



c a v e o n TM

Test Security

Caveon Data Forensics for District Assessment Coordinators

Steve Addicott, Vice President
Dennis Maynes, Chief Scientist

September 7, 2011

Key Takeaways

- This program is really important, and timely
- It's about valid test results
- We will get you comfortable with the program and its output

Outline of Presentation

- FDOE Data Forensics Goals
- Data Forensics Overview
- 2011-2012 FCAT Data Forensics Program
 - Students
 - Schools
- Q&A

HUFFPOST EDUCATION

THE INTERNET NEWSPAPER: NEWS BLOGS VIDEO COMMUNITY

Atlanta Cheating Scandal Unveiled By Reporters

The CHRISTIAN SCIENCE
MONITOR

America's biggest teacher and principal cheating scandal unfolds in Atlanta

At least 178 teachers and principals in Atlanta Public Schools cheated to raise student scores on high-stakes standardized tests, according to a report from the Georgia Bureau of Investigation.



More resign or retire in face of APS cheating scandal



c a v e o n™
Test Security

The Washington Times

Duncan: No link between cheating, NCLB

Places the blame on school leaders

“I think there is a **morally bankrupt culture** there,” Mr. Duncan said, referring specifically to Atlanta, where a recent government probe found that 44 schools and 178 teachers and principals had been faking test scores for the past decade.

August 25, 2011

Education on  msnbc.com

Pennsylvania is latest to face school cheating scandal



DCPS asks inspectors to review cheating charges

LIGHT FOR ALL
THE BALTIMORE SUN

EMT cheating scandal: Another black eye for Baltimore

Our view: It's better that officials are finding and punishing corruption than sweeping problems under the rug

CNN U.S.

34 N.J. schools investigated for possible cheating



e on™
Test Security

What Are We Talking About Here?

- Caveon's John Fremer says 1%-2%

"I estimate the proportion of educators involved in high stakes state testing who engage in cheating to be between one and two percent.

*You can find higher estimates. In the important and very widely read book "Freakonomics" a testing misbehavior estimate of five percent is provided, but I think that overstates the actual amount."**

*Caveon Security Insights, 8/24/2011

A Quick Exercise

- Assumption: 1% of teachers may “bend” rules
- Source: Nearly 170,000 teachers in FL, 2009-2010 School Year

$.01 \times 170,000 = 1,700$ teachers of concern, potentially

More Troubling News

Michigan Educator Survey*, July 2011

- 34% felt pressure to change grades for the better
- 29% felt pressure to cheat on standardized tests
- 21% know of an educator that changed scores on student tests
- 8% admitted to changing students' grades due to outside pressure

**Detroit Free Press, July 26, 2011*

US Department of Education

- **Special US DOE Summit on test security issues**
- **Key policy letters from Secretary of Education urging states to:**
 - **Make assessment security a high priority by reviewing and, if necessary,**
 - **Strengthen efforts to protect assessment and accountability data, to ensure the quality of those data, and enforce test security.**
 - **Includes running Data Forensics**

Use of Data Forensics

- Many high stakes testing programs now using Data Forensics
- Standards for Testing, e.g.,
“CCSSO’s Operational Best Practices for State Assessment Programs”

This work is important!

FDOE Data Forensics Goals

- Uphold fairness and validity of test results
 - Identify risks and irregularities
 - Take action based on data and analysis
 - “Measure and Manage”
 - Communicate Zero Tolerance to students and educators

Caveon Data Forensics™ Process

- Analyses of student results
- Create a “model” of typical test taking behaviors
- Identify EXTREMELY unusual behaviors with potential of unfair advantage

Caveon Data Forensics Process (cont)

- Examples of “Unusual” Behavior
 - Very high similarity among pairs or groups of test takers
 - Very unusual number of erasures, particularly wrong to right
 - Very substantial gains or losses from one occasion to another
- Focused on impact on scores

Similarity

- Our Most Powerful & “Credible” Statistic
 - Measures degree of similarity between two or more test instances
 - Analyze each test instance against all other test instances in the school
- Possible causes of extremely high similarity:
 - Answer Copying
 - Test Coaching
 - Proxy Test Taking
 - Collusion

Erasures

- Based on estimated answer changing rates from:
 - Wrong-to-Right
 - Anything-to-Wrong
- Find answer sheets with unusual WtR answers
- Extreme statistical outliers could involve tampering, “panic cheating”, etc.

Important! No student-level score invalidations based on erasure analysis; erasure analysis limited to school-level flagging for additional review.

