

Thesis Equivalency Project -- Educator Cheating: Classification, Explanation, and Detection

Brad Thiessen

Tests are only useful if they provide accurate information. To ensure the scores obtained from a test will be accurate, test developers spend considerable time, effort, and money to carefully develop test specifications, test items, and administration and scoring procedures. This effort is wasted if examinees cheat on the test. Cheating invalidates the decisions made on the basis of test scores.

In general, cheating is defined as:

Cheating: to deprive of something valuable by the use of deceit or fraudulent means to violate rules dishonestly (Merriam-Webster online dictionary)

Gregory Cizek, author of the book *Cheating on Tests: How To Do It, Detect It, and Prevent It*, provides a definition of cheating specific to educational testing. Cizek defines cheating as:

Cheating: an attempt, by deception or fraudulent means, to represent oneself as possessing knowledge that one does not actually possess (Cizek, 1999, p.3)

Student Cheating

Research in cheating on educational tests (the prevalence of cheating, reasons for cheating, and techniques to prevent cheating) has typically focused on the examinees – the students. This research usually takes one of the five following forms:

1. Confession Surveys: Students are asked to confess to cheating they have done in the past. Obviously, it may be difficult to get accurate information from these surveys (even if randomized response techniques are used to encourage students to tell the truth).
2. Accusation Surveys: Students are asked about the behaviors of their peers in regards to cheating.
3. Observations: Student behaviors are observed during test-taking. Researchers focus on behaviors that may indicate a student is cheating on a test.
4. Experiments: Researchers place students into situations where they are allowed to cheat if they desire. In one study, students were allowed to grade their own tests after the tests were previously graded secretly (Spiller & Crown, 1995). In another experiment, students were administered a multiple-choice test with no real correct answers (Houser, 1982). Students whose answers closely matched the developed “answer key” were thought to have cheated.
5. Statistical Detection: Analyzing student answer strings on multiple-choice tests to find improbable answer patterns. Several statistical indices have been developed to analyze answer patterns for potential cheaters.

Statistical detection methods calculate the probability that a student would answer items as he or she actually did. The most popular indices include:

Angoff's β -Index (1974): Angoff developed 8 indices, labeled A-H, to detect potential cheaters. His β -Index examines the number of "errors in common" between two examinees. If students choose the same incorrect answer to an item, they are said to have an error in common. The formula for calculating the β -Index is:
$$\beta = \frac{(\text{common errors}) - (\text{avg. common errors for all examinees})}{(\text{standard deviation of common errors})}$$
.

Outlying values of β are used to identify potential cheaters.

ESA (Belleza, 1989): Using a binomial distribution, this index calculates the probability that a pair of examinees would have identical errors on test items (errors in common). If N represents the number of items answered incorrectly by a pair of examinees, and k represents the number of those items with the same incorrect response,

$$\text{ESA is calculated as: } ESA = \frac{N!}{(k!)(N-k)!} p^k (1-p)^{N-k}$$

MESA: MESA stands for the *Modified Error Similarity Analysis*. While the details of this index remain a mystery, this index is calculated in the *Integrity* software package sold by Assessment Systems Corporation.

g_2 (Frary, 1977): This index treats one examinee as "the copier" and another examinee as "the source." Using the copier's score and the popularity of each item option, this index calculates the probability that the copier would select the source's answer across all items. In using information from all items, not just items answered incorrectly, this index has been found to provide better results than the ESA index.

ACT Pair 1 & Pair 2 (Hanson & Brennan, 1987): The Pair 1 method uses two indices for each pair of examinees: (1) the number of matching errors and (2) the length of the longest string of matching responses. The Pair 2 method uses: (1) the number of errors in the longest string of matching responses and (2) the number of test items and identical responses across all examinees.

K-Index (Holland, 1976); u -Index (Wollack, 1997); Z-Index (van der Linden, 2002); S_1 and S_2 indices (Sotaridona, 2001): These indices are slight modifications of previous indices. For example, the S_1 index is similar to the K-index except it uses a poisson distribution instead of the binomial distribution.

ω -Index (Wollack, 1997): This index uses Item Response Theory to detect students with large numbers of unusual responses. Other aberrant response indices using IRT include IMS, OMS, and L_z (Karabatsos, 2003).

NBME (Cizek, 1999): The National Board of Medical Examiners compares the response patterns of examinees sitting adjacent to one another to the patterns of nonadjacent examinees.

Through the use of these five techniques, researchers have been able to estimate just how often students cheat on tests. In his book on student cheating, Gregory Cizek summarizes 17 studies into the prevalence of student cheating. These studies, mostly surveys, consistently reveal that more than 50% of students admit to some form of cheating. Furthermore, more than two-thirds of college students admit to cheating (Cizek, 1999).

This widespread cheating can take many forms. In fact, Cizek documents 62 unique cheating methods (not counting cheating using modern technology) students have used to artificially inflate their scores. To gain an understanding of the variety of cheating methods, Cizek created a taxonomy of student cheating to classify cheating methods into one of the three categories:

- (1) Giving, taking, receiving: Typically involves students sharing answers with one another
- (2) Using forbidden materials: Students obtaining the answer key before a test is administered
- (3) Taking advantage of the process: Students not following test administration procedures.

Because cheating invalidates test scores, it is important that test users are able to detect and prevent student cheating on tests. To aid in this goal, the Assessment Systems Corporation developed two software applications to detect cheating: *Scrutiny!* and *Integrity*. The importance of cheating prevention is also demonstrated by the existence of *Caveon*, a company specializing in detecting and preventing cheating on tests.

Educator Cheating – Prevalence & Types

While attention has been focused on the cheating behaviors of students, perhaps more attention should be focused on teachers and principals who cheat for their students on tests. Recently, many educators have been discovered to cheat on tests. In researching newspaper stories on cheating on tests, Nichols & Berliner found 26 published stories on student cheating... they found 83 stories of educator cheating (Nichols & Berliner, 2004).

In developing this paper, I searched newspaper articles for stories of educators cheating on tests. In the first six months of 2005, I found an additional 15 stories. These articles, summarized in Appendix A, give an idea as to the nature and prevalence of educator cheating on educational tests.

A 1990 survey by Gay found that 31.5% of educators either observed cheating or engaged in cheating themselves. A 1992 survey by *Educational Measurement* reported that 44% of educators said that colleagues cheated on tests for their students and 55% were aware that fellow teachers cheated on tests. The newspaper articles and research results indicate that educator cheating is a real problem that must be addressed.

