Missouri End-of-Course Assessments

Data Forensics Procedure

Presented to the
Missouri Department of Elementary and Secondary Education
by
Questar Assessment, Inc.

August 22, 2013

5550 Upper 147th Street West
Apple Valley, MN 55124
(952) 997-2700
www.questarAI.com

Table of Contents

Overview 1
Data Forensics Procedures 2
Aberration 2
Similarity 2
Answer Change 3
Response Time 3
Outliers and Flags 4
Reporting Results from Data Forensics 4
State Policy and Communications 5
References 6

Overview

Results of high-stakes assessments are frequently used as part of high-stakes decisions for students, such as high school graduation or promotion to the next grade level. Many states also use the results for teacher evaluation, resource allocation for schools and districts, and state accountability. To protect the integrity of scores on the Missouri End-of-Course (MO EOC) assessments, the Missouri Department of Elementary and Secondary Education (DESE) has requested that Questar Assessment, Inc. (Questar) conduct data forensics analysis starting with the Summer 2013 test administration to identify possible responses that warrant further investigation of potential testing misconduct.

This document describes Questar's plan for data forensics for the MO EOC assessments. Questar plans to conduct the analyses using the following four analytic procedures[1]:

1. **Aberration:** identifies response patterns that are inconsistent with a student's ability level
2. **Similarity:** checks for similar or identical student response patterns beyond probability of occurrence
3. **Answer Change:** detects large amounts of right-to-wrong and wrong-to-right answer changes
4. **Response Time:** detects abnormally fast responses to items

Olson and Fremer (2013) and Wollack and Fremer (2013) also identify these methods for detecting testing irregularities in *TILSA Test Security Guidebook: Preventing, Detecting, and Investigating Test Security Irregularities* and the *Handbook of Test Security*.

Each of these procedures will generate statistics regarding test scores where extreme, or outlier, values (i.e., values determined to be "unusual") will be flagged, indicating an atypical occurrence but not necessarily a cause for further investigation; data forensics flags simply point to a potential problem but do not validate the problem. A group of students such as a school or a class that receives a rather large, and very unlikely, number of flags will be identified for a possible follow-up investigation in which the state will try to determine the cause with or without other external assistance. Irrespective of the cause, Questar will identify such results as unusual and not as "cheating" or "forgery" because, as stated previously, flags from data forensics indicate only possible, not definitive, testing violations.

Testing irregularities might include "sharing answers during testing, teachers helping before or during testing, [and] illicit use of stolen test questions" (i.e., similarity analyses); "changing answers by educators [and] inappropriate assistance during testing" (i.e., answer change analyses); and "inconsistent response patterns" (i.e., aberration) (Olson & Fremer, 2013, p. 28).

Data forensics can be applied to multiple levels of test data such as the student, class, proctor, teacher, school, district, or state, but it can only be applied to a level that is uniquely and consistently coded in the test data. For the Missouri EOC assessments, Questar's four data forensics procedures will use "school" and "district" as the primary groupings. While individual students could be flagged based on several statistics, data forensics procedures are not designed

[1] Under DESE's request, Questar will not conduct gain score analysis.

to detect, nor will they likely find, two students that independently colluded in the test-taking apart from all other students. Instead, they are designed to detect group-level irregularities.

This document describes the details of the four data forensics procedures Questar plans to implement for the MO EOC assessments. Considerations for flagging and reporting of flagged responses are also included.

Data Forensics Procedures

Aberration

Aberration analysis flags test scores that have inconsistent response patterns and refers to situations when students do not respond to test items according to their ability level. For example, it would be unusual if a student with a Below Basic ability level answered a group of very difficult items correctly.

Described by Hulin, Drasgow, and Parsons (1983) and termed as an Appropriateness Index to identify scores as appropriate or inappropriate, this analysis can be applied to all item response patterns regardless if the test is scored with a Rasch model, raw score, percentile, or even a pattern scoring approach. Essentially, the item responses are calibrated into Item Response Theory (IRT) parameters that are then used in a Maximum Likelihood Estimation (MLE) scoring algorithm to determine an ability, or theta, score. The Likelihood in the MLE procedure is then standardized to produce an individual aberration statistic, L(z). An unusually low L(z) value reflects a response pattern that is not consistent with the difficulties of the items. Paired with a high test score, this would be an aberration flag for a student. If a school has an unusually large number of aberration flags, it could reflect a testing problem. Other statistics such as the mean and standard deviation of L(z) can provide additional flags to gauge the seriousness of any pattern.