Unusual Gains/Losses

- Predict score using prior year information.
 - Measure large score increases/decreases against predicted score
 - Which score truly reflects the student's actual ability or competence?
- Extreme Gains/Losses may result from:
 - Pre-knowledge, i.e., “Drill It and Kill It”
 - Coaching
 - Student development—visual acuity

Important! No student-level score invalidations based on gains analysis; gains analysis limited to school-level flagging for additional review.

2011-2012 FCAT Data Forensics

- Focus on two groups
 - Student-level
 - School-level

- Flagging only the most extreme results

A Quick Discussion of Extreme Results....

- Chance of being hit by lightning = 1 in one million
- Chance of winning the lottery = 1 in 10 million
- Chance of DNA false-positive = 1 in 30 million
- Chance of students being flagged for similarity, but doing own work = 1 in a trillion

Student-level Analysis

- Similarity Analysis only
 - Most credible, strongest
 - No flagging for erasures or gains
- Invalidate test scores with Similarity Index ≥ 12

Similarity Index ≥ 12

- Extremely remote chance of a false positive
 - Chances of seeing two (or more) students' tests so similar, with each doing his/her own work:

0.00000000000001

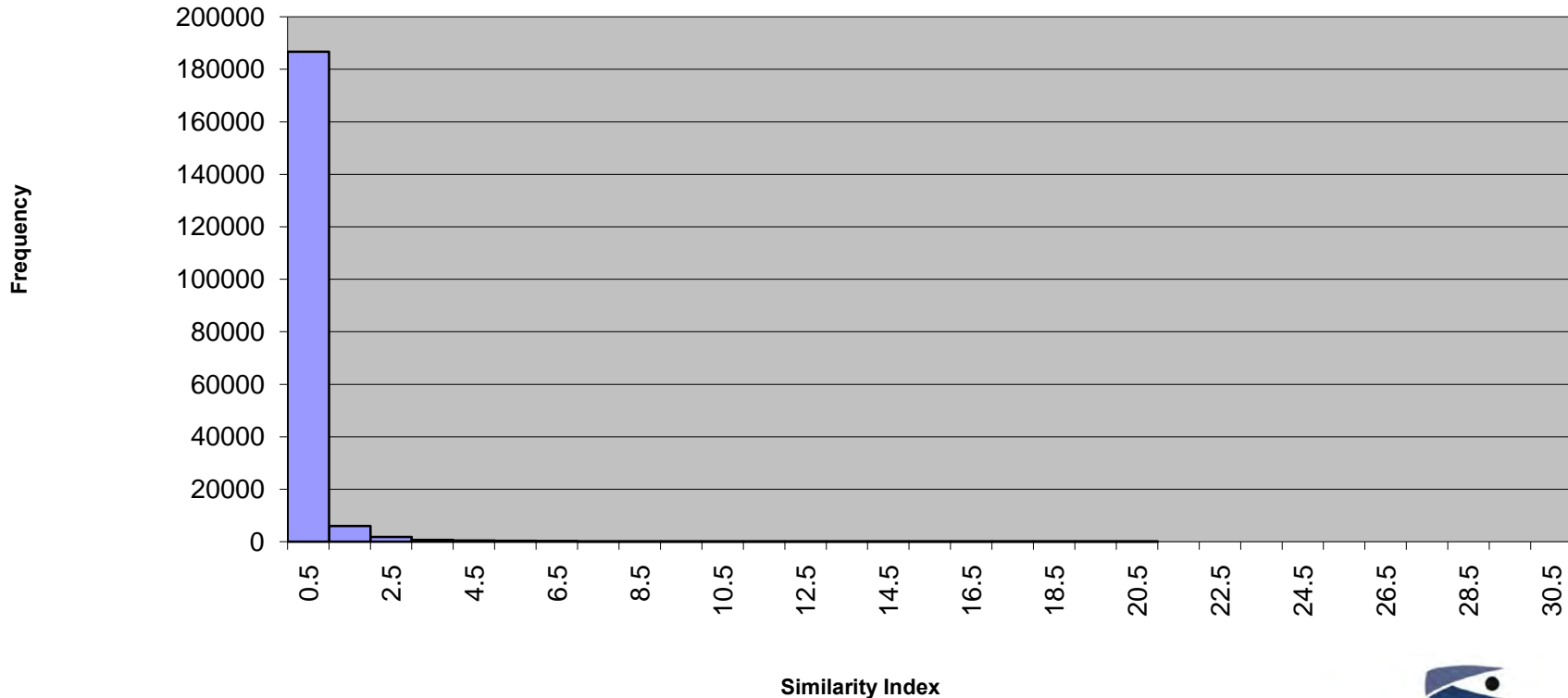
- Does not equal cheating, means test score is not trustworthy
 - Fairness and Validity of test instance must be questioned
 - Appeals process to be implemented

