Data forensics: A compare and contrast analysis of multiple methods

Christie Plackner

Vince Primoli

Data Recognition Corporation

Paper presented at the 2012 Conference on Statistical Detection of Potential Test Fraud in Lawrence, KS.

All correspondence and citation permission should be directed to Christie Plackner, Data Recognition Corporation - 13490 Bass Lake Road, Maple Grove, MN 55311 (cplackner@datarecognitioncorp.com).

While test cheating behavior has been acknowledged as an issue in high stakes testing for some time, recent news coverage has brought it to the forefront of the minds of assessment professionals – and everyone else – across the nation. Recently, *The Atlanta Journal-Constitution* (March 2012) published a report identifying school districts across the nation who had unlikely patterns in their test scores, such as a large increase one year followed by a large decrease the next year. Until this time, the articles garnering most of the attention have focused on the use of erasure analysis to provide evidence of test taking misconduct. Major newspapers have reported on the suspicious test score gains identified by an erasure analysis in Washington D. C. as well as in Georgia. Additionally, an Internet search results in additional stories regarding erasure analysis results in other states – New Jersey, Pennsylvania, New York, for example. Despite the prevalence of stories regarding erasure analysis, the majority of states are not conducting it. A survey of states by the USA Today (September 12, 2011) found that only twenty states and Washington D.C. conducted an erasure analysis on all paper-and-pencil tests during the 2010-2011 school year. An explanation given as to why one was not performed included the belief that their state does not have a cheating problem. Another explanation is the lack of money. Some states have had a history of conducting an erasure analysis for many years and then halted the practice due to budget constraints. Then there are the states that have gathered erasure analysis data for years but have done nothing meaningful with the results, for whatever reason.

If there are not methods in place to attempt to identify them, it is difficult to know if any testing irregularities exist. It also seems like an unwise assumption that cheating isn't a problem in some areas of the country yet so prevalent in others. A survey administered to Arizona teachers directly asked teachers about cheating practices. The survey found that more than 50% of the respondents knew colleagues who cheated. More than 50% reported having engaged in these practices themselves (Amrein-Beardsley, Berliner & Rideau, 2010). The authors caution generalizing the results, but they do stress that they deliver a strong message regarding cheating under high-stakes testing.

Much media attention has been given to adults erasing student test answers as a means to increase scores. It is naïve, however, to assume that the only way a teacher or administrator will influence a student's test answer is to erase a wrong answer and replace it with a correct one. It has been found, for example, that student answers may also be influenced by having answers written on the board, making copies of the test ahead of time and teaching to the questions, having lower achieving students sit next to higher achieving ones, or, by giving instructions to not fill in the answer sheet until confirming the answer with the teacher. In the high-stakes testing environment it is best to ensure that the results from

the administered tests are based on effective instruction and true student achievement and not the examples previously given. It is also encouraged. U.S. Secretary of Education Arne Duncan issued a policy letter (dated June 24, 2011) that urges states to "make assessment security a high priority" and "ensure that assessment development contracts include support for activities related to test security, including forensic analysis." Additionally, it is recommended by the Association of Test Publishers and the Council of Chief State School Officers (2010) that rules and procedures be adopted that respond to instances of test administration irregularities.

To identify cheating behaviors that go beyond erasures, multiple data forensic methods are seem prudent. If erasure analyses are considered a budget constraint, then this realization may be concerning. This exploratory study uses principal component analysis to investigate the interrelationships among multiple data forensic methods. In addition to erasure analysis the data forensic methods included in the analysis are:

- Scale score gains and losses using cohort and non-cohort groups
- Performance level gains and losses using cohort and non-cohort groups
- Measurement model misfit (e.g., Rasch residuals)
- Across and within subject predictions using cohort and non-cohort groups, and
- Pattern analysis (Jacob and Levitt, 2003).

The first goal in the exploratory analysis is to examine to what extent each method accounts for variability in the data set. The second goal is to identify a more efficient – and perhaps more cost effective – set of reduced methods.

Data Forensic Methods

The data forensic methods in this study have been used in a high-stake assessment's data forensic approach for grades 3-8 and 10 and across mathematics and reading. Each was selected to look at the data from a different perspective in order to (hopefully) identify the various situations in which student test scores may have been influenced by someone else, be it from erasing and replacing responses to writing answers on the board. The majority of the forensic methods applied a score to the school if the school's results were statistically improbable, with probable defined as the state average of the event or occurrence (i.e., the baseline).

The statistical probabilities were transformed into a score with a range of 0 to 50. School results that were closer to the baseline have a score closer to zero, while schools with the most extreme results were closer to 50.

The score is computed using the following equation:

1. $$Score = \left| 1.086 \ln\left(\frac{p}{q}\right) \right|,$$

where *p* is the probability of the occurrence of the behavior, and *q* is 1-*p*.