So if educators are cheating, how are they doing it? After examining the newspaper stories and the surveys by Gay (1990), Shepard & Dougherty (1991), and Nichols & Berliner (2004), I developed a taxonomy of educator cheating. The methods used by educators when cheating on tests fall into one of the four following categories:

1. Manipulating answersheets
2. Manipulating the test administration process
3. Manipulating the score reporting process
4. Manipulating the teaching process or teaching philosophy.

Table 1 on the next page displays the taxonomy along with specific methods educators have used to cheat (based on the newspaper articles in Appendix A). The second and third columns report the percentage of educators who admit to using each specific cheating method. The final column reports the number of newspaper stories Nichols & Berliner found that describe educators using each cheating method.

Looking at Table 1, it is apparent that most cheating by educators involves manipulating the test administration, score reporting, or teaching processes. Fewer educators admit to taking part in the more blatant form of cheating – manipulating student answersheets.

While the rest of this paper will look into methods used to detect and prevent educator cheating, it is worthwhile to consider reasons why educators cheat in the first place. One reason why educators choose to cheat may be the increasing pressure they face from high-stakes testing and the *No Child Left Behind* legislation. This was expressed as Campbell's Law:

Campbell's Law: The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor (Campbell, 1975).

Table 1: Taxonomy of Educator Cheating			
	Shepard & Dougherty (1991)	Gay (1990)	Nichols & Berliner (2004)
Manipulating Answersheets			
* Altering a student's answersheet	6.1%	1.5%	16 stories
* Giving students answers	8.3%		8 stories
* Pointing out mismarked items		10%	11 stories
Manipulating Administration Process			
* Giving students hints on test items	22.7%	10%	7 stories
* Rephrasing test items for students	18.0%		
* Providing students extra time	19.6%	15%	3 stories
* Reading items that are supposed to be read by students	14.1%		
* Excluding students from testing	7.4% - 13.3%		41 stories
* Answering questions about test content	11.7%		
* Instructing students to fill-in a specific answer for unanswered items			
* Providing accommodations or inappropriate special ed. placement			1 story
Manipulating Reporting Process			
* Removing student test scores from the official records			1 story
* Providing students with false ID numbers so their scores won't count			1 story
* Misrepresenting data			8 stories
* Changing the criteria for proficiency			21 stories
* Conflicting accountability ratings			15 stories
Manipulating Teaching Process or Philosophy			
* Teaching students specific test items or teaching during test	30.4%	5%	4 stories
* Practicing with actual test items	11.3%		17 stories
* Teaching to the test			15 stories
* Narrowing the curriculum			13 stories

Detecting Educator Cheating

While several statistical indices have been developed to detect student cheating, there is a need to develop indices to detect educator cheating. Whereas student cheating can invalidate one test score, an educator who cheats can invalidate an entire set of test scores at once.

The only attempt to detect educator cheating on large-scale standardized tests was completed by Jacob & Levitt in 2003. Jacob & Levitt set out to create a statistical index to identify educators, specifically teachers, who manipulate answersheets (which is the most blatant and least-used form of educator cheating). After developing their index, they analyzed Iowa Tests of Basic Skills scores from Chicago public schools. They conclude their research by stating, "Empirically, we detect cheating in approximately 4 to 5 percent of the classes in our sample" (Jacob & Levitt, 2003, p. 846).

The method used by Jacob & Levitt to detect potential cheating teachers is actually a combination of two indicators: (1) unexpected test score fluctuations and (2) unexpected patterns in student answers. Rather than focus on individual students, these indices are calculated at the classroom level.

Index #1: Unusual Test Score Fluctuations

On a vertically scaled test like the ITBS, we expect student test scores to increase by a relatively constant rate each year. The test score gains depend on a variety of factors (student ability, curriculum, teacher quality, etc.), but most students will experience growth in achievement each year. An *unexpected test score fluctuation* would occur when many students in a classroom experience large score gains one year followed by small gains (or even negative growth) the next year. To calculate this unexpected test score fluctuation index, Jacob & Levitt do the following:

- If we are interested in determining if a teacher manipulated answersheets during the “Year t” test administration, we must gather test scores for year t, the previous year (t-1) and the next year (t+1).
- Find the average test score gains (in grade equivalent units) for all students in each classroom (the gain from year t-1 to year t along with the gain from year t to year t+1).
- Find the percentile rank of each classroom’s average test score gains relative to all other classrooms in that same subject, grade, and year. The percentile rank of growth from year t-1 to year t will be called “rank gain_t” while the rank of growth from year t to year t+1 will be called “rank gain_{t+1}”
- The index is calculated as: $\text{Index \#1} = (\text{rank gain}_t)^2 + (1 - \text{rank gain}_{t+1})^2$. As the formula shows, classrooms with relatively large score gains followed by low gains the next year will yield large values of this index. The percentile ranks are squared in computing the index to give relatively more weight to big score gains and big score declines.
- Teachers whose classrooms yield values in the top 95th percentile of this index are identified as having unusual test score fluctuations.

Table 2 shows an example of this index. The test score gains for three classrooms are displayed. Classroom 1 represents an average classroom, growing from 2.9 to 4.1 to 5.2 grade equivalent units over the course of two years. Classroom 2 represents a classroom with an excellent teacher (during year t). This excellent teacher was able to drastically increase test scores from year t-1 to year t. The teacher of Classroom 3 also was able to get drastic increases in test scores. The

difference between Classrooms 2 and 3 can be seen by the changes in test scores once the students moved on to the next grade (with a different teacher). Whereas the students from Classroom 2 were able to build upon their score gains (as we would expect from students who were taught by an excellent teacher), the students in Classroom 3 experienced test score declines. This decline in test scores may indicate that the gains experienced the previous year did not represent genuine gains in achievement. Thus Classroom 3 earns a high score on Index #1 and the teacher of that classroom is identified as a potential cheater.