Steps for Calculating Similarity

1. Use student's performance to compute the probability of an incorrect/correct answer on all items
2. Calculate the probability that two students will answer an item identically
 - Expected Identical Correct/Incorrect
 - Observed Identical Correct/Incorrect
3. Tests are flagged when the number of identical responses is *much* greater than the expected number
 - *Identical incorrect answers are much more revealing than identical correct*

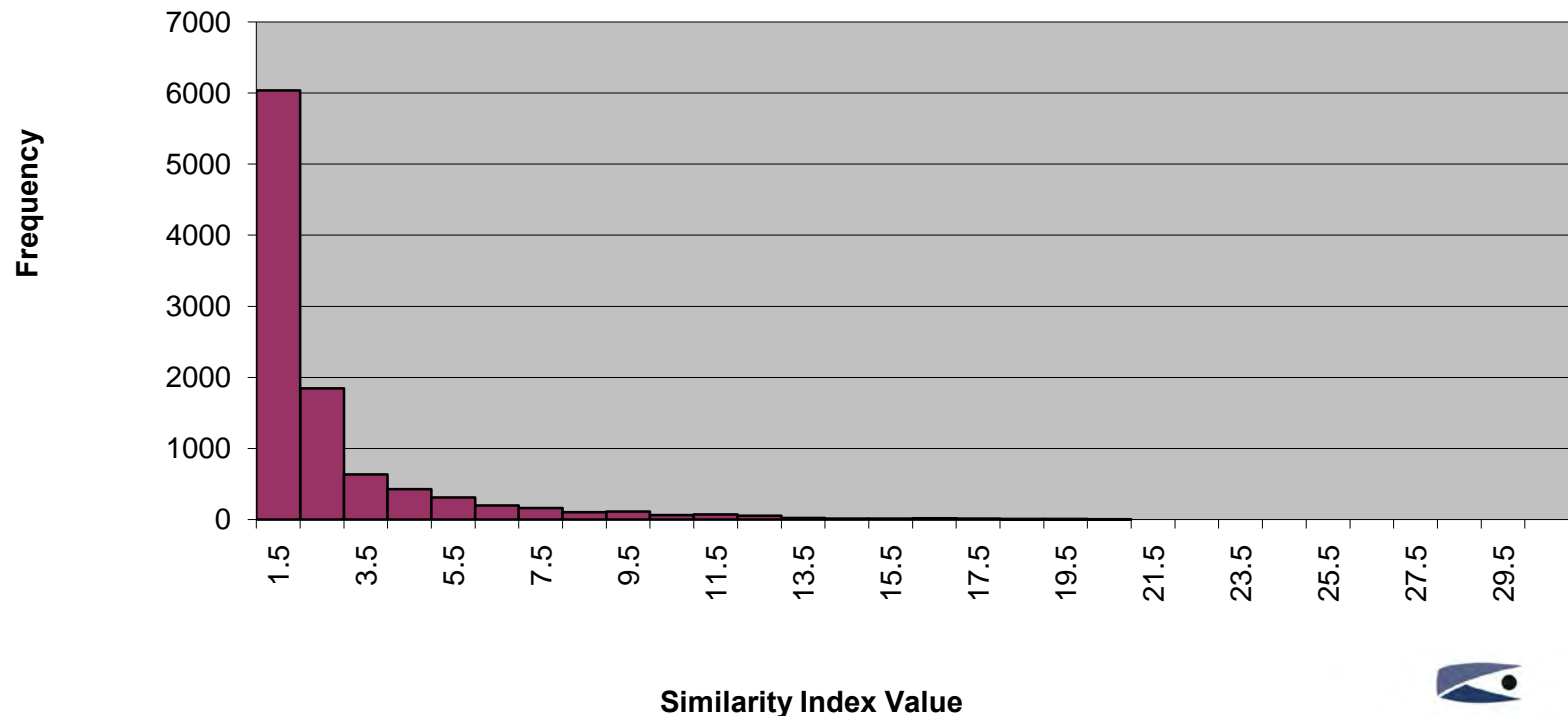
A Visual Representation

FCAT Spring 2011 Grade 8 (Reading)
N=196,866, Above 12=163 (0.08%)
Mean = 0.3



A Visual Representation

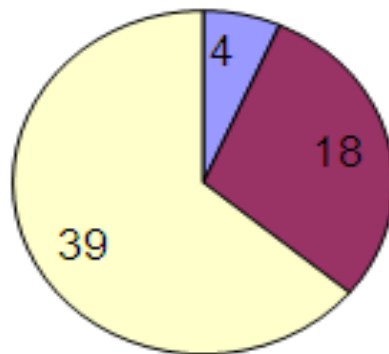
**FCAT Spring 2011 Grade 8 (Reading)
With first bar removed
Index ≥ 12 is 15 standard deviations above the
mean of 0.3**



What does the mean look like?

