Spring 2016 EOC Forensic Decision Rules

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

July 29, 2016

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

Table of Contents

1. Introduction 3
2. Stable Decision Rules 3
    2.1. Aberration 4
    2.2. Similarity 4
    2.3. Answer Change 5
    2.4. Time 5
3. Unusual Identical Answer Patterns 6

1. Introduction

Each administration of the MO End of Course (EOC) assessments is subjected to a test data forensics analysis. This analysis aims to identify unusual or outlier test answer characteristics that may call into question the validity of individual test results. This analysis does not detect cheating or other misuse of the test which can only be determined by an onsite investigator. The analysis does point out those instances that might posit an investigation. However, before concluding that an investigation is needed, local knowledge of the testing administration must be considered to possibly explain the findings. For instance, tests that are administered outside of normal school hours may simply reflect an evening class. Or a high percentage of similarly high scoring students in a class may simply reflect an honors class where most students give the same correct answers to test questions.

From time to time, Questar refines the analysis to better provide meaningful information. This document describes the decision rules used to identify outlying test answer behavior which has not changed from past decision rules as well as some new rules reflecting recent refinements to the analysis.

2. Stable Decision Rules

The following analyses are identical to those in the past. The analytical logic is to calculate most of these statistics at the student level and flag those students who exceed 3 Standard Deviations (SD) beyond the mean across all students who took the same test. This is done separately for each subject matter test. For example, a student would be flagged for excessive answer changes if there answer changes were at or exceeded 3 SD of observed changes compared to all other students. The one exception to the 3 SD rule is the average time in seconds to answer the Multiple-Choice (MC) test items. A fixed 10 seconds, or quicker, was determined to be the rule for flagging a student. Ten seconds was considered to be superhuman speed as an average to answer all MC questions on a test.

The 3 SD decision rule was recommend by the MO Test Advisory Committee after reviewing results from 2, 3 and 4 SD decision rules. The 2 SD rule was considered too open, resulting in very many false positives or falsely flagged students. The 4 SD decision rule was considered too strict and may potentially miss an egregious situation. It was recommend proceeding with the 3 SD rule and interpreting the results with caution.

While individual students are flagged with the 3 SD rule, the forensics analysis is rolled up by class to reflect the aggregate behavior that occurred in the test administration group. The most egregious testing violations around the country have been at the teacher, class, school and district levels often attributable to educational staff. By aggregating over students within the class or test administration group, violations will manifest in the number of aggregated flagged students. The decision rule for flagging a class is 3 or more flagged students in

that class testing situation. This rule is applied to identify the potentially egregious outliers of testing behavior and typically results in 6 to about 20 classes in an administration window that may need additional examination. Even with this rule several of the classes identified are explained away, typically evening classes that test outside of normal school hours. Other times a basic inquiry can explain the flagged class. For example, a teacher who instructs students in a test taking strategy to quickly answer all the questions then go back through them slowly making changes will likely be flagged for excessive answer changing. This test taking strategy, if verified, would explain the forensic flag.

Forensic analysis and its statistics are just indicators of somehat unusual test taking behavior. They are not evidence of improper testing behavior but may, if sufficiently extreme, indicate that follow-up questions should be asked of the people involved in the testing and possibly a more detailed investigation to determine the cause of the unusual testing behavior.

The following forensic statistics are provided in the report of those classes identified with 3 or more flagged students.

2.1. Aberration

Aberration 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 subset of very difficult items correctly.

Described by Hulin, Drasgow, & Parsons (1983)[1] and termed as an Appropriateness Index, this analysis can be applied to all item point patterns regardless if the test is scored with a Rasch model, raw score, percentile, or even a 3PL scoring approach. Essentially, the item point 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 Log Likelihood from the MLE procedure is then normalized 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 an unusually large number of students in a class have this aberration flag, it might reflect a testing problem.

