



Data Forensics: Assessing Test Score Validity for EOC Spring 2014

Joseph A. Orban, PhD



Overview

- ❑ High school End of Course (EOC) Spring 2014 testing for subjects: Algebra I, Algebra II, American History, Biology, English I, English II, Geometry and Government
- ❑ Spring 2013 testing for determining gains
- ❑ Analysis conducted at the school, class and student level for each district
- ❑ Major revisions since prior TAC meeting to incorporate feedback and changes.
- ❑ Analysis and reporting to be simplified for ease of understanding
- ❑ Report design and analysis QA in progress
- ❑ Reports to districts tentatively 10/30/14

Student Selection

- Not all students are in the analysis, between 3% and 5% are excluded if they:
 - did not complete the entire test and/or did not receive a valid achievement level
 - had session start to end times exceeding 2 hours and other time inconsistencies not under control of the student
- Classes and schools with fewer than 5 students testing in a subject were not evaluated on that subject.

Approach

- Evaluate the validity of test scores and flag those that may not reflect the student's ability by looking for:
 - Unusual answer choices
 - Unusual answer patterns
 - Unusual answer changes
 - Unusual answer timings
 - Unusual changes over administrations
- Students are compared to students, classes to classes and schools to schools

Aberration: Unusual Answer Choices

- Aberration – incongruence between answer correctness and question difficulty.
 - Measures the extent to which answer correctness is consistent with item difficulty.
 - Normalized $l(z)$ statistic (Levine & Drasgow, 1982; British Journal of Mathematical and Statistical Psychology)
 - Aberration greater than 4 standard deviations with examinee achieving Proficiency are flagged.
 - Classes and schools are flagged with mean aberration greater than 4 standard deviations compared to all classes and schools in the state

Rare Responses: Unusual Answer Choices

- ❑ Rare Responses – the number of least chosen answers for each MC item across the state.
- ❑ Detects random or otherwise choices independent of the question and suggesting that the test score is not valid for a measure of the student ability
 - ❑ ABCDABCDABCDABCDABCDABCD
 - ❑ CCCCCCCCCCCCCCCCCCCCCCCC
- ❑ Student counts of rare responses are flagged 5 standard deviations above the mean of all students
- ❑ Student level analysis only, not done for class or school

Similarity: Unusual Answer Patterns

- Similarity is done at the class level for classes with 5 or more students using only operational MC items.
 - Each student's answer choices are compared to every other student and the number of identical answers are tallied.
 - Total Similarity across all operational MC items
 - % Similarity of Wrong answers across all operational MC items
 - Identical answer patterns are not flagged but are captured for any subsequent investigation
- Mean similarity in a class is flagged if more than 4 standard deviations from the mean across all classes

Answer Change: Unusual Answer Changes I

- ❑ Answer changes are infrequent with online assessments, averaging less than 1 per student depending on subject and test.
- ❑ Total answer changes is a count of the number of times the student changed and answer to a MC question
- ❑ Percent W-R answer changes is the percent of total answer changes that were changed from wrong (incorrect) to right (correct)
- ❑ Students flagged with 4 standards above the mean of other students in the state.
- ❑ Classes and schools are flagged with the mean of their students at 4 standard deviations above the means of classes and schools respectively

Answer Change: Unusual Answer Changes II

- ❑ Student flagging of Total Answer Changes was the most common flag in the data across subjects for the EOC tests in Spring 2014
- ❑ Must be interpreted with caution as some teachers train students to use this as a test taking strategy to answer all questions quickly then go back and review to change to best answer
- ❑ For a class with many students flagged for Answer Changes a follow-up would be to interview students and the teacher about using answer changing as a testing strategy
- ❑ Answer changing can also be a sign that during testing information was communicated about the items to students while testing but that cannot be inferred from the forensics data.

Time:

Unusual Answer Timings

- ❑ Unusual Time of Testing Session is flagged for a student if the student started a session before 7 AM or after 4 PM
 - For both MC and PE sessions
 - Interpret with caution in cases of night classes or late afternoon test sessions
- ❑ Speed of Answering is flagged at the student level if the student has an average answer time to MC questions of 10 seconds or quicker and achieved the Proficiency standard.
- ❑ Classes and Schools are flagged if the average student MC response time in seconds is less than 4 standard deviations below the mean of all classes and schools respectively.

Gains: Unusual Changes Over Administrations

- Reports will provide test counts and parentages at each Achievement Level for the current Administration and the most comparable prior Administration
- Flags are not computed on these changes but the data is provided for review
- Counts and percentages are based on the same student sample as the flagged statistics and therefore may be different from official test counts and performance
- Look for large swings in test counts per subject but cautiously interpret
 - School might be new and growing or being closed or relocated, or changing schedules
 - Some subjects may be tested in different test windows and vary throughout the year.
 - There should be a reasonable explanation for changes in test counts and large swings in achievement levels

What Do These Flags Mean?

- ❑ The flags for students, classes and schools within a district and across subjects do not mean something nefarious has occurred.
- ❑ They do mean that certain test scores may not be a valid indicator of the student's ability. The flags do not indicate any other underlying motivation or behavior.
- ❑ Students, classes and schools with a few flags are not a reason for concern.
- ❑ Classes and schools and associated students that have numerous flags are a reason for concern but with caution.
- ❑ Only a trained investigator conducting site interviews can reveal if there are any underlying inappropriate motivations or behaviors

Policy Implications

- ❑ A clear policy distributed to all involved in testing, including a NDA, is highly recommended.
- ❑ Prevention is the highest priority. Consider the TILSA and Caveon publications as recommendations to implement
- ❑ Define a documented process for any follow-up investigations.
- ❑ Never use the word “cheating” as it has legal implications subject to libel and defamation laws. Data Forensics does not detect cheating, only an investigation can detect cheating.

Average Statewide Forensic Data by Subject Area

Subject	N	Aberration*	Total MC Answer Changes	MC W-R Changes	Rare MC Answers	MC Start Time Flag	MC Avg Answer (sec)*
ALGEBRAI	59377	-1.00694	1.00	.72	2.12	0.015%	85.44
ALGEBRAII	24038	-.86398	1.29	.99	2.58	0.046%	93.26
AMERICANHISTORY	49967	-1.10733	1.25	.78	3.23	0.108%	36.56
BIOLOGY	60757	-.72679	1.14	.76	2.04	0.026%	34.90
ENGLISHI	58672	-1.07332	1.61	1.06	2.68	0.020%	69.67
ENGLISHII	60102	-.65731	1.45	.98	1.63	0.042%	68.91
GEOMETRY	34868	-.98890	1.28	.95	2.51	0.006%	77.30
GOVERNMENT	44682	-.73182	1.28	.86	2.35	0.114%	36.46

* Lower values are more unusual

Recommendations

- ❑ Eliminate from the district report those classes and schools that only have one student flagged in a subject
- ❑ Provide a 30 minute WebEx type training to Districts, record for future use to ensure proper understanding
 - ❑ What the reports are & how to read them
 - ❑ What to do and not do with the information
- ❑ List all schools in the district report that are not flagged as “No Critical Flags”

References for Test Security and Data Forensics

- ❑ Wollack, J.A. & Fremer, J.J. (2013). *Handbook of Test Security*. Routledge.
 - ❑ <http://www.amazon.com/Handbook-Test-Security-James-Wollack/dp/0415816548/>
- ❑ Olson, J., and Fremer, J. (2013). *TILSA Test Security Guidebook: Preventing, Detecting, and Investigating Test Security Irregularities*. Washington, DC: Council of Chief State School Officers.
- ❑ Levine, M.V. & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology*, 35, 42-56.