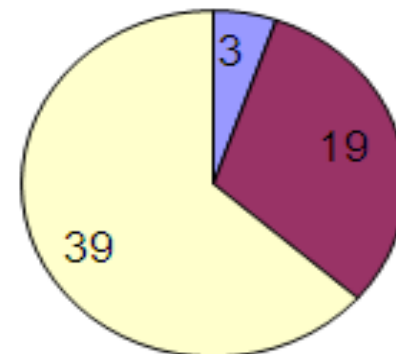
Index	Dominant Incorrect	Dominant Correct	Non-dominant	Expected Incorrect	Expected Correct	Expected Non-dominant
0.2	4	18	39	3.2	19.0	38.8

Observed Agreement



- Dominant Incorrect
- Dominant Correct
- Non-dominant

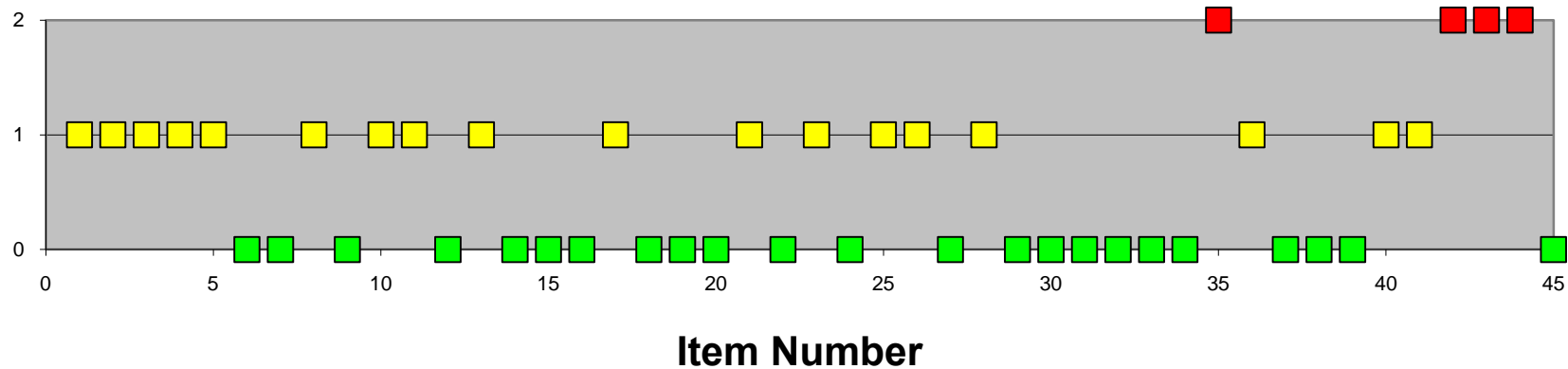
Expected Agreement



- Expected Incorrect
- Expected Correct
- Expected Non-dominant

Another Look at the Mean...

Index=0.25; Scores=303 & 309



■ No Match

■ Same Correct

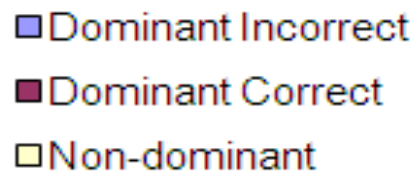
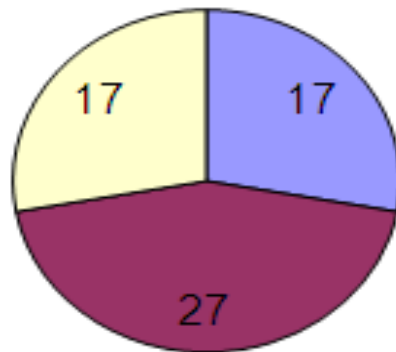
■ Same Incorrect



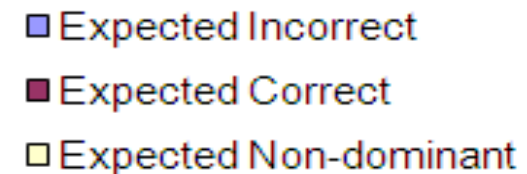
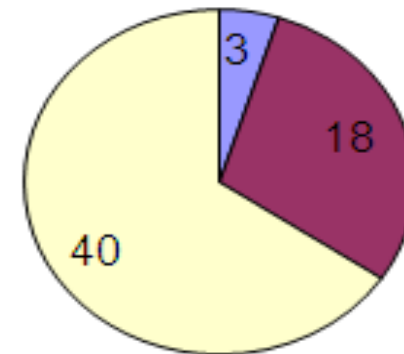
What does an index >12 look like?