Similarity

Similarity refers to the extent to which students' answers are the same. This analysis procedure will require comparing every student's responses to every other student's responses within a group, such as a school. All the groups in a state have a distribution of similar response patterns, but only those groups that have a very unusual number of similar, and perhaps identical, response patterns will be flagged for concern. Multiple-choice answer options (e.g., A, B, C, and D) and the point values achieved in constructed-response items or performance events can be used in the similarity analysis. The point value similarity analysis relies on the same raw data used in the aberration analysis. Similarity is not a direct indication of a problem since it could result, for example, from a class of Advance Placement students who all learned the material well and independently provided similar answers. Therefore, any similarity flags will need to be considered along with aberration flags, as well as the other analyses described in this document. Sufficient concern that a problem exists only arises when multiple flags from multiple indicators point to the same group. Three implementations of the similarity statistics will occur as follows:

1. The first is to calculate the Euclidean distance between two vectors of item point values. This is basically the square root of the sum of the squared differences between two

students' responses. This approach applies to constructed-response as well as multiple-choice items.

2. The second implementation is to compute the percentage of items between the two students that have identical answer options for multiple-choice items and point values for constructed-response items (i.e., are the A's, B's, C's, D's, and point values the same regardless if the item was answered correctly or not?).
3. The third is to determine the ratio of the number of identical correct answers to identical incorrect answers. This statistic is sensitive to a shared test key that may include incorrect answers to some of the test items.

High similarity statistics indicate higher degree of similarities between responses. As with the other data forensics statistics, supplemental statistics can also be computed, including the mean and standard deviation of the similarity and other ratios of student pairs in the larger group such as a school. These statistics will provide flags for outlier or unusual values. Only the groups with multiple flags across all of the data forensics statistics should be considered for possible follow-up investigations.

Answer Change

Answer change analysis examines the ratios of answer changes from right-to-wrong and from wrong-to-right based on the history of the captured responses. As with the other analyses, answer change for online assessments will flag the unusual or outlier changes and will help to identify groups requiring follow-up investigations. While answer changes from wrong-to-right will be of primary interest, the total number of changes, which includes both right-to-wrong and wrong-to-right, and those that are just wrong-to-right will be calculated. Extremely large amounts of changes in comparison with baseline answer change frequencies will be flagged.

The answer change analysis is similar in concept to erasure analysis for paper-and-pencil testing and can use the same statistical analysis, which is generally a t-score beyond 4 standard deviations earning a flag. The state-level distribution of answer changes will provide the baseline against which individual students are flagged. Schools with a high percentage of flagged students will be flagged for excessive answer changes, which may, for example, be attributed to a test-taking strategy employed by many of its students. Therefore, answer change analysis is an important aspect of data forensics and must be considered along with other analyses.

Response Time

Response time refers to how long it takes students to answer each item on an online test. From when the item is presented on the monitor to when it is submitted will vary, but it should always be in a human timescale and measured in the student workstation to avoid network and other timing factors. A response that is answered in "superhuman" time (i.e., much faster than one can possibly read the item) is flagged and indicates that the student was possibly working from an answer key or some other cue. Time is also an important component when examining flags from other analyses such as aberration and similarity; for example, an unusually large number of difficult items answered quickly and correctly by a student with a low aberration statistic, L(z), could indicate the student received help while taking the test.

As with the other data forensics statistics, a time flag does not indicate a definitive testing problem, although all flags must be considered. However, if a group has an unusually large number of “super-humans” answering items in very short periods of time, concern should be raised.

In addition to individual item response times, a more reliable measure of time is from the start to the completion of the exam, as well as the overall testing time and average time per test item for an individual student. Outliers in both very fast and very slow response times would generate a flag. However, factors such as extra testing time will need to be considered and the related time flags will need to be discounted.