The natural log of *p/q* was taken to make the scale symmetric around small and large probabilities. The constant 1.086 was used to make the probability of 0.0001, or .0002 for two-sided, equal to a score of 10. Thus, for ease of use and interpretation, any behavior that results in a score of 10 or greater was considered different from the baseline

Erasure Analysis

Schools were identified that had an erasure rate higher than the state average than was expected from random events. This analysis was done for each year and grade within a subject. The baseline for the erasure analysis is the average wrong-to-right (WR) erasures made by all students in the state, the state average. Each school's WR average is calculated and then compared to the state average.

The following statistics are calculated:

2. $$Mean_{State} = \frac{\sum X_i}{N}$$, where X_i is the number of WR erasures for student *i*, and *N* is the total number of students in the state.

3. $$SD_{State} = \sqrt{\frac{\sum_{i=1}^{k} (X_i - Mean_{State})^2}{N-1}}$$

4. $$Mean_{School} = \frac{\sum X_i}{n}$$, where X_i is the number of WR erasures for student *i* in the school, and *n* is the total number of students in the school.

5. $$SD_{School} = \sqrt{\frac{\sum (X_i - Mean_{School})^2}{n-1}}$$

6. $$t = \frac{Mean_{School} - Mean_{State}}{\frac{SD_{School}}{\sqrt{n}}}$$, where degrees of freedom (df) is n-1

The probability of the *t* statistic, *p*, is converted into the score using Equation 1.

Scale Score Changes

Schools were identified that had scale score changes that were either statistically higher or lower than the previous year. This analysis was done using non-cohort groups and cohort groups. For the non-cohort approach, scores were compared from a specific year's grade to the prior year's grade (e.g., this year's fourth graders to last year's fourth graders). Alternatively, the cohort analysis matched students from the previous year to the current year (e.g., this year's fourth graders to a matched set of the same students who were in third grade last year). The cohort analysis was made possible by matching students within a school by their student identification number. If a student could not be matched across years, then they were not used in the analysis. The statistics used were the same regardless of the approach.

To determine whether a school's change in scale score is statistically different than the baseline's, the means of two independent samples are compared by conducting a *t*-test. The following statistics are calculated:

7. $$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$, where

$\bar{X}_1$ is the mean score of each school of previous year,
$\bar{X}_2$ is the mean score of each school of current year,
μ_1 is the mean score of state level of previous year, and
μ_2 is the mean score of state level of current year.

8. $$s_p^2 = \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2}{n_1 + n_2 - 2}$$, where

s_1^2 is the variance of each school of previous year,
s_2^2 is the variance of each school of current year,

n_1 is the number of students of each school of previous year, and

n_2 is the number of students of each school of current year.

The probability of the *t* statistic, *p*, where *df* is ($n_1 + n_2 - 2$), is converted into a score using equation 1.

Performance Level Changes

Schools were identified that had a percentage of students in the top two achievement categories that was considered to be statistically higher or lower than the previous year. As with the scale score comparisons, two methods were applied to this examination of change in percent in achievement levels. The first method compares the percentage in the same grade across years. The cohort approach focuses on the same group of students from one year to the next.

To determine the extent of change in the percentage of students in the top two achievement categories, considered the percentage proficient, the log odds ratio is used to compare the percentage of students in the current year to the percentage in the previous year.

For a probability π of proficient, the odds are defined to be

9. $$\Omega = \frac{\pi}{1-\pi}.$$

The ratio of the odds Ω_1 (for current year) and Ω_2 (for previous year) is

10. $$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/1-\pi_1}{\pi_2/1-\pi_2}.$$

The odds ratio is also called the cross-product ratio, since it equals the ratio of the products $\pi_{11}\pi_{22}$ and $\pi_{12}\pi_{21}$ of cell probabilities from diagonally opposite cells. The odds ratio can equal any nonnegative number. The conditions $\Omega_1 = \Omega_2$ and $\theta = 1$ correspond to independence of the current and previous years. When $1 < \theta < \infty$, students in the current school year are more likely to have proficient scores than are students in the previous school year. For example, when $\theta = 4$, the odds of proficiency in current school year are four times the odds in previous school year.

Values of θ farther from 1.0 in a given direction represent stronger association. Different direction of values represents the same association, but in opposite direction, when one is the inverse of the other. For example, θ=0.25, the odds of proficiency in the current school year are 0.25 times the odds in the previous school year, or equivalently, the odds of proficiency in the previous school year are 1/0.25=4.0 times the odds in the current school year.

For inference it is convenient to use $\log\theta$ (Agresti, 2002). Independence corresponds to $\log\theta = 0$. The log odds ratio is symmetric about this value. The following statistics are used to calculate the log odds ratio:

11. $$\ln(\hat{\theta}) = \ln\left[\frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{12}+0.5)(n_{21}+0.5)}\right], \text{ where}$$

n_{11} is the number of students proficient in previous year,

n_{12} is the number of students NOT proficient in current year,

n_{21} is the number of students proficient in previous year, and

n_{22} is the number of students NOT proficient in current year.