Index	Dominant Incorrect	Dominant Correct	Non-dominant	Expected Incorrect	Expected Correct	Expected Non-dominant
15.6	17	27	17	3.0	18.0	40.0

Observed Agreement

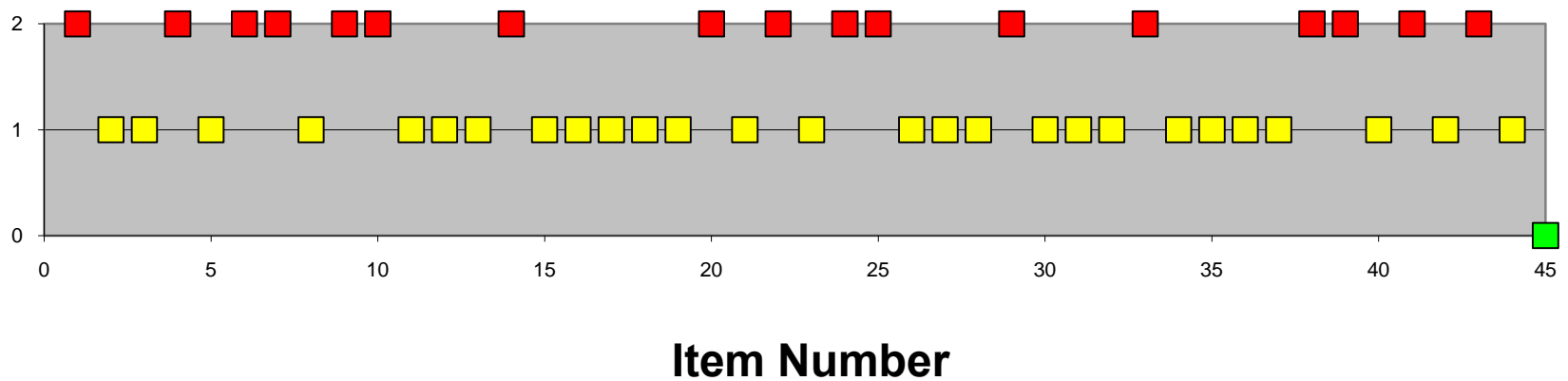


Expected Agreement



Another Look at Collusion...

