b) Detection

Statistical methods to detect cheating do not, in fact, detect cheating. These methods, first developed in the 1920s to detect student cheating on multiple-choice tests, measure the likelihood of observing score gains, erasures, or answer patterns from student answer sheets. While the methods can identify unlikely large score gains, an improbably number of erasures, or unusual patterns of answers to items, they cannot determine if these events were due to cheating or simply due to chance. Furthermore, the methods cannot identify all forms of cheating. They can only attempt to detect cheating due to educators giving answers to students or changing student answer sheets.

While most statistical detection methods were developed to detect student cheating, they can also be used to identify possible educator cheating. After all, if multiple students within a class or school are flagged as having unusual patterns of answers or erasures, the educator in charge of that class or school may have cheated.

Early Developments

Saupe (1960) summarizes the development of statistical detection methods to detect students who copy answers from other students. Bird (1927, 1929) developed three empirical approaches to detect possible copying in which the number of matching incorrect answers between two student tests is compared to the distribution of identical incorrect answers observed from a large random sample of answer sheet pairs (Saupe, 1960, p. 476). Because the number of incorrect answers depends upon the ability level of the student, the empirical distribution was based on random samples of test pairs from students with similar total scores to the suspected cheater. If the tests from the suspected cheater and the source student (from whom the cheater allegedly copied) were found to have an unusually large number of identical incorrect responses in comparison to this empirical distribution, then the suspected cheating could be verified.

In an application of his method, Bird describes a test administration in which the test proctor observed suspicious behaviors from four examinees. Bird calculated an average of 4.0 identical incorrect answers from a random sample of pairs of tests from examinees not suspected of cheating. The suspected cheaters had 17, 25, 28, and 31 identical incorrect answers on the 149-item test. As validation of his method, Bird reports that three of the suspected cheaters "confessed guilt when confronted with the evidence" (Bird, 1927, p.261).

From Empirical to Chance Models

Rather than taking the time to develop the empirical distribution, Dickenson (1945) developed a method to determine the likelihood of identical answers occurring by chance. This method simply compares the actual number of identical incorrect answers on a pair of answer sheets to the expected number based on the number of possible responses to each item. Under this method, it is assumed that each incorrect item response is equally likely to be chosen by students. If k is the number of possible responses to each item, then $(k-1)/k^2$ is the expected proportion of incorrect answers on one test that are identical to another test. Dickenson suggested that if the observed proportion of identical incorrect answers is more than twice the expected proportion, then copying is implied (Saupe, 1960, p. 476).

Anikeeff developed another chance model using the binomial distribution to determine the likelihood of observing a specific number of identical incorrect answers between two tests. The number of observed identical incorrect answers between a pair of tests is compared to a binomial distribution with a mean of N, and a standard deviation of $\sqrt{Np(1-p)}$, where N is the number of wrong responses by the suspected cheater and p is the reciprocal of the number of possible responses to each item (Saupe, 1960, p. 476). A low likelihood of observing that number of matching incorrect answers may indicate copying.

In an application of his method, Anikeef concludes that his method is not effective at detecting copying. He concluded that his method would be useful in situations in which an examinee copies more than 16% of the answers from another examinee (Anikeeff, 1954).

Bellezza and Bellezza (1989) developed a method similar to Anikeef's method called Error Similarity Analysis (ESA). This method, used by the *Scrutiny!* software package (Advanced Psychometrics, 1993), calculates the total number of times all pairs of examinees chose identical incorrect answers for each item. The probability of observing a given number of identical incorrect answers is estimated by the binomial distribution:

$$\frac{w!}{c!(w-c)!}P^c(1-P)^{w-c},$$

where c is the number of common items answered incorrectly by a pair of examinees, w is the number of items for which the pair of examinees had identical incorrect responses, and P is the estimated probability of two examinees selecting an identical incorrect answer. Using this equation, or a method based on a standard normal approximation, Bellezza and

Bellezza were able to determine the likelihood of observing a specific number of identical incorrect answers between two examinees.

Holland (1996) describes another popular method to detect possible cheaters called the K-Index. This index, used by the Educational Testing Service (ETS), may be the most popular method currently used (Cizek, 1999, p. 141). Although limited information about this index exists, Holland describes it as a method used to "assess the degree of unusual agreement between the incorrect multiple-choice answers of two examinees" based on an estimate of the probability two examinees would agree on a response by chance (Holland, 1996, p. 5). The index uses the binomial distribution to model this probability. Sotaridona (2001) developed two indices, S_1 and S_2, similar to the K-Index except using the poisson distribution to model the probability of observing a specific number of identical examinee responses.

Incorporating More Information

Acknowledging the limitation in methods that only analyze matching *incorrect* answers, Saupe (1960) developed his method to detect copying on multiple-choice tests. In this method, the total number of items on the test, *K*, is partitioned into:

$$K = R_i + R_j - R_{ij} + W_{ij}$$

where R_i and R_j are the number of correct responses for students *i* and *j*, respectively; R_{ij} is the number of items both students answered correctly; W_{ij} is the number of items both students answer incorrectly; and w_{ij} is the number of items in which both students gave matching incorrect answers.