12. $$\hat{\sigma} = \left(\frac{1}{n_{11}+0.5} + \frac{1}{n_{12}+0.5} + \frac{1}{n_{21}+0.5} + \frac{1}{n_{22}+0.5}\right)^{1/2}$$

13. $$z = \frac{\ln(\hat{\theta})}{\hat{\sigma}}$$

The probability of the *z* statistic, *p*, which is converted into the score using equation 1.

Note that the 0.5 in equations 10 and 11 are used to deal with problems that may occur when using log odds ratios with small sample size. This method is preferred when the cell counts are very small or any zero cell counts occur. Gart and Zweifel (1967) showed that these amended estimators behave well.

Measurement Model Misfit

Schools were identified that had performed better or worse than expected. Although student-level residuals were not reported, they were calculated and then summed to the school level. The defining characteristic of Rasch measurement is *specific objectivity*. For the individual, this means that the probability of success on any item depends only on the person's ability and the item's difficulty. After extracting the information associated with the sufficient statistics for estimating ability and item difficulty, the remainders can be used to control the model by assessing the degree to which specific objectivity is obtained. Inspecting the residuals can provide valuable diagnostic information (e.g., it can help detect students who score unexpectedly higher on some items given their overall ability and item difficulties). The fundamental unit of data is the response of one person to one item. Because both the

person ability and the item difficulty have been estimated from larger data sets, they can be effectively removed from the observation leaving the typical person-item residual. The residuals are summed across operational items for the school. To adjust for unequal school sized, the mean outfit z for a school is regressed on the square root of the school's enrollment count. The following statistics are computed:

14. $y_{vi} = x_{vi} - p_{vi}$, where $p_{vi} = \frac{e^{b_v - d_i}}{1 + e^{b_v - d_i}}$

15. $z_{vi} = \sqrt{\frac{1 - p_{vi}}{p_{vi}}} = \sqrt{\frac{e^d}{e^b}}$, if item *i* is answered correctly, where *b* is the student ability, and *d* is item difficulty.

16. $z_{vi} = -\sqrt{\frac{p_{vi}}{1 - p_{vi}}} = -\sqrt{\frac{e^b}{e^d}}$, if item *i* is answered incorrectly, where *b* is the student ability, and *d* is item difficulty.

17. $MS = \frac{\sum_{i=1}^{N} z_{vi}^2}{N}$, where *N* is the number of points.

18. $MS_{school} = \frac{1}{n_{school}} \sum_{v}^{n_{school}} \sum_{i=1}^{N} z_{vi}^2 -$, where *N* is the number of points.

19. $S = \frac{\left[\sum \frac{1}{w_{ni}} - 4N\right]^{\frac{1}{2}}}{N}$

20. $S_{school} = \frac{1}{n_{school}} \sum_{v}^{n_{school}} \left[\sum \frac{1}{w_{ni}} - 4N\right]^{\frac{1}{2}}$

21. $w = p_{vi}(1 - p_{vi})$

22. $outfit_{zstd} = (MS^{\frac{1}{3}} - 1)(\frac{3}{S}) + (\frac{S}{3})$

The probability of the $outfit_{zstd}$ statistic based on the cumulative standard normal distribution, *p*, is converted into a score using equation 1.

Subject Regression

To determine if the difference between the observed and predicted scores are statistically different, a regression analysis was conducted to identify schools that have large differences between observed and

predicted scores. The equations below show how the regression is done to predict the reading scores from the mathematics scores. Similar analyses were conducted to predict the current year's mathematics scores from the previous year's score. For the within subject regression a matched cohort set of students was used across years.

23. $\hat{Y}_j = \beta_0 + \beta_1 X_j$, where

$\hat{Y}_j$ is the estimated reading scale score of student *j* in the state level,

X_j is the actual mathematics scale score of student *j* in the state level,

β_0 is the intercept for the state level, and

β_1 is the slope for the state level.

24. $\hat{\bar{y}}_k = \beta_0 + \beta_1 \bar{x}_k$, where

$\hat{\bar{y}}_k$ is the predicted mean reading scale score of school *k,* and

$\bar{x}_k$ is the actual mean mathematics scale score of school *k.*

25. $t = \frac{\bar{y}_k - \hat{\bar{y}}_k}{s}$, where

$\hat{\bar{y}}_k$ is the predicted mean reading scale score of school *k*

$\bar{y}_k$ is the actual mean reading scale score of school *k,* and

26. $s_{\bar{y}} = \sqrt{\frac{\sum (y - y_{State})^2}{n-1}}$, where

y_{State} is the state mean reading score, and *n* is the number of students in the school.

The probability of the *t* statistic, *p*, where *df* is ($n-1$), which is converted into a score using equation 1. Negative values of *t* are converted to a score of 0, since we only want to flag a school when the observed score is much greater than the predicted score.

Modified Jacob and Levitt

This was the only method that did not result in a school receiving a score. This method included a combination of two indicators: (1) unexpected test score fluctuations across years using a cohort of students, and (2) unexpected patterns in student answers. Schools were indentified that have both large score fluctuations across years and unexpected patterns in student answers. This analysis was a modified application of that described by Jacob and Levitt (2003).