Table 2: Example of Index #1

	Avg. Grade Equivalent For Classroom			Change in Grade Equivalent Units		Index #1 (%ile rank)
	t-1	t	t+1	From t-1 to t (%ile rank)	From t to t+1 (%ile rank)	
Classroom 1	2.9	4.1	5.2	1.2 (59)	1.1 (56)	0.54 (20)
Classroom 2	3.4	5.5	6.8	2.1 (92)	1.3 (77)	0.8993 (70)
Classroom 3	3.1	5.2	4.7	2.1 (92)	-0.5 (1)	1.8265 (99*)
Avg. Classroom	3.3	4.4	5.3	1.1	0.9	

Jacob & Levitt experienced with other measures of unusual test score fluctuations (such as regressing test score gains on previous gains and student demographics). They concluded that these other, more complicated measures yielded information similar to their simple index.

Index #2: Unexpected Patterns in Student Answers

Unusual fluctuations in test scores do not prove that a teacher manipulated student answersheets. In fact, since the 95th percentile is used as a cut-off, Index #1 will always identify 5% of the classrooms as having unusual fluctuations. To determine which of these classrooms cheated (and which just experienced improbable fluctuations), Jacob & Levitt developed a second index to identify unexpected patterns in student answers.

The logic is this: The quickest way for a teacher to cheat is to alter the same block of consecutive items for students in the class (or instruct students in the classroom to change their

answers to the same set of items). Thus, if a classroom experiences unusual test score fluctuations and the students in the classroom have unusual answer patterns (identical answers to the same block of items or unexpected correct answers to difficult items), then we have more reason to believe the teacher cheated.

To identify unexpected answer patterns, the researchers combine four measures of suspicious answer strings to calculate Index #2. These four measures will be briefly discussed:

- The first measure focuses on identifying 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. This likelihood is based on the student's past test scores, future test scores, and demographic (gender, race, etc). All combinations of students and consecutive items are searched to find the block of identical answers that were least likely to have arisen by chance (controlling for classroom size).

Given student s in classroom c with answer j on item i , the model is:
$$P(Y_{isc} = j) = \frac{e^{\beta_j X_s}}{\sum_{j=1}^J e^{\beta_j X_s}}$$

Where X represents the vector of past test scores, future test scores, and demographics.

The likelihood of a student's answer string for item m to item n is calculated as:
$$P_{sc}^{mn} = \prod_{i=m}^n P_{isc}$$

This likelihood is multiplied across students in the class with identical responses in the string:

$$\tilde{P}_{sc}^{mn} = \prod_{\text{students with identical string}} P_{sc}^{mn}$$

If each student in the classroom has unique responses from item m to item n , then there will be a distinct value of this index for each student in the class. If all students in the classroom have identical responses across these items, then there will only be one value of this index (and the value will be extremely small).

The calculations are repeated for all strings from a length of 3 items to a length of 7 items.

Notice that the values of yielded by these calculations will be smaller as: (1) the number of students with identical responses increase, (2) the length of the string of identical responses increase. Thus, smaller values are associated with more improbable answer strings within a classroom.

The minimum value of this measure for each classroom is recorded as measure #1:

$$\text{Measure \#1} = \min_s (\tilde{P}_{sc}^{mn})$$

- The second measure calculates the degree of correlation in student responses across the test, especially for unexpected answers. The logic is that teachers who cheat will have students with highly correlated answers. To calculate this measure, the residuals for each item choice are calculated:

$$e_{ijsc} = \begin{cases} 0 - P(Y_{isc}), & \text{for unchosen options} \\ 1 - P(Y_{isc}), & \text{for the chosen answer} \end{cases}$$

Then, the residuals for each option are summed across students within the classroom: $e_{jic} = \sum_s e_{ijsc}$

The option residuals for the classroom are then summed for each item. At the same time, the residuals are (1) squared to accentuate outliers and (2) divided by the number of students in the class to normalize for class size (n):

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

Measure #2 is simply the average of these item residual values. Higher values indicate classrooms with highly correlated answers.

$$\text{Measure \#2} = \bar{v} = \frac{\sum_i v_{ic}}{n}$$

- The third measure calculates the variance in the degree of correlation across test items. With Measure #2, we might expect high correlations among student answers if a teacher emphasizes certain topics during the school year. If a teacher cheats by changing answers for multiple students on selected questions, the within-class correlation on those particular questions will be extremely high, while the degree of within-class correlation on other questions is likely to be typical. Thus, a teacher who change answers on selected items will have a classroom with a large degree of variance in the correlation of responses across items.

This measure is calculated as the variance of item residuals from Measure #2:

$$\text{Measure \#3} = \sigma_v = \frac{\sum_i (v_{ic} - \bar{v}_c)^2}{ni}$$

- The fourth measure compares the answers of students within a classroom to the answers from other equally-able students in the sample. This measure can then detect students who miss easy items while answering difficult items correctly. Students whose answers follow this pattern may have had their answers influenced by a cheating teacher.

To calculate this measure, students are grouped by their total number correct scores on the test. Let A_s represent a specific total correct score. Let $q_{ic} = 1$ if a particular student answers item i

correctly and zero otherwise. Then determine the proportion of students with total score A_s who answered each item correctly (call this quantity \bar{q}_A).

The deviations between a student's item score and the expected item score (based on equally-abled students) are squared and summed across items: $Z_{sc} = \sum (q_{isc} - \bar{q}_A)^2$

This deviation between this Z-value for each student and the average Z-value for all equally-abled students is then summed for all students within a classroom. Thus, high values of this index indicate the answers from a large number of students in the classroom deviated from equally-abled students in other classrooms.

$$\text{Measure \#4} = \sum (Z_{sc} - \bar{Z}_A)$$

After completing the calculations, the classrooms are ranked on each of the four measures. The percentile ranks for each classroom on each measure are then combined to form the second index:

$$\text{Index \#2} = (\text{Measure 1 rank})^2 + (\text{Measure 2 rank})^2 + (\text{Measure 3 rank})^2 + (\text{Measure 4 rank})^2$$

Classrooms falling above the 95th percentile on this index are identified as having unusual answer patterns.

Putting It Together: Detecting Potential Cheating Classrooms

Jacob & Levitt argued that taken individually, the above two indices do not detect teachers who manipulate answersheets. After all, there are always going to be (innocent) classrooms with unexpected score fluctuations and there are going to be (innocent) classrooms with improbable answer patterns. The key is to identify classrooms that yield high values on both indices.

In non-cheating classrooms, there is no reason to believe that the two indices would have a strong correlation. If a teacher manipulates student answersheets, we would expect a strong correlation between the two indices. Therefore, educators whose classrooms appear above the 95th percentile on both indices are identified as potential cheaters.