Under chance conditions, the expected number of items answered correctly by both students would be the proportion of all items answered correctly by student *i* multiplied by the number of items answered correctly by student *j*:

$$ER_{ij} = \frac{1}{K} R_i R_j$$

Thus, the regression of R_{ij} on the product $R_i R_j$ is of interest. This regression line can be written as:

$$\hat{R}_{ij} = b_{r1} R_i R_j + b_{r0}$$

The distance of an observed point $\left(R_i R_j, R_{ij}\right)$ from the regression line can be used to evaluate the observed degree of correspondence between the items answered correctly by a pair of students. If this distance exceeds ts_r, where *t* is the appropriate value from the t-distribution and s_r is an appropriate estimate of the standard error of estimate of R_{ij}, then the assumption of chance correspondence can be rejected at a specified confidence level (assuming a bivariate normal distribution of R_{ij} and $R_i R_j$). A correspondence index can be written as:

$$CI = \frac{R_{ij} - b_{r1} R_i R_j - b_{r0}}{ts_r}$$

A correspondence index greater than 1.00 is equivalent to rejecting the null hypothesis of chance correspondence between the items answered correctly by a pair of students.

The same logic is used to determine the correspondence between the incorrect answers from two students. If each item has *k* possible responses, the expected number of matching incorrect answers due to chance is:

$$Ew_{ij} - \frac{1}{k-1} W_{ij}$$

Using the regression of w_{ij} on W_{ij}, the correspondence index would be:

$$CI = \frac{w_{ij} - b_{w1} W_{ij} - b_{w0}}{ts_w}$$

Saupe suggests an advantage to analyzing the correspondence of correct and incorrect answers separately is that the evidence provided by both indices is non-overlapping and, therefore, complementary. In applying his model to a random sample of 150 pairs of tests, Saupe's correspondence indices identified 6 suspicious pairs. In an attempt to validate the results, Saupe examined seating charts and discovered that 5 of the 6 suspicious pairs came from students in adjacent seats. Saupe admits

the main disadvantage of his method is its use of a chance model – it is not reasonable to assume students randomly answer test questions (Saupe, 1960).

Attempting to overcome this disadvantage, Angoff (1974) developed 8 more indices to detect examinees copying on tests. Angoff's methods were all based on developing distributions of identical responses made by pairs of non-cheating examinees. The methods only differ in the combinations of independent and dependent variables used to develop the bivariate distributions. The degree to which an examinee's observed value on the dependent variable, conditioned on the observed value of the independent variable, deviates from the mean of the dependent variable from the distribution provides an index of cheating.

Angoff found that six of his indices were not effective in detecting cheating. Of the remaining indices, Angoff favored the method called the B Index. To use this index, the bivariate distribution of $W_i W_j$ and Q_{ij} is estimated from all examinees, where $W_i W_j$ is the product of the number of incorrect answers from two examinees and Q_{ij} is the number of identical incorrect answers for both examinees. For a pair of examinees, *A* and *B*, the observed values $W_a W_b$ and Q_{ab} are calculated. The following test statistic can then be used to determine whether the observed value of Q_{ab} is significantly different from the mean value of Q_{ij} :

$$t = \frac{Q_{ab} - \bar{Q}_{ij}}{S_{Q_{ij} \cdot W_i W_j}}.$$

While Saupe and Angoff used information from both incorrect and correct responses, Frary (1977) developed two indices based on estimating the probability of an examinee choosing a correct response, choosing each incorrect response, or choosing to omit each item. After dismissing his first index, Frary developed the following formula for his g_2 index:

$$g_2 = \frac{C - \sum_i \hat{P}\left(k_{ia} = k_{ib}\right)}{\sqrt{\sum_i \hat{P}\left(k_{ia} = k_{ib}\right)\left[1 - \sum_i \hat{P}\left(k_{ia} = k_{ib}\right)\right]}},$$

where C is the number of identical answers for a pair of examinees and $\hat{P}\left(k_{ia} = k_{ib}\right)$ is the probability that an examinee would choose the identical response of another examinee. Frary used piecewise linear functions of total test scores to estimate this probability.

After applying his method to actual test data and recommending its use to prevent cheating, Frary acknowledged three limitations. First, in order to use his method, one examinee must be identified as "the copier" and another examinee must be identified as "the source." This will not always be practical in large-scale testing situations. Second, the g2 index assumes that the probabilities of an examinee choosing each response to an item are constant, regardless of examinee ability. Third, Frary found that his method decreased in effectiveness for easier tests, stating, "If no examinees can answer as many as 90% correctly, the potential for detection is greatly enhanced" (Frary, 1977, p.253).

Hanson and Brennan (1987) continued to compare responses between pairs of examinees in their development of two more indices to detect possibly copying. The first method, Pair 1, uses the number of identical incorrect responses between a pair of examinees along with the length of the longest string of identical responses. The second method, Pair 2, uses the same information along with the percentage of maximum possible identical incorrect responses between two examinees.