In the original study indices are calculated at the classroom level; however, since classroom data was not available in the current data set, results of this study are reported at the school-grade-subject level, referred to as cohort. Because of this difference, larger variations existed in sample sizes than did in Jacob and Levitt's data. To accommodate this, modifications were made to the analysis including the use of quantile regression to rank-order data otherwise heavily dependent on sample size.

Quantile regression was used on each of the four measures used by Jacob and Levitt as a final step to account for statistical differences associated with sample size. Also, changes were made to the analyses to accommodate the need for expedient results. Jacob and Levitt's work used previous, current, and future student scores in the multinomial and score fluctuation calculations. Future scores refer to the score achieved the year after the year of interest under investigation. The following year's score was important because it was expected to drop (or show minimal gain) since the theorized unsanctioned behavior would have stopped. This analysis uses only one-year's previous and current scores for these calculations. The omission of future student scores was a considerable data loss but it would have been unfortunate to have to wait more than a year past the testing administration for results.

Index 1

The first indicator ranked each school's average test score gains relative to other schools' gains for a particular grade and subject. The mathematical form of this index is Index 1 = rank gain, where rank gain[1] is the percentile rank for average test score gains for all students in each cohort from years *t-1* to *t* as ordered by the probability of obtaining a given change or more extreme in deviations from the mean across years assuming the distribution of the test score change follows the t distribution. Cohorts that yield values in the top 95th percentile of this index are identified as having unusual test score fluctuations.

[1] The grade equivalence (GE) was used by Jacob and Levitt for gain scores over years.

Index 2

The second index ranked schools regarding unexpected patterns in student answers. The student answer pattern analyses were examined in four ways. Schools' rankings on the four measures were combined to provide an overall index of unexpected patterns in student answers. The analyses identified the:

1. most unlikely block of identical answers,
2. highly correlated answers across the test,
3. degree of variance in the correlation of responses across items, and
4. cases in which students miss easy items while answering difficult items correctly.

Measure 1 identifies the most unlikely block of identical answers given by students on consecutive items using a multinomial logit model. The likelihood of each student choosing each possible answer on every item is calculated based on the student's current year's (t) test responses and past test scores (year t-1). All combinations of students and consecutive items are compared to find the block of identical answers that were least likely to have arisen by chance.

First, a multinomial logit model is used to calculate every student's likelihood on each item:

27. $$P(Y_{isc} = k) = \frac{e^{\beta_k x_s}}{\sum_{j=1}^{J} e^{\beta_j x_s}}, \qquad \text{k = 1, ..., J}$$

where *s* is the student, *c* is the cohort, *k* is the selected answer option, J is the total number of options, and *x* is the vector of past test scores.

Second, the likelihood of a student's answer for item *i* is found by selecting the appropriate value from expression 1:

28. $$P_{isck} = \frac{e^{\beta_k x_s}}{\sum_{j=1}^{J} e^{\beta_j x_s}}$$, where *k* is the response actually chosen by student *s* on item *I*.

Third, identify strings of items, *m* to *n*, for which the cohort gave identical responses; then the likelihood of this string for student s is the product of the item likelihoods from expression 2:

29. $$P_{sc}^{mn} = \prod_{i=m}^{n} P_{isck}\ .$$

Fourth, the product across all students in the cohort who had identical responses in the string is:

30. $\tilde{P}_{sc}^{mn} = \prod_{s\in\omega} P_{sc}^{mn}$, where ω is the group of students who have identical responses to items *m*. to *n.* The calculations are repeated for all strings of five consecutive items.

Finally, the minimum value of this measure for each cohort is recorded as measure 1:

31. $Measure\ 1 = \ min_s\left(\tilde{P}_{sc}^{mn}\right)$

The smallest values are associated with more improbable answer string within a cohort.

Measure 2 examines the degree of correlation in student responses across the test, particularly for unexpected answers. It was based on the assumption that teachers who cheated will have students with highly correlated answers. Measure 2 is the average of the item residual values. Higher values indicate cohorts with highly correlated answers.

32. $$e_{ijsc} = \begin{matrix} 0 - P_{isck} & if\ j \neq k \\ 1 - P_{isck} & if\ j = k \end{matrix}$$

Then residuals for each option are summed across students within the cohort:

33. $$e_{jic} = \sum_{s} e_{ijsc}$$.

Then, the sum of squared residuals are divided by squared number of students to normalize for cohort size.

34. $$v_{ic} = \frac{\sum_j e_{jic}^2}{n^2}$$

35. $$Measure\ 2 = \bar{v} = \frac{\sum_j v_{ic}}{ni}$$

Measure 3 calculates the variance in the degree of correlation across test items. If a teacher cheated by changing or providing answers for multiple students on selected questions, the within-class correlation on those particular questions will be extremely high, while the within-class correlation on other questions is likely to be normal. Thus, a large degree of variance in the correlation of responses across items would occur. The variance is calculated as follows:

36. $$Measure\ 3 = \sigma_v = \frac{\sum_i (v_{ic} - \overline{v}_c)^2}{ni},$$

where *ni* is the number of items on the exam.