Index=15.7; Scores=303 & 309

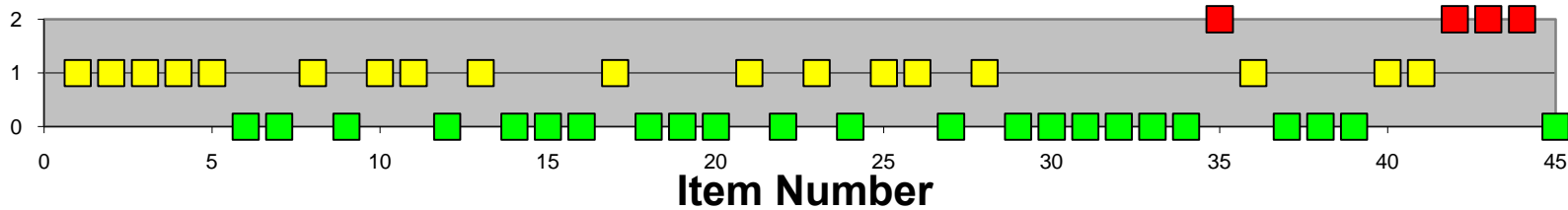


■ No Match ■ Same Correct ■ Same Incorrect



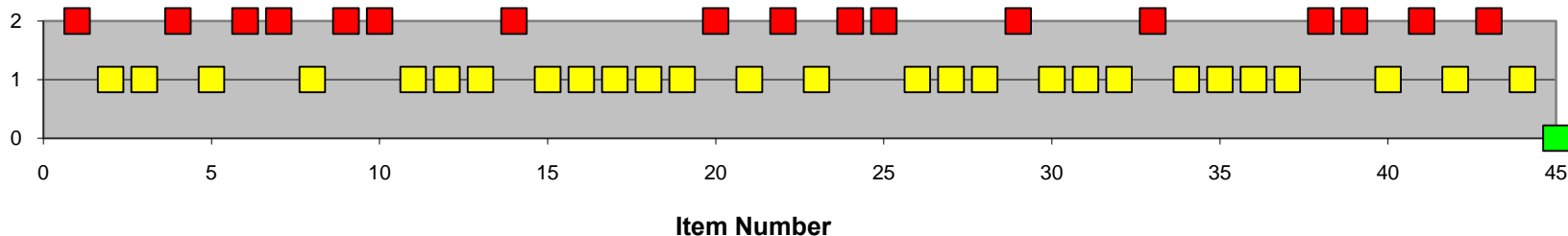
A Comparison...

Index=0.25; Scores=303 & 309



■ No Match ■ Same Correct ■ Same Incorrect

Index=15.7; Scores=303 & 309



■ No Match ■ Same Correct ■ Same Incorrect



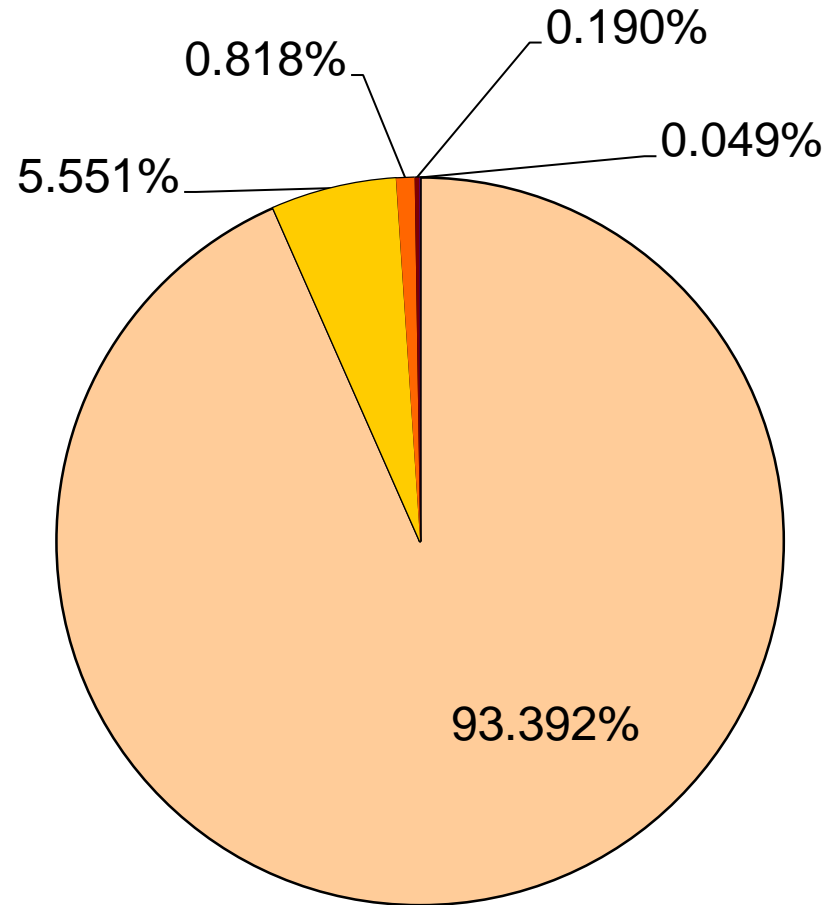
School-Level Analysis

- Similarity, Gains, AND Erasures
- Flagged schools will lead to school district review
- Unsatisfactory reviews may lead to an inspection by the Inspector General's Office

Baseline

- Presumption of normal test taking
- Baseline for entire state
 - Similarity
 - Presumption—matches occur randomly
 - Concentrations of clusters in a school warrant review
 - Erasures
 - Presumption—erasures are random
 - Concentrations of extreme numbers of erasures warrant review
 - Gains & Losses
 - Last year's mean score vs this year's mean score
 - Anomalous changes warrant review