Conclusions From Jacob & Levitt (2003)

Jacob & Levitt used their methodology to analyze scores from the Iowa Tests of Basic Skills from the Chicago public school system. Based on their analysis, they reached the following conclusions:

- 1.1% to 2.1% of classrooms in their sample cheated on any particular ITBS subtest
- 3.4% to 5.6% of classrooms cheated on *at least one* ITBS subtest
- Teachers found to cheat on one subtest were 10 times more likely to cheat on other tests
- A significant correlation was found in cheating within schools – a school with one cheating teacher was likely to have more cheating teachers
- A teacher found to cheat one time was 9 times more likely to cheat again in the future.
- The following factors were found to be significant predictors of cheating in classrooms:
 - Classrooms that performed poorly the previous year were more likely to cheat
 - Classrooms with higher poverty rates and more African-American students were more likely to cheat
 - Classrooms with a higher proportion of students whose scores count (non-LEP, non-IEP) were more likely to cheat
 - Teachers who administer the exam to their own students were 50% more likely to cheat
 - Younger teachers were more likely to cheat
 - Teachers not from prestigious undergraduate institutions were more likely to cheat
 - A high-stakes testing environment was found to increase the probability that a teacher would cheat
- When the researchers randomly assigned students to “classrooms,” the researchers found no cheating classrooms. This simulation provides some evidence that their methodology is effective in detecting cheating classrooms.
- The researchers sampled 100 classrooms (classrooms identified as cheating and classrooms identified as innocent) and retested those classrooms. The classrooms identified as potential cheaters had their score gains disappear in the retest. This provides further evidence of the effectiveness of their methodology.

These results support the survey results of Gay (1990) and Shepard & Dougherty (1991): it appears as though between 1-8% of educators cheat on standardized tests by manipulating their students’ answersheets or by giving answers to their students (two of the most blatant forms of cheating). It seems reasonable to guess that a higher percentage of teachers cheat for their students by manipulating the test administration process, the score reporting process, or the teaching process.

Educator cheating on standardized tests must be prevented and educators found to have cheated on tests must be punished. By invalidating student test scores and artificially inflating estimates of student achievement, this cheating behavior penalizes students – students with artificially high test scores will not receive resources set-aside for low-achieving students. Also, this cheating behavior makes it impossible to get an accurate snapshot of the effectiveness of educational policies, teaching

strategies, and curriculum changes. Furthermore, cheating teachers (who should serve as role models for their students) cast doubt on all teachers and the educational process.

The following recommendations can be made to detect and prevent each of the four methods of educator cheating (manipulating answersheets, administration, reporting, and teaching):

How to Prevent & Identify Educator Cheating	
Manipulating Answersheets	Use statistical indices as a deterrent; have a clear policy; have independent proctors; track the time between test completion and collection; look at erasures; surveys
* Altering a student's answersheet	
* Giving students answers	
* Pointing out mismarked items	
Manipulating Administration Process	Have a clear policy; control the number of test copies given and collected; surveys
* Giving students hints on test items	
* Rephrasing test items for students	
* Providing students extra time	
* Reading items that are supposed to be read by students	
* Excluding students from testing	
* Answering questions about test content	
* Instructing students to fill-in a specific answer for unanswered items	
* Providing accommodations or inappropriate special ed. placement	
Manipulating Reporting Process	Have a clear policy; have independent proctors; train proctors; have clear administration procedures
* Removing student test scores from the official records	
* Providing students with false ID numbers so their scores won't count	
* Misrepresenting data	
* Changing the criteria for proficiency	
* Conflicting accountability ratings	
Manipulating Teaching Process or Philosophy	Have a clear policy; get scores directly from the scoring service instead of from educators
* Teaching students specific test items or teaching during test	
* Practicing with actual test items	
* Teaching to the test	
* Narrowing the curriculum	

Empirical Study: Detecting Cheating in Two Iowa Districts

Data Description

Scores from the Math Concepts and Estimation test in the Iowa Tests of Basic Skills were gathered from two school districts in Iowa (Clinton Community School District and Davenport Community School District) in an attempt to detect 4th grade teachers who may be cheating for their students. Scores were gathered from 3rd grade students in 2001-02 (2002 data), 4th grade students from 2002-03 (2003 data), and 5th grade students from 2003-04 (2004 data). Student scores for the three-year period were collected by matching student names, student ID numbers, and student dates of birth.

After cleaning the data, eliminating students with missing scores, and eliminating all classrooms with 5 or fewer students, a total of 98 classrooms were contained in the data. These classrooms consisted of, on average, 16 students. To simplify percentile rank calculations, two additional classrooms were added to the data by randomly sampling 30 students from the entire data set, duplicating their data, and dividing those 30 students equally into the two created classrooms.

To detect potentially cheating teachers, the methods developed by Jacob & Levitt (2003) were implemented with this data set. It should be noted that there is no reason to believe this data set contains teachers who actually cheated on the ITBS; this data set was used because it was readily available to the researcher.

Index #1

The average change in student grade equivalent scores 2002-2003 and from 2003-2004 was calculated for each classroom in the sample. The classrooms were ranked by their average growth scores and each classroom was assigned a percentile rank score for 2002-2003 and for 2003-2004.

Then Index #1 was calculated for each classroom: $\text{Index \#1} = (\text{rank gain}_t)^2 + (1 - \text{rank gain}_{t+1})^2$.