Measure 4 compares the answers of students within a cohort to the answers from other students with same total scores in the sample. It detects students who missed easy items while answering difficult items correctly.

The mathematical form is as follows.

37. $$Measure\ 4 = \sum (z_{sc} - \bar{z}_A)$$

38. $$\text{where, } z_{sc} = \frac{\sum (q_{isc} - \bar{q}_A)^2}{n},$$

where A is the total number correct score; q_{ic} equals one if the student answers item *i* correctly and zero otherwise; and $\overline{q}_A$ is the proportion of students with total score A_s who answered each item correctly. The deviations between this value for each student and the average value for all students with same total score are then summed for all students within a cohort. High values of this index indicate the answer from a large number of students in the cohort deviated from students with same total scores in other cohorts.

The cohorts are ranked on each of the four measures. For this, quantile regression is used to remove the effect of sample size inherent in the indices used by Jacob and Levitt (2003). While *least squares* regression minimizes the sum of squared deviations from the regression line and passes through the mean, quantile regression on the median minimizes the sum of absolute values of the deviations from the line which is the median (Koenker, 2005). The result is that exactly 50 percent of the data points will be above line and 50 percent below. It is also true that 50 percent of the points are expected to be above the line for any value of the independent variable. In other words, the quantile regression line is the median of the dependent variable conditional on the independent variable.

In the quantile analysis, the JL measures were the dependent variables and school enrollment was the independent variable.

Quantile regression can be generalized to any percentile, typically denoted as τ between 0 and 99. The preceding discussion used the median ($\tau = 0.5$.) By iterating on τ using the *R* package *quantreg* (R Development Core Team, 2003; Koenker, 2011), it is possible to determine the percentile rank for any value of the measure conditional on school enrollment.

Index 1 + Index 2

It was possible for a school to experience a large increase in tests scores due to, for example, the introduction of a new curriculum or after-school program. It was also possible for unexpected answer patterns to appear without inappropriate behavior having occurred. For these reasons, a school had to be in the 95th percentile on both indices to be flagged. Having to be within the 95th percentile on both indices, in this context, was a way to limit the schools being identified due to Type I error; the schools are protected from being falsely identified. The percentile ranks for each cohort on each measure are then combined to form the Index 2 as follows:

$$39. \quad Index\ 2 = Measure1\_rank^2 + Measure2\_rank^2 + Measure3\_rank^2 + Measure4\_rank^2.$$

Principal Component Analysis

Ten data forensic methods have been applied to the data set. Eight of these use a score to represent the probability of unusual testing behaviors and one, Modified Jacob and Levitt, ranks schools based on two indexes. These indexes are treated separately in the analysis. These analyses use the scores and percentile rankings resulting for a data forensic analysis of a grade 4 math test. The data forensic methods used were:

1. Erasure Analysis (mER)
2. Scale score changes using non-cohort groups (mSS)
3. Scale score changes using cohort groups (mmSC)
4. Performance level changes using non-cohort groups (mPL)
5. Performance level changes using cohort groups (mPLC)
6. Model misfit using Rasch Residuals (mRR)
7. Across subject regression using reading scores to predict mathematic scores (mRG)
8. Within subject regression using a cohort's previous year score to predict current score (mCR)
9. Index 1 of the Modified Jacob and Levitt evaluating score changes (mJL1)
10. Index 2 of the Modified Jacob and Levitt evaluating answer sheet patterns (mJL2).

Principal component analysis is used since it is a technique in which many somewhat correlated variables can be reduced to smaller set of variables while maintaining most of the information from the larger set (Dunteman, 1989; Joliffe, 2002). Through the analysis a greater understanding of how each method contributes to the overall variance will be gained. Additionally, a reduced set of methods accounting for the majority of the variance will be identified.

The analysis was conducted on the correlation matrix using the principal component extraction method in the SPSS 17 factor analysis function. The means and standard deviations for 8 of the data forensic scores are summarized in Table 1. The mean and standard deviation for the two Modified and Jacob Levitt (MJL) indexes are also provided, but note that these statistics are from percentiles and not scores as the other methods. As a reminder the schools' scores ranged from 0 to 50. A total of 1692 schools' grade 4 math scores were used in this analysis.

Table 1. Data Forensic Methods Descriptive Statistics

	Mean	Std. Deviation	Analysis *N*
mSS	2.494	2.3771	1692
mPL	1.747	1.4836	1692
mRG	1.353	2.5161	1692
mRR	.998	1.8506	1692
mER	1.580	3.7375	1692
mSC	3.429	3.9141	1692
mPLC	1.486	1.4927	1692
mJLI	.50278	.288319	1692
mJL2	.49774	.287703	1692
mCR	4.4495	5.14603	1692

Table 2 provides the correlations between the original variables. The largest degree of correlation among the methods is .986 for the scale score cohort method (mSC) and the cohort regression method (mCR). This isn't surprising as both methods rely on a cohort's performance from one year to the next. The smallest correlations are between the two MJL indexes, -0.005. The high correlation may pose a

problem in the analysis as mSC and mCR may be measuring the same thing. Likewise, the correlations below 0.1 may also be problematic as they may form their own individual principal components.

