Paper 257-26

Techniques for Detection of Cheating on Standardized Tests using SAS®

Sean W. Mulvenon, University of Arkansas, Fayetteville, AR
Ronna C. Turner, University of Arkansas, Fayetteville, AR
Shawn P. Thomas, University of Arkansas, Fayetteville, AR

ABSTRACT

A general perception in the United States of a declining quality in our public schools has, in part, led to legislated requirements from both Federal and State governments demanding K - 12 educational institutions demonstrate their effectiveness through accountability programs. Accountability programs have typically required public schools to demonstrate consistent improvements in student performance using standardized achievement testing. Further, the use of rewards and sanctions has contributed to a renaming of this approach as "High-Stakes" testing, because students may be denied a diploma or promotion to the next grade for a low test score or a school may lose a portion of their funding for low overall student performance. The pressure of these programs may have contributed to an increase in the incidence of cheating on standardized exams in K - 12 schools. The most common method of identifying cheating has been at the individual student level using direct observation by an observer or response string evaluations of students in the same testing location by testing companies. However, these procedures do not address the secondary cheating issue in "high stakes" testing of systemic cheating at the school or classroom levels. The purpose of this paper is to demonstrate how through the use of SAS® products SAS/STAT and SAS/IML it is possible to identify indicators of cheating or other improprieties on standardized tests, focusing on the system-wide cheating situation, but also providing examples for student level cheating. This paper should apply to the public and academic sectors and is designed for novice to advanced users of SAS® products.

INTRODUCTION

The release of reports such as *A Nation at Risk* (1983) and the Third International Math and Science Study (TIMSS) (1996), that evaluate the progress of students in the U.S. public school system, has led to a debate on the quality of education provided in the U.S. The basic premise of these reports is students in the U.S. are not performing at levels consistent with many of their international counterparts. Further, areas where U.S. students had previously excelled, such as mathematics and science, are now areas identified as weaknesses. The results from these studies are in contrast to self-evaluations of educators who cite improved pedagogical practices and educational programs as evidence of the strength of the U.S. public school system. Legislators and school boards, aware of the decline in demonstrated student performance, have begun implementing "high stakes" testing and accountability programs.
"High Stakes" testing commonly refers to the use of exit examinations to assess the achievement level of students. If a student does not pass an exit examination with the requisite score, no formal diploma is awarded. An alternative outcome of "high stakes" testing is that students who do not pass the test may not be permitted to move to the next grade. Accountability programs are more elaborate versions of "high stakes" testing directed at the school or district levels with a more formal system for evaluating the overall performance of students. Accountability programs will typically employ a form of a standardized test as part of the assessment of the performance of students within a specific school or district. The primary difference between "high stakes" testing and accountability programs is the level of responsibility, or the student versus the local school system, respectively.

In addition to local efforts, the U.S. Department of Education in 1993 passed a new requirement for Title I, a program to distribute Federal monies to low socioeconomic schools, that required the creation of accountability programs for any state requesting Title I funding. Further, the development of accountability programs required the implementation and use of a term identified as "adequate yearly progress." The U.S. Department of Education allows each State to define "adequate yearly progress," with the proviso the definition must be measurable and included in the accountability

program. Most State educational agencies use performance on a standardized achievement measure to determine if there has been an increase in performance as the basis for defining adequate yearly progress.

The increase in accountability programs, requiring the demonstration of "adequate yearly progress" through the use of "high stakes" testing may have contributed to the increased number of documented cases of cheating in K - 12 public schools (e.g., Driscoll, 2000). Most recently, teachers in Maryland, Massachusetts, and Arkansas have been identified as cheating on exams. However, detection of cheating has typically been at the student level using the direct observation of behavior or response string comparisons of students in the same testing location. The purpose of this paper is to demonstrate methods available to identify potential areas of cheating at the classroom and school levels using SAS® statistical software.