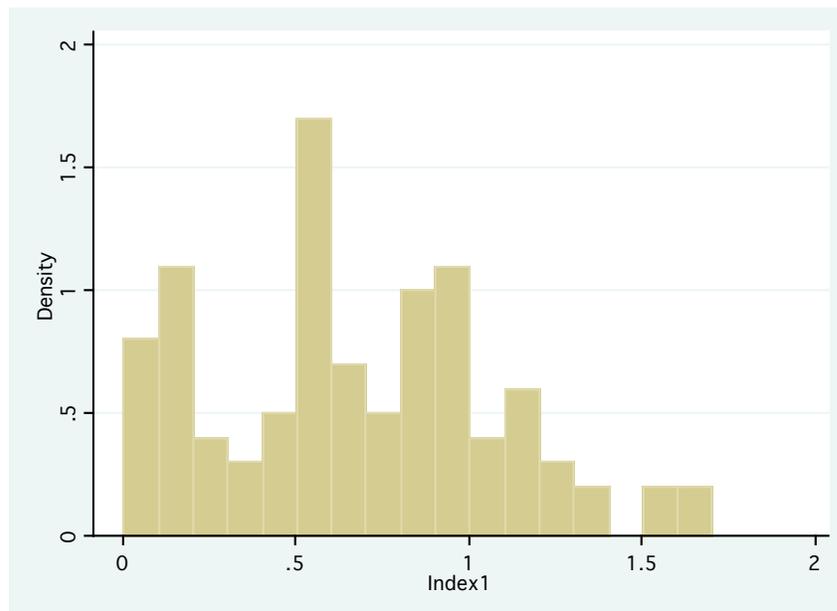
Before calculating the index, the following information was discovered:

- The average growth from 2002-2003 for students in this sample was 0.50 GE units
- The average growth from 2003-2004 for students in this sample was 0.96 GE units

- Highest student GE growth from 2002-03 was 5.0 units; the lowest was -7.8 units
- Highest student GE growth from 2003-04 was 5.9 units; the lowest was -2.9 units (the -7.8 is troubling; this data may need to be cleaned even more)
- The average growth for a classroom from 2002-2003 was 0.42 units
- The average growth for a classroom from 2003-2004 was 0.92 units
- Highest classroom GE growth from 2002-03 was 1.3 units; the lowest was -0.96 units
- Highest classroom GE growth from 2003-04 was 2.2 units; the lowest was -0.75 units

The following results were obtained from calculating Index #1 on the data:

- The lowest Index #1 value for a classroom was 0.004.
This classroom grew -0.6 units from 2002-03 and 2.15 units from 2003-04.
This provides evidence that this teacher did not cheat.
- The highest Index #1 value for a classroom was 1.664
This classroom grew 1.1 units from 2002-03 and 0.2 units from 2003-04.
This teacher may be a potential cheater; growth was unusually high for one year.
- The average value of Index #1 was 0.67. This index can range in value from 0.0002 to 1.96. With an observed minimum of 0.004, an observed maximum of 1.664, and an observed mean of 0.67, it appears as though no errors were made in the calculations.
- The following histogram displays the values of Index #1 for all 100 classrooms in the sample.



Keeping with the guidelines developed by Jacob & Levitt (2003), the 5 classrooms with the highest values of Index #1 are identified as having unusual test score fluctuations. After assigning a

identification number to each classroom, the following table displays the classrooms that had unusual test score fluctuations:

Classroom	2002-2003 Growth	2003-2004 Growth	Index #1 Value
15	0.5	-0.1	1.664
74	0.9	0.3	1.613
26	0.9	0.2	1.573
28	1.2	0.4	1.506
67	1.1	0.2	1.488

Index #2

The previous index uncovered five classrooms with unusual test score fluctuations. To identify teachers who may have cheated on the test, we must also uncover classrooms with unexpected test score fluctuations. To do this, the four measures discussed by Jacob & Levitt (2003) were computed for this data set.

First Measure: Most unlikely block of consecutive identical answers

Student responses to the Math Concepts and Estimation items were entered into STATA along with each student's race, gender, and 2001/2003 Math Concepts and Estimation grade equivalent scores. After recoding the student response strings, a multinomial logit model was used to calculate the probability that each student would select each answer choice for each item.

Selecting the first block of five items on the test (items 1-5), the likelihood of each student's answer string to those five items was calculated by multiplying the probabilities of each response to each item. These likelihoods were then multiplied across students within each classroom (who happen to have identical response strings) to obtain the likelihood that the students within a classroom would have selected those answers for the first five items on the test. These calculations were then repeated for items 2-6, items 3-7, items 4-8, and so on, such that the index was calculated for all blocks of five consecutive items on the test. The minimum value of this measure for each classroom across a 5-item block is then recorded for each classroom. The values of this measure,

which could theoretically range from 0-1, identify the degree to which a classroom's answer strings were likely to happen by chance. For the data set in question:

- The lowest value of the first measure was 1.67×10^{-12} (for classroom #28)
This identifies a classroom with highly unlikely answer strings for a 5 item block. It also identifies a classroom in which many students had identical answer strings.
- The highest value of the first measure was 0.0047 (for classroom #13)
Based on students' abilities and backgrounds, there was a relatively high likelihood of observing the answer strings provided by the students in this classroom. Therefore, there is no evidence to suggest that the teacher manipulated student answersheets.

Second Measure: Correlation among items within each classroom

Using the same multinomial logit model, residuals for each possible answer choice to each item were calculated. These residuals are then summed across students within each classroom for each answer choice. Next, the answer choice residuals are summed for each item. Plugging these values into the formula for the second measure provides a measure of the degree of correlation in student responses across the entire test. Assuming the calculations were correct, the following results were found:

- The lowest value of the second measure was 3.78 (classroom #6).
The students in this classroom did not have highly correlated answers.
- The highest value of the second measure was 16.41 (classroom #83).
The students in this classroom had a relatively high correlation across items. This provides evidence that perhaps the teacher manipulated student answersheets. On the other hand, it may just mean that the teacher stressed a particular topic in class.

Third Measure: Variance in the degree of correlation across test items within a classroom

This third measure is simply the observed variance of the item residuals from the previous measure. A high value on this measure indicates the students within a classroom had some blocks of highly correlated items and some blocks of uncorrelated items. This would indicate that perhaps a teacher altered the answers to a select group of items on the test. For this data set:

- The lowest value of the third measure was 6.41 (classroom #6).
The correlations among student responses did not vary across 5-item blocks

- The highest value of the third measure was 28.62 (classroom #37).
The correlations among student responses did vary across 5-item blocks
This teacher may have altered student answersheets.

Fourth Measure: Compare students to other equally-abled students in the sample

The final measure in Index #2 compares each student's answer string to the answers from other students of equal ability in the data set. First, the 2001-2002 Math Concepts and Estimation raw scores were used to place students into groups of equal ability. Then within each ability group, the proportion of students answering each item correctly was calculated. This proportion represents the expected score for each item (for a particular group of students with equal ability). Squared deviations between a student's actual item score and the expected item score are then computed and summed across items (call these SSD scores). Next, the deviation between these SSD scores and the average SSD score for all students of equal ability are computed. Finally, these values are summed across students within each classroom to provide a value for measure #4. For this data set:

- The lowest value of the fourth measure was 6.22 (classroom #22).
The students in this classroom had response strings similar to equally-abled students in other classrooms.
- The highest value of the fourth measure was 17.25 (classroom #67).
A large number of students in this classroom deviated from equally-abled students in other classrooms. These unusual student responses might indicate the teacher manipulated answersheets.