Table 2. Data Forensic Correlation Matrix

	mSS	mPL	mRG	mRR	mER	mSC	mPLC	mJLI	mJL2	mCR
mSS	1.000									
mPL	.503	1.000								
mRG	.029	-.042	1.000							
mRR	-.033	-.020	-.033	1.000						
mER	.084	.034	.070	.128	1.000					
mSC	.121	.048	.322	.050	.087	1.000				
mPLC	.085	.184	-.010	.056	.094	.465	1.000			
mJLI	-.028	-.078	.518	-.054	.092	-.045	-.122	1.000		
mJL2	.097	.006	.133	.235	.212	.400	.217	-.005	1.000	
mCR	.135	.058	.285	.056	.091	.986	.474	-.107	.405	1.000

The Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) tests of the strength of the partial correlations of the variables. Values close to one are desirable; however, as seen in Table 3, a value of 0.6 is adequate for the analysis. Also reported in Table 3 is Bartlett's Test of Sphericity, another indicator of strength of the relationship among variables. If this is significant it can be concluded that the strength of the relationship among variable is strong and the analysis may proceed.

Table 3. KMO and Bartlett's Test Statistics

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.599
Bartlett's Test of Sphericity	Approx. Chi-Square	8806.642
	df	45
	Sig.	.000

Examination of Variance

Ten components are summarized in Table 4 as ten data forensic methods are used in the analysis. The first component explains almost a third of the variance in the ten forensic methods. The first four components account for approximately 70 percent of the total variation. It is clear that as the component number increases the contribution to the total variance decreases. Components 5 through 10 each contribute little more to the total variance. Figure 1 is a plot of the components and their eigenvalues. This scree plot illustrates the decreasing contributions as the number of components increases.

Table 4. Principal Component Statistics

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.758	27.582	27.582	2.758	27.582	27.582
2	1.625	16.247	43.829	1.625	16.247	43.829
3	1.407	14.074	57.902	1.407	14.074	57.902
4	1.204	12.045	69.947	1.204	12.045	69.947
5	.847	8.471	78.418			
6	.724	7.238	85.656			
7	.604	6.040	91.696			
8	.459	4.587	96.283			
9	.360	3.601	99.884			
10	.012	.116	100.000			

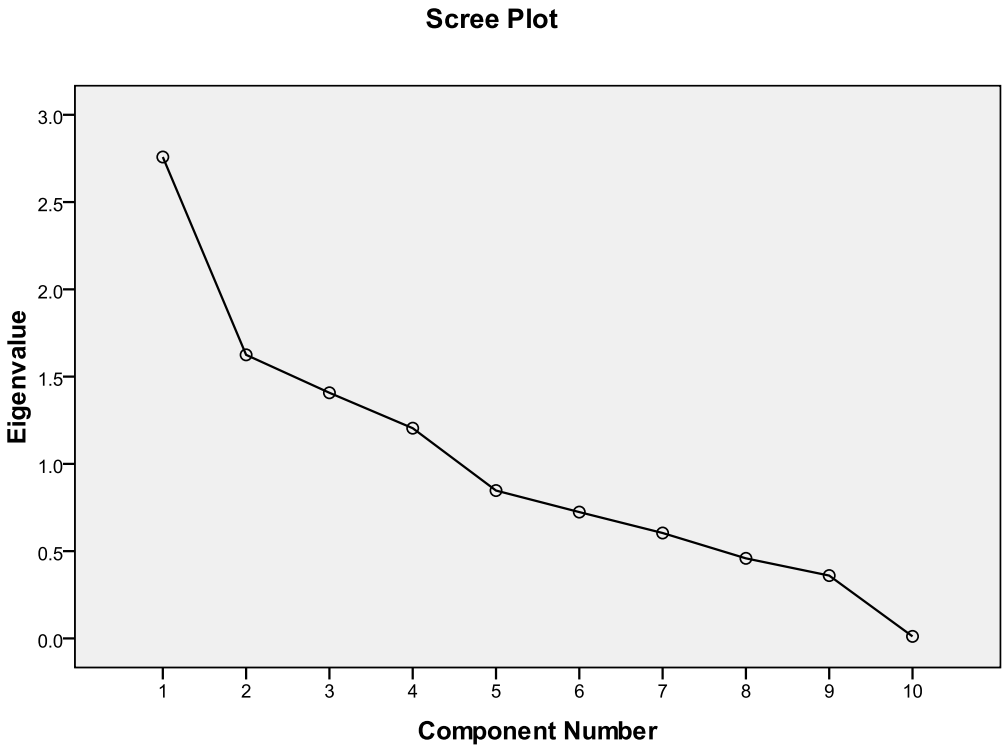

Figure 1. Scree Plot of 10 Components

The first principal component is sensitive to cohort analyses. As can be seen in Table 5 cohort regression (mCR) and cohort scale score change (mSC) correlate quite high with the first component. The second component, accounting for 16 percent of the variation, correlates highest with the score change index from the MJL and the subject regression method. In Table 5 the forensic methods which correlate the highest with each of the components are shaded gray.