Rationale

The incidence of cheating on standardized exams, given the more recent accounts from the news media, may be evidence of a greater systemic problem, but is more likely representative of isolated cases. Regardless, the culpability associated with student progress and the willingness of officials to deviate from testing protocols may put individual educational agencies at risk for litigation. For example, if School "A" fails to test several low performing students, thus receives higher scores and a larger reward in the accountability system, who is responsible? If School "B" challenges the distribution of rewards and reports the malfeasance of School "A", local schools, districts, and state agencies, given the much publicized nature of those places where cheating has occurred, may be held culpable. Several simple statistical procedures can be completed using SAS® to identify the potential for greater problems, at which time the testing company can be alerted to complete a more formal investigation to determine any improprieties. A series of easy to complete steps to identify potential problems is provided:

Step 1: Creation of a Combined Data Set
Step 2: Detection of Eliminated Students
Step 3: Assessment of Scores
Step 4: Likelihood of Scores
Step 5: Orientation of Students & Response Patterns
Step 6: Overall Integrity of Scores

Step 1: Creation of a Combined Data Set

The first step is to combine standardized test data obtained from the testing company with local demographic and descriptive data. Several variables recommended to include are:

ADM	=	Average Daily Membership
SGPA	=	Student Grade Point Average
SST	=	Special Services Status
TI	=	Title I participation
ID	=	Student Identification Numbers
ATT	=	Student Attendance
Read98	=	Student Performance on a Prior Reading Exam (e.g., Reading Score for 1998)

These variables can be merged with data from the standardized test using some basic merge commands:

```
Title "Detection of Irregularities in Testing";

Data Test;        /* Standardized Test Data */
 input @1 (SCHOOLNO) (7.) @9 (ID) (9.) ...
plus other relevant variables ;
run;

  /* We recommended you create and use format and
data libraries */

Data Local;
 input @1 (SCHOOLNO) (7.) @9 (ID) (9.) ... plus
other relevant variables;
run;

Proc sort data= test; by schoolno id; run;
proc sort data= local; by schoolno id; run;

Data new;
 merge test local;
 by schoolno id;
run;
```

The resulting data set will include all the testing information as well as the local information for each student. This new data set provides the foundation for completing steps two through seven.

Step 2: Detection of Eliminated Students

At the school level analyses, one area of constant criticism is the selection of students actually tested. A commonly propagated assertion is some schools do not test all of their students as required. The result is scores that may not be reflective of the actual performance of the school in delivery of education. To control for this effect, it is recommended you use the Average Daily Membership (ADM) data. To receive Federal and State funds, a school must report the average number of students in daily attendance. Thus, it is in their benefit for this value to be as high as possible, and if this is done correctly, it should be within a few percentage points of the actual number of students tested. The following SAS® program is an example of how to determine if the number of students not tested exceed what would be expected due to random chance.

```
Data Total_N;
 set new;

 Title2 "Determination of Number of Students
Tested";

Proc Sort; by schoolno; run;

Proc Means n mean std; by schoolno;
 var ADM Readtest ATT;
/* Readtest are the results for a standardized reading
test */
run;
```

The "n" or sample size values reported with the Readtest variable and the mean of the ADM are the statistics of interest. If the ratio of these two values (Readtest sample size / ADM value) is < .95, some students may have been inappropriately excluded from testing. Further, check the ratio value with average daily attendance figures (ATT) for the days preceding and proceeding the administration of the test, as well as the actual test day. If the ratio of students tested to the average daily attendance figures are less than .98, you have further evidence of potential problems with students not being tested.

Step 3: Assessment of Scores

The goal of step 3 is to identify the validity of students selected to be including in any accountability scoring system. For example, Title I legislation allows for the creation of an alternative assessment for students with disabilities or who need special accommodations. These students may not be included in the regular assessment information, and thus the percentage of students tested as described in Step 2 may be higher than what would be expected. Using the theory of an approximate normal distribution of student abilities within any given school, an estimation of the percentage of students that should be included in regular testing can be computed. Depending upon the criteria designated for a given state, students below 1.5 standard deviations of the mean on a given criterion may be expected to be removed from regular testing status, for example. The percentage of students remaining would be approximately 93 percent. Because of the natural variability in student populations between schools, it is recommended that a confidence interval be computed for comparison purposes. Thus, if the lower bound of the confidence interval was 10 percent, any schools with a percentage of students included in testing that is lower than 90 percent would be flagged for further investigation. The code to be used to determine percentage of students tested would be that identified in Step 2.