Index #2: Sum of the squared percentile ranks on all four measures

After calculating the four measures for each classroom, the classrooms were ranked on each measure and percentile ranks were calculated. These percentile ranks were then squared and summed to give the value for Index #2. Again, the 5 classrooms with the highest values on this index were identified as potentially cheating on the test. Noting that this index can range from 0.0004 to 3.9204, the following table displays the 5 classrooms with the highest Index #2 values:

Classroom	Index #2 Value
28	2.942
83	2.518
67	2.360
37	1.950
08	1.735

Potential Cheaters

Now that both indices have been calculated for the 100 classrooms in this data set, teachers who may have cheated on the test can be identified. Any classroom that falls in the highest 5% of values for Index #1 and Index #2 is identified as a potential cheater. The following table displays the top 5% classrooms on each measure.

Index #1 Classrooms	Index #2 Classrooms
15	28
74	83
26	67
28	37
67	08

One can see that two classrooms – Classroom #28 and Classroom #67 – can be identified as potential cheaters.

Since it is possible that these classrooms would be in the top 5% on both indices due to random chance, further evidence must be gathered to determine if the teachers did, in fact, cheat on this test. This further evidence could include an interview with the teacher or surveys from other teachers or the students themselves. In this project, this further evidence cannot be gathered. We can, however, examine the items responses from students in this classroom to see why these classrooms are identified as potential cheaters.

Classroom #28 consisted of only 12 students (after missing data were removed). With such a small number of students, it is easy to examine the answer strings for this classroom. The

following table displays the responses from all 12 students on 3 blocks of 5-items from the 2003 Math Concepts and Estimation test:

Student	Item Block A	Item Block B	Item Block C
1	A 4 C B A	D B A 4 B	D C 3 C C
2	A 4 C B A	D B A 4 B	D C D 4 C
3	A 4 C B A	D B A 4 B	2 1 D C 4
4	A 4 C B A	D B A 4 B	C 2 3 1 A
5	A 2 C B A	D B A A B	D 2 D C C
6	A C C B A	D B A A B	D C D C C
7	A 2 C B A	2 B A A B	D 1 D C C
8	A 4 C B A	D B A 4 B	D C 1 C C
9	A C C B A	D 3 A B B	D C D C C
10	A 2 C B A	D B A 4 B	D C D C C
11	A 4 C B A	D B A 4 B	D C D C C
12	A C C B A	D B A 4 B	D C D C C

In the table, a letter response indicates a correct answer (A, B, C, D) and a number response indicates an incorrect answer (1 = option A was incorrectly selected; 2 = option B was incorrectly selected, etc.). One can immediately see that for the three blocks of 5-items displayed in the table, most of the 12 students in the class had identical answer strings.

In Item Block A, six students had identical answers for all five items. The other six students differed by only one answer. In Item Block B, seven students had identical answer strings and another two students differed by only one answer. Over the five items in Block C, five students had identical answers and five more students differed by one answer. Out of the 180 item responses displayed in the above table, 156 of them (87%) are identical. The fact that so many students had very similar answer strings along with the fact that students in this class grew in achievement by only 0.2 GE units the following year seems to indicate that this teacher may have cheated on this test.

Weaknesses of this study

While the purpose of this study was simply to implement Jacob & Levitt's (2003) method for detecting cheating teachers on a real data set, it is important to note the limitations of this study. First, this data set was extremely small; consisting of only 98 actual classrooms. Although the sample size is small, the conclusions might yet be valid. If we assume that no classrooms cheated and that Index #1 and Index #2 are independent indices, then each classroom had only a 0.25% chance of being incorrectly identified as a potential cheating classroom.

Appendix A: Newspaper Stories on Educator Cheating (01/2005 – 06/2005):

- 01/2005: How public schools lie to parents and betray our children (Turtel, 2004).
 Ways in which teachers can cheat:
 1. Exclude/discourage poor students from taking the test
 2. Assign tests as homework or teach test items in class
 3. Minimal or nonexistent test security
 4. Allow students more time to complete the exam

In December 1999, a special investigation of NY City schools revealed two principals and dozens of teachers and assistant teachers were helping students cheat on standardized math and reading tests. <http://www.pressmethod.com/releasestorage/547.htm>
- 01/11/2005: When tests' cheaters are the teachers
 Kris Axtman – Christian Science Monitor
- An Indiana 3rd-grade teacher was suspended after being accused of tapping students on the shoulder when they marked wrong answers.
- Mississippi threw out portions of test scores at nine schools after discovering more than two dozen cases of alleged cheating. One 5th-grade teacher was fired after allegedly helping students on the writing portion of the test.
- Nine Arizona school districts invalidated portions of their test scores after teachers allegedly either read sections of the test to students or gave students extra time to finish. It was the state's 21st case of cheating since 2002.
- In Ohio, teachers are required to sign a code of ethics and are warned that their licenses may be revoked. Kentucky uses 6 different versions of the exam in one classroom. Mesa, AZ hires retired principals to wander among classrooms, monitoring the test.
- Dr. Levitt estimates that serious cases of teacher or principal cheating occur in about 5% of elementary school classrooms.