Outliers and Flags

“Unusual” or “outlier” statistical values may not be specified in advance since they may depend on a distribution of those statistical values from the state overall. Data forensics statistics such as aberration and similarity will require additional statistics to be calculated using classical and resampling statistics. For example, each school within a state will have an average L(z) value. Across the schools there is a standard deviation of school L(z) values. That standard deviation is programmed and computed automatically, and a school with mean L(z) values below a critical value, such as 4 standard deviations below the mean, would be flagged for aberration. This is the same with similarity. For example, a group might have a large number of student pairs that each contributes to similarity statistics. As each pair is processed, the software will keep running values used to compute the mean and standard deviation of each similarity statistic, which will then be associated with that group. After all schools are processed, each can then be compared to the standard deviation across schools to identify the outlier value.

When a group receives multiple flags, the question that arises is “How likely could this have occurred by chance?” or “What is the probability that this occurred by chance?” Several approaches can be taken to answer those questions, and it is complicated by groups that differ in the number of students tested. One approach requires resampling (e.g., using the bootstrap method), a procedure to sample a test’s entire data set many times by randomly sampling the group size from students across the state and calculating the distribution of the data forensics statistics. A computer can do such sampling as a background task. The original group statistics are then compared to those from the resampling data to determine the probability of randomly obtaining the group statistic. If it is a very small probability, such as one in a million, it can be concluded that the flag is not a random occurrence. Implementing this approach to probability is valuable in identifying outliers and documenting problems, Probability information is often used in court when cheating is challenged by an examinee and is considered to be strong evidence, as indicated in various Caveon, Inc. papers and presentations (Wollack & Fremer, 2013).

Reporting Results from Data Forensics

A critical factor regarding Questar data forensics is to report the results so they have valuable and convincing information yet do not accuse anyone of cheating, since data forensics is not direct evidence of a wrongdoing but is simply circumstantial evidence. The approach described herein is to evaluate multiple forensic statistics and, using outlier flags, highlight suspicious

situations for further examination and investigation. The data forensics report should provide information about only those groups with substantial multiple flags in a table format. The statistics underlying the flags would be shown in a graphical histogram distribution with markers indicating where the flags are located so DESE and other users of this information can see how extreme the flagged groups are compared to the norm or expected probability. Again, it is important to note that data forensics statistics are not direct evidence of wrongdoing but that multiple flags, or combinations of certain flags, should be considered for follow-up investigations.

In his Association of Test Publishers (ATP) presentation in 2013, Dr. John Fremer, a noted authority in data forensics from of Caveon, Inc., indicated that it is not in the best interest or expense of a state or contractor to investigate every oddity in data forensics, and instead the most egregious situations should be chosen for investigation. The investigation and its subsequent actions can do more to prevent future problems than chasing after every unusual occurrence in test response data. For every test window comprised of hundreds of groups or schools, there might be one, a few, or none that have multiple flags. Depending on the severity and number of flags, all or none may be investigated.

The Questar report will be streamlined with a combination of text, tables, and graphics. It will report only the most egregious forensic findings. The report will be designed by Questar, reviewed and approved by DESE, and should remain confidential.

State Policy and Communications

To assist DESE with the implementation of data forensics, Questar has prepared the following list of questions regarding state policy and communications related to test security and data forensics.

1. Will a test security policy be established?
2. What are the purposes of the test security policy?
3. Does the test security policy include preventative measures? If so, what are those measures?
4. Does the policy include clear consequences for violations? If so, will they be enforced?
5. Is there a clear process in place to conduct investigations of test security breaches?
6. How will the test security policy be communicated to districts, schools, teachers, students, and parents?
7. Will test security training be provided to staff (e.g., test administrators)?
8. Will the state establish a data forensics policy? If so, what purposes will it serve?
9. Will data forensics results trigger investigations or score holds?
10. What results will trigger such investigations or score holds?
11. What are the investigation procedures?
12. Who will be conducting investigations?
13. Who will have access to information? Under what circumstances?
14. How will the data forensics policy be communicated to stakeholders?

References

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item Response Theory: Applications to Psychological Measurement*. Homewood, IL: Dow Jones-Irwin.

Olson, J., & Fremer, J. (2013). *TILSA Test Security Guidebook: Preventing, Detecting, and Investigating Test Security Irregularities.* Washington, DC: Council of Chief State School Officers.

Wollack, J. A., & Fremer, J.J. (2013). *Handbook of Test Security. Routledge, NY.*