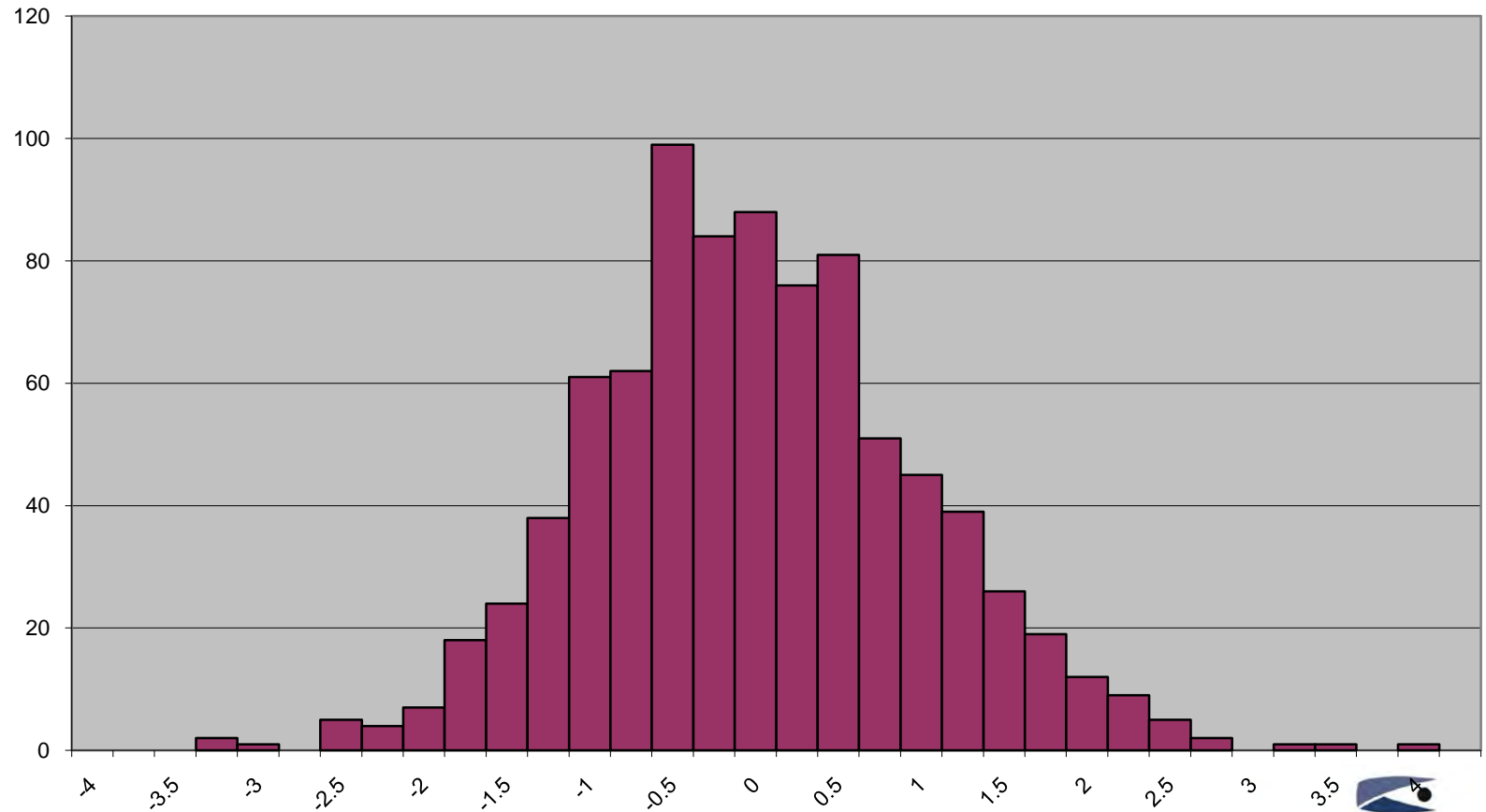
WTR Erasures - Grade 4 - Math



Low/Normal Moderate High Very High Extreme



Standardized School Gains
Grade 8 Science
Flagging Threshold = ± 4.02 ; Largest Gain = 4.08



Flagging Schools

- Output shows schools with extreme results
- Identifies the source of irregularity, ie similarity, erasures, and/or gains