- 1/31/2005: WRAL Investigates teachers accused of changing EOG test scores
 Kelcey Carlson - Wilson, North Carolina
 Investigators are looking into testing irregularities on last year's end-of-grade tests for Sallie B. Howard School for the Arts and Education. In the past five years, there have been at least 10 investigations into testing irregularities. In that time, two teachers have had their licenses revoked. A third case is in litigation.
<http://www.wral.com/news/4148782/detail.html>

- 2/18/2005: Changed test answers lead to teacher resigning
 Robert Blair, a fourth grade teacher with 19 years experience at Palmer Elementary School, resigned after administrator's discovered altered answers on his pupils' state English Language Arts tests. Of the 22 students in the class, 17 or 18 tests were altered (based on erasures).
- There were 14 proven cases of teacher cheating in 2003-04 in New York.
<http://www.theithacajournal.com/news/stories/20050218/localnews/2003002.html>

- 03/2005: Principal suspended while cheating charge investigated
 Associated Press
 Jayne Gibbs, a principals with the Edison Schools in the Chester Upland School District has been placed on a paid leave while officials investigate allegations that she helped students cheat on standardized tests. 8th graders reported that she had given them the correct answers to questions on the Pennsylvania System of School Assessment test.
<http://www.observer-reporter.com/372790862273676.bsp>

- 3/24/2005: Teachers ordered cheating
Philadelphia Inquirer - Mezzacappa, D., Langland, C., & Hardy, D
State Education Commissioner Shirley Neeley moved to dissolve the Wilmer-Hutchins Independent School District board after 22 educators were found cheating. Teachers ordered students who finished the TAKS early to fix answers on other students' answer sheets.
<http://www.dfw.com/mld/dfw/news/local/11431968.htm>
- 4/12/2005: Miami-Dade principal faces cheating allegations
NBC6
Nicholas Emmanuel, principal of West View Middle School, was reassigned after allegations that he helped students cheat on the FCAT exams.
<http://www.nbc6.net/education/4370439/detail.html>
- 4/13/2005: Principal's test scores probed
Carrie Watters – Rockford Register Star
Tiffany Parker, principal of Lewis Lemon Elementary School in Rockford, is being investigated for allegedly manipulating the school's celebrated test scores in 2003.

Christian Science Monitor article refers to 21 cases of alleged cheating in Arizona in 2002, 24 in Mississippi in 2004, and 3 allegations in the past three years in Indiana.
<http://www.rstar.com/apps/pbcs.dll/article?AID=/20050222/NEWS0107/502220321/1004/NEWS>
- 4/19/2005: Students may have been helped
Cynthia Garza – Star-Telegram
Two A.M. Pate Elementary School teachers are no longer working after an investigation revealed that they may have provided students with answers to TAKS tests last year. Last year, 89% of Pate's 5th graders passed the test. This year, only 29% passed – the lowest rate among the district's 77 elementary schools.
Georgia Johnson = a 25-year veteran. 18 of 19 of her students passed the test; 6 had perfect scores
Mildred Lawrence-Medearis = 17 years experience. All 29 of her students passed the reading and math tests.

The Dallas Morning News reported that wide swings in TAKS scores indicated that nearly 400 schools may have helped students cheat on the TAKS.
<http://www.dfw.com/mld/dfw/news/local/11431968.htm>
- 4/26/2005: Cause of diving TAKS scores unclear
Holly Hacker – Dallas Morning News
A.M. Pate Elementary School in Fort Worth fell from 2nd highest in the district to dead last. Two teachers who administered the test last year were placed on administrative leave when an investigation found test irregularities.

Houston officials are investigating 15 schools. At Sanderson Elementary, two teachers helped students cheat on last year's 5th grade math test and possibly on reading and science tests. Based on statements from students, other teachers were involved in testing "irregularities" in 2003 and 2004.

Dallas announced cheating occurred at Harrell Budd Elementary.
<http://www.dallasnews.com/sharedcontent/dws/dn/education/stories/042605dnmetcheating.491c8d28.html>
- 5/03/2005: Test stress and cheating
Patti Ghezzi – Atlanta Journal-Constitution
No high-profile cheating case has emerged in Georgia, though the state has sanctioned some 159 educators for test administration problems in the last five years.
http://www.ajc.com/metro/content/custom/blogs/education/entries/2005/05/03/test_stress_and_cheating.html

05/04/2005: Teacher resigns over standardized testing procedures
ITBS Science test results for a 7th grade class have been invalidated and a teacher has resigned after administrators discovered he quizzed students on materials found in the ITBS. Teacher Gene Zweifel, of the Adel-Desoto-Minburn school district, had worked for the district for nearly two decades.

David Frisbie said similar things have occurred at four other Iowa schools.
<http://www.whotv.com/Global/story.asp?S=3302244>

05/09/2005: 'Cheating' probe halts Wai'anae school tests
Treena Shapiro – Honolulu Advertiser
8th graders apparently were given some test questions and answers in preparation for the high-stakes exams. The review sheet contained 12 questions from one section of the assessment.
<http://the.honoluluadvertiser.com/article/2005/Apr/09/In/In03p.html>

05/16/2005: Do teachers coach during WASL testing?
Jessica Blanchard - Seattle Post Intelligence
During a WASL math test, a student was made to erase and rewrite her answers to a question so many times that she wore a hole through the booklet page and had to reinforce it with scotch tape.
http://seattlepi.nwsourc.com/local/224429_was16.html

05/18/2005: Foothills, AZ teacher resigns over cheating incident.
Explorer News – Erin Schmidt
Allowed student to take exam one day early
<http://www.explorernews.com/articles/2005/05/18/education/education01.txt>

References

- Associated Press (2005). Changed test answers lead to teacher resigning. *The Ithaca Journal*, 2/18/2005. <http://www.theithacajournal.com/news/stories/20050218/localnews/2003002.html>
- Axtman, K. (2005). When tests' cheaters are the teachers: probe of Texas scores on high-stakes tests is the latest case in series of cheating incidents. *Christian Science Monitor*, January 11, 2005.
- Bay, L. (1995). *Detection of cheating on multiple-choice examinations*. Paper presented at the annual meeting of the American Educational Research Association.
- Blanchard, J. (2005). Do teachers coach during WASL testing? *Seattle Post Intelligence*, 5/16/2005. http://seattlepi.nwsourc.com/local/224429_was16.html
- Brumfield, R. (2005). High-stakes cheating spawns new market. *ESchool News*, March 9, 2005.
- Campbell, D. (1975). Assessing the impact of planned social change. *Social Research and Public Policies: The Dartmouth/OECD Conference*, Hanover, NH; Dartmouth College, The Public Affairs Center, p. 35.
- Carlson, K. (2005). WRAL investigates teachers accused of changing EOG test scores. 1/31/2005. <http://www.wral.com/news/4148782/detail.html>
- Cizeck, G. (1999). *Cheating on tests: how to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum.
- Cizeck, G. (2001). *An overview of issues concerning cheating on large-scale tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Cizeck, G. (2001). Cheating to the test. *Education Matters*, 1(1), 40-47.
- Cozin, M. (1999). *Cheating and its vicissitudes*. Research report from Raritan Valley Community College, NJ.
- Dwyer, D. & Hecht, J. (1994). Cheating detection: statistical, legal, and policy implications. Illinois State University
- Figlio, D. & Getzler, L. (2002). *Accountability, ability, and disability: gaming the system?* Working paper, University of Florida, 2002.
- Garza, C. (2005). Students may have been helped. *Star-Telegram*, 4/19/2005. <http://www.dfw.com/mld/dfw/news/local/11431968.htm>
- Ghezzi, P. (2005). Test stress and cheating. *Atlanta-Journal Constitution*, 5/03/2005. http://www.ajc.com/metro/content/custom/blogs/education/entries/2005/05/03/test_stress_and_cheating.html
- Green, J & Winters, M. (2003). Testing high stakes tests: can we believe the results of accountability tests? Civic Report 33, Center for Civic Innovation, Manhattan Institute, Manhattan, NY
- Hacker, H. (2005). Cause of diving TAKS scores unclear. *Dallas Morning News*, 4/26/2005. <http://www.dallasnews.com/sharedcontent/dws/dn/education/stories/042605dnmetcheating.491c8d28.html>

