educational Measurement: Issues and Practice****.*

Fremer also served as President of the Association of Test Publishers (ATP) and the Association for Assessment in Counseling (AAC). He was co-chair of the Joint Committee on Testing Practices (JCTP) and of the JCTP work group that developed the testing-industry-wide ***Code of Fair Testing Practices in Education****; one of the most frequently cited documents in the field of educational measurement. Fremer is a co-editor of* ***Computer-Based Testing: Building the Foundations for Future Assessments****, (2002, Erlbaum.) and author of "Why use tests and assessments?" in the 2004 book,* ***Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators.***

John has a B.A. from Brooklyn College, City University of New York, where he graduated Phi Beta Kappa and Magna Cum Laude, and a Ph.D. from Teachers College, Columbia University, where he studied with Robert L. Thorndike and Walter MacGinitie.