Sample School-Level Analysis

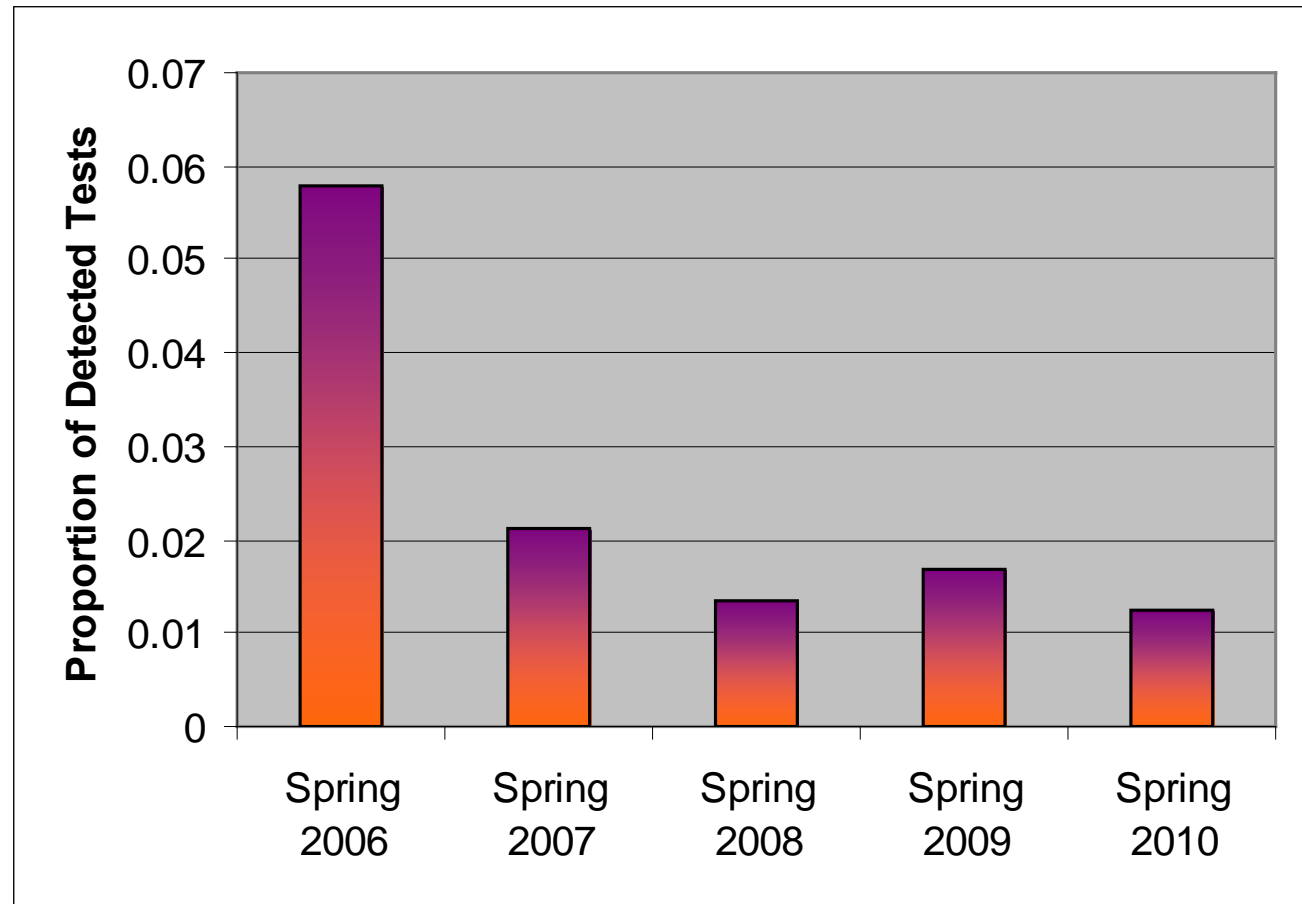
District-School	Subject	Number of Tests	Pass Rate	Pass Index	Mean Score	M4 Similarity Rate Index	Erasures Rate Index	Gains Rate Index
xxxx	M	338	0.43	0.0	286.20	34.9	0.6	9.1
yyyy	R	672	0.37	0.0	306.49	24.3	0.1	0.1
zzzzz	M	532	0.60	0.0	300.62	21.3	0.3	4.5
aaaa	M	664	0.67	0.2	310.77	17.5	0.1	3.5
bbbb	M	364	0.97	45.0	376.53	0.0	0.0	6.1
cccc	M	512	0.52	0.0	292.00	14.8	0.1	5.3
dddd	R	338	0.18	0.0	280.79	13.5	0.3	1.0
eeee	M	830	0.87	39.6	344.23	0.0	0.0	2.2
ffff	R	534	0.32	0.0	299.73	13.1	0.0	0.2
gggg	R	458	0.35	0.0	303.05	1.5	12.7	0.0
hhhh	M	197	0.28	0.0	265.98	12.5	2.4	10.3

What is coming...

- From Caveon
 - Data Forensics Reference guide
 - Descriptions of analyses in layman's terms
 - Descriptions of program details
 - Data Forensics Results training
 - Web-based training
 - Statistics “deep dive”
- From DOE
 - Appeals process detail
 - Additional information on conducting reviews of students and schools

Program Results

- Monitored behavior improves
- Invalidations deter cheating



Q & A