- Harrington-Lueker, D. (2000). When educators cheat. *The School Administrator*, December 2000.
- Houser, B. (1982). Student cheating and attitude: a function of classroom control technique. *Contemporary Educational Psychology*, 7(2).
- Jacob, B. & Levitt, S. (2003). To catch a cheat. *Education Next*,
- Jacob, B. & Levitt, S. (2004). Rotten apples: an investigation of the prevalence and predictors of teacher cheating. *Education Next*,
- Kantrowitz, B. & McGinn, D. (2000). When teachers are cheaters. *Newsweek*, June 19, 2000.
- Karabatsos, G. (2003) Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education* 16:4, 277-298.
- Kimmel, E. (1997). *Unintended consequences or testing the integrity of teachers and students*. Paper presented at the annual Assessment Conference of the Council of Chief State School Officers, Colorado Springs, CO, June 1997.
- Kohn, A. (1999). Tests that cheat students. *New York Times*, December 9, 1999.
- Ligon, G. (2000). Trouble with a capital t. *School Administrator*, 57(11), 40-44.
- Magnuson, P. (2000). High-stakes cheating: will the focus on accountability lead to more cheating. *Communicator*, February 2000.
- McCabe, D., Treviño, L. & Butterfield, K. (2001). Cheating in academic institutions: a decade of research. *Ethics and Behavior*, 11(3), 219-232.
- Mezzacappa, D., Langland, C., & Hardy, D. (2005). Principal accused of cheating on tests. *Philadelphia Inquirer*, 4/19/2005. <http://www.dfw.com/mld/dfw/news/local/11431968.htm>
- Million, J. (2000). When a principal cheats. *Communicator*, September 2000.
- Nathanson, C., Paulhus, D. & Williams, K. (2--5). *Predictors of behavioral measure of scholastic cheating: personality and competence but not demographics*. Unpublished.
- NBC6 (2005). Miami-Dade principal faces cheating allegations. 4/12/2005. <http://www.nbc6.net/education/4370439/detail.html>
- Nichols, S. & Berliner, D. (2005). The inevitable corruption of indicators and educators through high-stakes testing. Education Policy Studies Laboratory, Arizona State University, March, 2005. <http://www.greatlakescenter.org/pdf/EPSTL-0503-101-EPRU.pdf>
- Niels, G. (1996). Is the honor code a solution to the cheating epidemic. Research report from The Klingenstein Center, Teachers College, Columbia University, NY.
- Roberts, P., Anderson, J., & Yanish, P. (1997). *Academic misconduct: where do we start?* Paper presented at the annual conference of the Northern Rocky Mountain Educational Research Association, Jackson, WY, October 1997.
- Rooks, C. (1998). www.2cheat.com. Paper presented to the Teaching in the Community Colleges Online Conference, Honolulu, HI, April 7-9, 1998.

- Schmelkin, L, Kaufman, A., & Liebling, D. (2001). *Faculty assessments of the clarity and prevalence of academic dishonesty*. Paper presented at the annual meeting of the American Psychological Association. San Francisco, CA, August 24-28, 2001.
- Schmidt, E. (2005). Foothills, AZ teacher resigns over cheating incident. *Explorer News*, 5/18/2005. <http://www.explorernews.com/articles/2005/05/18/education/education01.txt>
- Shapiro, T. (2005). 'Cheating' probe halts Wai'anae school tests. *Honolulu Advertiser*, 5/09/2005. <http://the.honoluluadvertiser.com/article/2005/Apr/09/ln/ln03p.html>
- Shepard, L. & Dougherty, K. (1991). Effects of high-stakes testing on instruction. Paper presented at the annual meeting of the American Education Research Association, Chicago, IL.
- Snell, L. (2005). How schools cheat: from underreporting violence to inflating graduation rates to fudging test scores, educators are lying to the American public. *Reason*, June 2005. <http://www.reason.com/0506/fe.ls.how.shtml>
- Sotaridona, L. & Meijer, R. (2001) *Two new statistics to detect answer copying*. Research Report RR-01-07, University of Twente Research, Enschede, Netherlands
- Sotaridona, L. & Meijer, R. (2002) Statistical properties of the K-index for detecting answer copying. *Journal of Educational Measurement*, 39, 115-132.
- Spiller, S. & Crown, D. (1995). Changes over time in academic dishonesty at the collegiate level. *Psychological Reports*, 76(3)
- Turtle, J. (2004). How public schools lie to parents and betray our children. *Public Schools, Public Menace*. <http://www.pressmethod.com/releasestorage/547.htm>
- United Press International (2005). Teachers ordered cheating. *Washington Times*, 3/24/2005. <http://washingtontimes.com/culture/20050324-114418-7654r.htm>
- van der Linden, W. & Sotaridona, L. (2002). *A statistical test for detecting answer copying on multiple-choice tests*. University of Twente Research Report 02-04. Enschede, Netherlands
- Watters, C. (2005). Principal's test scores probed. *Rockford Register Star*, 4/13/2005. <http://www.rstar.com/apps/pbcs.dll/article?AID=/20050222/NEWS0107/502220321/1004/NEWS>
- WHO TV (2005). Teacher resigns over standardized testing procedures. 5/04/2005. <http://www.whotv.com/Global/story.asp?S=3302244>