Table 5. Principal Component Loading Matrix for all Components

	Component									
	1	2	3	4	5	6	7	8	9	10
mCR	.924	.026	-.147	-.230	.013	-.086	-.193	-.017	.140	.077
mSC	.923	.079	-.125	-.236	.019	-.069	-.183	-.015	.162	-.076
mPLC	.617	-.266	-.126	-.120	-.206	.576	.291	.176	-.168	.000
mJL2	.592	.065	-.147	.405	.117	-.390	.535	-.083	-.040	.000
mJLI	-.023	.732	.476	.085	.031	.234	.205	.142	.334	.005
mRG	.359	.706	.385	-.089	.147	.031	-.123	-.143	-.397	.000
mSS	.267	-.438	.679	.095	.122	-.236	-.086	.428	-.074	.000
mPL	.194	-.555	.636	.046	.108	.180	.050	-.438	.097	7.526E-5
mRR	.158	-.033	-.282	.688	.533	.288	-.227	.048	.013	.000
mER	.239	.088	.115	.646	-.676	-.016	-.209	-.048	-.007	.000

To assist in determining if a general pattern exists between the components, a simpler presentation is offered. In Table 6 a method's loading on a component is replaced with a + or – if the loading's absolute value is greater than half the maximum coefficient in that component. If the loading coefficient's absolute value is between a quarter and a half of the maximum loading in that component it is represented as a (+) or (-) (Jolliffee, 2002). This recoding was completed for only the first four components as they represent most of the variance.

Of the ten methods, seven at least partially loaded onto the first component. Two of the methods that did not load on the first component, non-cohort performance level changes and MJL index 1, do correlate with the second. The final method, Rasch residuals, did load, although not highly, on the third component. Each method seems to account for some of the variability in identifying misconduct. However, those loading highest on component one may be contributing the most.

Table 6. Simplified Principal Component Loading Matrix

	1	2	3	4
mCR	+			(-)
mSC	+			(-)
mPLC	+	(-)		
mJL2	+			+
mJLI		+	.+	
mRG	(+)	+	+	
mSS	(+)	_	+	
mPL		_	+	
mRR			(-)	+
mER	(+)			+

Data Forensic Selection

Determine How Many Principal Components

There are multiple ways to reduce the variable set to a more efficient grouping. The first step is to determine how many principal components to retain. One method is to make a selection based on the cumulative percentage of total variation. Recommendations include selecting components until the percentage of total variation contributes 80% or 90% (Jolliffe, 2002). Recommendations of 70% can be found as well (Dunteman, 1989). Jolliffe (2002) indicates that the amount of preferred variance will become smaller as the number of observations becomes greater. Additionally, the preferred level may also change due to practicality. In this study seven of the ten components would be retained to achieve at least 90% of the total variation. This is not practical with the objective of reducing methods. Since the count of observations, 1692, is high and 70% is a sensible amount, this is considered a plausible method to determine how many components to keep.

Another method to determine how many components to maintain is to review the scree plot in Figure 1. The scree plot recommends selecting two components as a steep slope is evident from the first to the second components. Beyond that, a clear slope does not exist.

A third approach to resolve how many principal components are necessary are to review the eigenvalues. A general rule of thumb, although not necessarily the best, is to retain a component if its eigenvalue is over 1 (Osborne, 2008). An eigenvalue of one may also occur due to chance.

Table 7 summarizes the various methods of selection and the number of components it would retain. Of the methods proposed retaining two or four components seems practical. It could be argued that using a more conservative eigenvalue as a cut off would result in the retention of two components as well.

Table 7. Number of Components Retained by Method

Method	Number of Retained Components
90% Cumulative Variance	7
70% of Cumulative Variance	4
Scree Plot	2
Eigenvalue	4

Determine Which Data Forensic Methods to Maintain

Two methods will be used to select the forensic methods. Both methods will associate one variable with each of the selected components. In the first method, the selected variable would be the one loading highest on a component. (Dunteman, 1989; Jolliffe, 2002). If a variable is the highest loading on two components and is already selected to represent a component, then the second highest loading variable will be chosen. A variable is selected for the first component, then the second, and so on until one variable represents each component. The second method is a discarded principal components method (Dunteman, 1989) in which variables correlating highest with the smallest component are removed. This continues until the same amount of variables remain as components selected. If the variable has already be discarded due to is loading on another component then remove the next highest loading variable. The rationale for this approach is that variables with high loadings on small components reflect redundancy among the variables with high weights.