2.2. Similarity

Similarity refers to how students' answers, both correct and incorrect, are the same. This analysis procedure compares every student's response pattern with every other student's response pattern within a class. Across all the students in a state there is a distribution of response pattern similarity with classmates, but

---

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

only those groups that have a very unusual number of similar response patterns will be flagged for concern using the 3 SD rule.

The Multiple Choice answer options of A, B, C and D are used in the similarity 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 identical correct answers. So any similarity flags will need to be considered along with other student flags. Sufficient concern that a problem exists only arises when multiple flags from multiple indicators point to the same group such as high similarity paired with high number of answer changes.

For this analysis, Similarity is the number of identical answers choices for the operational MC items. This is calculated two ways i) over all MC items and ii) over incorrectly answered MC items. Similarity must be interpreted with caution as very high performing students will of necessity have similar answer patterns by correctly answering many of the same items. Additional data must be considered when there is a class with many students who have identical 100% correct MC items such as starting and ending times for the test and the number of answer changes.

2.3. Answer Change

Answer change refers to analyzing the type and frequency of answer changes. As with the other analyses, answer change for both paper-and-pencil and online tests will flag the unusual or outliers and will help to identify classes requiring follow-up investigations. The analysis will flag students with 3 SD or more of overall answer changes for a given subject test and another flag for those with 3 SD or more changes from Wrong to Right (W2R). In general the typical W2R changes is a percentage in the range of 35-50% of the overall changes whereas W2R changes on the order of 90-100% are typically flagged.

The answer change analyses are rolled up to the class level and counts towards the decision rule to flag a class if 3 or more students are flagged. Answer changes have been very useful for detecting egregious testing situations. For example the Atlanta, GA case revealed answer changes using erasure analysis which is the paper and pencil version of this analysis.

2.4. Time

Time analysis is conducted three different ways. First is the average time to answer the MC items that are flagged if the average is 10 seconds or less per item, or what is considered superhuman time. The other two analyses are flags for starting a test outside of the normal testing time for the sessions that are part of the test: the MC or selected response session and the Performance Event (PE) session. Not all tests have a PE session. These flags are applied to a student without a standard deviation rule and then rolled up to the class level. Often an evening class, perhaps in an adult education setting, is flagged for

testing outside of normal testing hours. A follow-up examination of the exact times is done to insure it does not reflect egregious abnormal time such as 2 AM.

3. Unusual Identical Answer Patterns

In addition to the Similarity analysis flag described above, Questar has been refining the individual student answer similarity metric over time and the report contains a section on "Unusual Identicals." The decision rules for this analysis have been refined for the spring 2016 EOC Forensic Analysis and described herein. An "Unusual Identical" are pairs, or more, of students in a class taking the same test at the same time. While there are many identical answer patterns among the students, even within a class such as an advanced placement class, these unusual ones standout for three reasons. First, they typically did not answer 100% correct so their wrong answers are the same wrong answers. Second, one or more of the students have a noticeable and sometimes large number of answer changes. Third, they all complete the test within seconds of each other or exactly at the same time.

One example of this would be two students in the same class testing session starting the exact same second, ending the exact same second, answering 50% of the operational MC items correctly and have the exact same incorrect answers on the remaining 50% of the MC items with no answer changes at all. That circumstance seems unusual. While it could be coincidence it could also reflect that perhaps one or both of that pair has a test result not valid for the student. Pairs of students with such unusual identical patterns within a class are not an egregious situation in and of itself. However, it may reflect a glimpse of a broader issue in the class but not picked up by other forensic statistics.

Another example is a group of 4 students with identical 100% correct answers but with varying numbers of answer changes all ending within a few seconds of each other. Indeed all four maybe in a class of 14 showing many other students with similar identical answers patterns, although not all 100% correct, answer changes and time similarities. It may simply reflect an advance placement class or perhaps some type of answer cueing taking place during testing. Therefore these "Unusual Identicals" should be interpreted with much caution as they may reflect coincidence or have a reasonable explanation.

The "Unusual Identicals" forensics results should still be considered experimental and Questar would appreciate any guidance on making the results more useful or meaningful.