Step 4: Likelihood of Scores

The likelihood of scores can also be used as an indication of improprieties on a standardized exam. This can be assessed by using performance on prior achievement tests to predict the likelihood of a given performance on the current exam. Most school districts implement annual standardized testing at the local level and bi-annual testing at the state level. Research indicates that the performance on these exams is highly correlated. Thus prediction models can be utilized to assess the probability of a given score on a standardized test. If performance on an exam is outside the range of 2.5 standard errors, for example, this score would be flagged for potential problems. This assessment could be used as a tool for assessing individual improprieties in addition to potential classroom level improprieties, by grouping the residuals by classroom.

```
Title2 "Determination of Students Performing
Outside Their Estimated Range";

Data Predict;
 set new;
```

```
Proc Reg;
 model Readtest = Read96 / p r ;
 plot p.* r. ;
run;
```

/* Read98 represents the reading score from the prior exam (1998). The plot function allows for a quick visual overview to indicate the need for individual residual assessment. */

/* Grouping by classroom can be achieved by printing out the predicted and residual scores (p and r, respectively) as a new data set and grouping them by classrm. This SAS® code would follow the Proc Reg statement.*/

```
Title3 “Residuals by Classroom”;

Proc Print;
 var id schoolno classrm p r;
run;

proc sort; by schoolno classrm; run;
proc print; by schoolno classrm;
 var id schoolno classrm p r;
run;

proc plot; by schoolno classrm;
 plot p*r / vpos=45;
run;
```

Step 5: Orientation of Students and Response Patterns

A second type of assessment of potential improprieties in testing at the student level, is an investigation into duplicate response matrices of students completing exams in similar testing locations during test administration. This process can be completed using SAS/IML and a statistical procedure analogous to computing a Kronecker product. Matrices of item responses of students completing multiple sections of an exam can be analyzed to identify linear dependencies in different students' responses. In other words, do two or more students have the same string of item responses to a section of an exam? This process can be completed using either dichotomous item scores (0 = incorrect, 1 = correct) or polytomous response patterns of options selected (e.g., 1 = a, 2 = b, 3 = c, 4 = d) with the second example providing the strongest evidence. If at least two students have identical response strings for one component of an exam, then the inverse of the matrix will not be computed, indicating a linear dependency among the variables. This would indicate a follow-up analysis is necessary to identify the pair of students with the same response strings. It is recommended that matrices be created at the classroom level in which students are most likely to have the opportunity to cheat. The number of students to be concatenated needs to equal the number of items in the response string in order to create a square matrix for computing the inverse. The following example displays individual student vectors to be concatenated into one global vector using a manual entry procedure for explanation purposes. More commonly, the data will be read directly from a prior data set using IML statements. This procedure will be used in Step 6.

```
Title2 “Response Patterns by Test Location”;

Proc IML;
Start detect1;
Student1 = {1 1 2 4}; Student2 = {1 1 2 2};
Student3 = {2 2 1 4}; Student4 = {1 1 2 4};
Newdata = Student1//Student2//Student3//Student4;
Check = inv(Newdata);
finish detect1;
run detect1;
print Newdata Check;
```

The four student matrices are representative of their distractor selections to four items on one section of an exam. Newdata is the concatenation of the student matrices to create one matrix representative of responses of students within a class. If the SAS/IML program is computed, the Check matrix will give an error of “ERROR: (execution) Matrix should be non-singular.” This indicates that at least two of the students have identical response strings. Notice that students 1 and 4 indeed have the same selections. This assessment of cheating is rather strict, requiring complete levels of duplicity on the part of the examinees, but is a procedure that might prove useful as an global check for student collaboration.

Further evidence that this is an incidence of an inappropriate testing condition could be to use the predicted information from Section 4 to identify whether or not the scores received by the individuals appear reasonable given prior performance.

Step 6: Overall Integrity of Scores

Steps 2 and 3 address the incorporation of the appropriate population of students into the testing situation. Steps 4 and 5 involve the assessment of potential testing improprieties of individual students and/or classrooms of students. In Step 6, the issue of how to evaluate the overall integrity of scores for a classroom, school, or district is demonstrated.