Table 8 summarizes the retained variables. The column labeled “Positive Selection” refers to the first method in which variable selection begins with the largest component. Both selection techniques resulted in the same methods being chosen. When retaining four components the methods are

1. Cohort Regression,
2. Modified Jacob and Levitt Index 1,

3. Non-cohort scale score change, and
4. Model misfit using Rasch residual.

The two methods selected when retaining two components are cohort regression and Modified Jacob and Levitt index 1. Of the four methods, three of them are specifically detecting unexpected score changes. JL1 is specifically identifying those with large increases, although mCR and mSS will also identify decreases in scores.

Table 8. Data Forensic Method Selection Summary

Selection Method	Positive Selection		Discarded Principal Components	
Number of Components	4	2	4	2
mCR	X	X	X	X
mSC				
mPLC				
mJL2				
mJLI	X	X	X	X
mRG				
mSS	X		X	
mPL				
mRR	X		X	
mER				

Discussion and Conclusion

This paper used principal component analysis to gain understanding as to what extent the data forensic methods were accounting for variation in detecting test taking irregularities. All methods seems to be accounting for some as evident by seven of the ten methods loading on component one. Those that did not load onto component one loaded on to either the second or the third components. Of these methods, cohort regression, cohort scale score change, and cohort performance level change seem to be accounting for the most variation.

Two methods were applied in an attempt to reduce the amount of methods that could be applied in detecting irregularities while maintaining a level of accounted variance. Both approaches resulted in the same methods being selected.

Future studies have much to consider to improve and to build upon what was done. There are additional principal components selection methodologies which may conclude with different results. It is also prudent to examine the variables more carefully before including them in the analysis. Including all the variables in the analysis was purposeful in the study, but for future studies there are some considerations. For example, the high correlation between cohort regression and cohort scale cohort may indicate that one of them needs to be removed from the analysis. Determining which one to remove would be a subjective decision that should be thoughtfully made. Another adjustment to the included methods would be how to handle the Modified Jacob and Levitt indexes. They are designed to work together and splitting them in the study may have had implications.

This exploratory study serves as a first step into understanding how applying multiple forensic methods is beneficial as well as to which methods account for the most variablity.

References

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). New York: Wiley.

Amrein-Beatdsley, A., Berliner, D. C., Rideau, S., (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education Policy Analysis Archives, North America,* 18, jun. 2010. Available at: <http://epaa.asu.edu/ojs/article/view/714/841>. Date accessed: 11 Mar. 2011.

Bello, M., & Toppo, G. (2011, September 9). Few states examine test erasures. *USA Today.* Retrieved from http://www.usatoday.com/news/education/story/2011-09-12/states-analyze-test-erasures/50376902/1

Council of Chief State School Officers and the Association of Test Publishers (2010). *Operational Best Practices for Statewide Large-Scale Assessment Programs.* Council of Chief State School Officers and the Association of Test Publishers, Washington, DC.

Duncan, A. (2011). Key Policy Letters from the Education Secretary or Deputy Secretary. June 24, 2011 http://www2.ed.gov/policy/elsec/guid/secletter/110624.html

Dunteman, G. H. (1989). *Principal Components Analysis*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-069. Beverly Hills: Sage Publications.

Gart, J. J., & Zweifel, J. R. (1967). On the bias of various estimators of the logit and its variance with applications to quantal bioassay. *Biometrika*, *54*, 181-187.

Hays, W. L. (1994). *Statistic* (5th ed.). Fort Worth, TX: Harcourt College Publishers.

Jacob, B. & Levitt, S (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics, 118*(3), 843-877.

Jolliffe, I. T. (2002). *Principal Component Anlaysis* (2nd ed). New York: Springer.

Koenker, R. (2005). Quantile Regression. Cambridge U. Press

Koenker, R. (2011). *Quantile regression in R: A vignette*, version dated July 24, 2011. Retrieved November 1, 2011, http://www.econ.uiuc.edu/~roger/research/rq/vig.pdf .

Mead, R. (1976). *Analysis of fit of data to the Rasch model with residuals*. Unpublished dissertation. University of Chicago.

Mead, R. (1980). Using the Rasch model to identify person-based measurement disturbances. *Proceedings of the 1979 Computer Adaptive Testing Conference*. Minneapolis, MN: University of Minnesota.

No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).

Osborne, J. W. (2008). *Best Practices in Quantitative Methods.* Los Angeles: Sage Publications

R Development Core Team (2008). *R: A language and environment for statistical computing.* http://www.R-project.org.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Smith, R. (1982). Detecting measurement disturbance with the Rasch model. Unpublished dissertation. University of Chicago.

Smith, R. (1986). Person fit in Rasch model. *Educational and Psychological Measurement*, *46*, 359-372.

Smith, R. (2004). Fit analysis in latent trait measurement models. In E. V. Smith & R. M. Smith, (Eds.), *Introduction to Rasch measurement*.(pp 73-92) Maple Grove, MN: JAM Press.

Vogell, H., Perry, J., Judd, A, & Pell, M.B. (2012, March 25). Cheating our children: Suspicious school test scores across the nation. *Atlanta Journal-Constitution.* Retrieved from http://www.ajc.com/news/cheating-our-children-suspicious-1397022.html

Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: Mesa Press.