In contrast to Steps 2 and 3 that address whether or not the proper sampling of students has been assessed, Step 6 addresses whether or not the assessment of these students appears reasonable from a longitudinal perspective. In this section, multi-year data is used to assess whether the difference in student performance, for a selected number of test administrations, is larger than what would be expected, statistically. A procedure that could be used to address this issue is to evaluate the differences in the average performance of students for a school (e.g., classroom, grade level) across multiple test administrations using a multivariate distance computation such as a Mahalanobis distance. Using a Mahalanobis distance, a z-score is computed for each school (e.g., classroom, grade level) for each test administration year. These values are compared to a global set of means, such as the statewide average for this set of schools or a national average, to determine if there have been significant differences in performance compared to the global comparison group across the years.

For example, when compared to a state average, School 1 has mathematics z-scores of -2.0, -1.5, -1.0, and -1.1 for the last four years of testing. This indicates a steady improvement in performance compared to the state average, and the changes seem reasonable over time. In contrast, School 2 has mathematics z-scores of -1.5, -1.8, 1.0, and -1.0 for the last four years. The third year of testing has a value different from the other three years that may not seem reasonable, given their other performances. The multivariate Mahalanobis distance for School 2 will be relatively large compared to School 1. In this situation, information from Steps 2 and 3 would be of interest to determine if the percentages of students testing were different for that year. Additionally, information from Steps 4 could provide additional insight into whether or not the students tested that year were performing at levels substantially different than what would be expected. The following SAS IML code is provided for computing the multivariate Mahalanobis distance for a multi-year dataset.

```
Title2 “Testing the Integrity of Multi-year Data”;

Proc IML;
reset print;
start detect2;
use new;   /* Read data from input statement */

read all var {Read96 Read97 Read98} into X1;
/* Read96 = 1996 Reading Scores, etc. */

n = nrow(X1);   sum = X1(| + , |);
mean1 = sum / n;
means = repeat(mean1, n, 1);
xpx = X1` * x1 - sum` * sum / n;
CovMat = xpx / (n - 1);
StdDev = diag(sqrt(CovMat));
/* IML statements for creating a Variance-
Covariance matrix and a vector of Standard
Deviations */

Mahal = (X1 - means) * inv(CovMat) * (X1 -
means)`;
CompMah = vecdiag(Mahal);
/* Computation and selection of a vector of
multivariate Mahalanobis distances */

zscores = (X1 - means) * inv(StdDev);
/* Provides the zscores for each year comparing how
a school performed in comparison to the group as a
whole */

Final = zscores||CompMah;
finish detect2;
run detect2;
/* Provides matrix of columns for annual school
performance, with the final column representing the
multivariate distance of performances each year as
compared to a centroid representing the composites
of means */
```

For example, if students within a school have average scores that are 1 standard deviation below the mean two of three years, but are 1 standard deviation above the mean on the third year, their Mahalanobis distance will be relatively large, signaling a pattern of scores not indicative of the group as a whole. Further assessments of the

possible reason for the difference in the annual scores would be investigated.

Educational Impact

The incidence of cheating on standardized exams may not be a systemic problem, but litigation and other potential problems associated with accountability programs and high-stakes testing has contributed to an environment where schools, district, and state educational agencies may need to protect themselves against allegations of these types. In addition to student level cheating, procedures are provided in this text for assessing testing improprieties at a classroom and school level. Some simple procedures, that can be completed using the SAS® software can help to provide this protection and help to obviate the need for more legal solutions.

REFERENCES

Driscoll, D. P. (2000). A statement on allegations of cheating on the MCAS. Internet document (http://www.doe.mass.edu/news/feb00/0225pr.html).

National Center for Education Statistics. (1996). Third international math and science study. Washington D.C.: U.S. Department of Education.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for education reform. Washington D.C.: U.S. Government Printing Office.

CONTACT INFORMATION

Sean W. Mulvenon, Ph.D.
254 Graduate Education Building
College of Education and Health Professions
University of Arkansas
Fayetteville, AR 72701
Phone: (501) 575-8727
Fax: (501) 575-2492
Email: seanm@uark.edu
Web: http://orme.uark.edu