In comparing their methods to the methods developed by Angoff (1974) and Frary (1977) on a simulated data set, Hanson and Brennan conclude that "it might not make a great deal of difference which of the statistical methods of investigating copying considered here are used" (p. 21). They do, however, recommend their method based on the interpretability of their indices.

Controlling for False Positives

In evaluating the effectiveness of the previously developed indices, Post (1994) concludes that while the indices may be used to scan for potential cheaters, "many existing statistical tests designed to detect copying on multiple-choice exams understate the Type I [false positive] error" (p. 140). Because this Type 1 error may be higher than specified, Post discourages using the indices to make accusations of cheating. Post attributes this inflated Type I error rate to the difficulty in estimating the probability of an examinee choosing each possible response to an item.

In an attempt to improve the estimation of item response probabilities and reduce the Type I error rate, Wesolowsky (2000) made a slight modification to Frary's method. Whereas Frary used piecewise linear functions of raw scores to estimate probabilities, Wesolowsky uses smooth distance iso-contours from location theory for estimation (p. 912). Also, while previous methods made assumptions about which examinee copied from another, Wesolowky's method simply examines the number of matching answers and ignores other suspicious patterns such as strings of identical answers. In developing a computer program to analyze answer sheets and employing a Bonferroni adjustment to control for overall Type I error rate, Wesolosky recommends his method as an effective way to screen for potential cheaters. This method, used in 2007 to scan for cheaters on the Texas Assessment of Knowledge and Skills, flagged more than 50,000 examinees as potentially having cheated (Benton & Hacker, 2007a, 2007b).

Incorporating Item Response Theory

Other researchers improved the estimation of the probability of an examinee choosing each possible response to an item by developing indices based on item response theory (IRT) models. In IRT models, the probability of an examinee choosing each response to an item is a function of the examinee's latent ability, θ, and characteristics of each possible item response. The item response characteristics of interest depend on the IRT model being used.

For the three-parameter logistic model, the probability of examinee a correctly answering dichotomously scored item i can be expressed as:

$$P_{ia} = P_{ia}(\theta_a) = P_{ia}(u_i = 1 \mid \theta) = c_i + \frac{1 - c_i}{1 + e^{-1.7a_i(\theta_a - b_i)}}, \text{ where}$$

a_i = the item discrimination parameter

b_i = the item difficulty parameter

c_i = the guessing parameter.

θ_a = the latent ability of examinee a

u_i = the examinee's scored response to the item

Using maximum likelihood or Bayesian methods, values for the item response and examinee ability parameters can be estimated under the assumptions of the specified IRT model. Under the assumption of local independence, the probability of observing a string of n item responses from an examinee with ability θ_a is equal to the product of the probabilities of the individual item responses:

$$P(u_1, u_2, \ldots, u_n \mid \theta) = \prod_{i=1}^{n} P_i^{u_i} (1 - P_i)^{(1-u_i)}.$$

Thus, given an observed string of responses from an examinee, the above formula can be used to estimate the likelihood of observing that response string or the probability of observing a different string of responses.

Rather than simply estimating the probability of an examinee answering dichotomously scored items correctly or incorrectly, IRT models can be used to estimate the probability of an examinee choosing each possible choice to a multiple-choice item. Bock's Nominal Model calculates the probability of choosing response *u* on multiple-choice item *g* with *m* possible responses as:

$$P_u = \frac{\pi_u}{\sum_{h=1}^{m} \pi_h},$$

where π_h represents a scale value directly related to the probability that response *h* is chosen on a specific test item. This model can be reparameterized as:

$$P_u(\theta) = \frac{e^{(a_u\theta + c_u)}}{\sum_{h=1}^{m} e^{a_h\theta + c_h}}$$

from which item response discrimination and difficulty parameters, *a* and *c*, and examinee ability parameter θ can be estimated. Again, under the local independence assumption, the probability of observing a specific string of item responses can be estimated from these item response and examinee ability estimates.

Person-Fit and Aberrant Response Indices

The application of IRT models to detect possible cheating has been theorized via person-fit and aberrant response indices. These indices measure the extent to which the observed pattern of responses from an examinee with ability level (θ) deviates from the response pattern expected under the chosen IRT model. For example, an examinee whose ability exceeds the difficulty level (b) of an item would have a high probability of answering that item correctly. Likewise, an examinee whose ability is less than the difficulty of an item would have a high probability of answering that item incorrectly. When an examinee's response string fits this pattern across most or all items on the test, the model "fits" the person. Aberrant response strings (high-ability examinees incorrectly answering easy items, low-ability examinees correctly answering difficult items, or examinees choosing unusual responses on a multiple-choice test) indicate poor model fit. Person-fit and aberrant response indices measure the degree to which the chosen IRT model fits the observed responses from an examinee.

More than fifty person-fit indices have been developed to detect aberrant responders (Karabatsos, 2003; Meijer & Sijtsma, 2001; Thiessen, 2007). These indices, displayed in **FIGURE ABERRANT,** attempt to detect students who provide unusual responses due to luck, language deficiencies, random guessing, low-motivation, misaligned answer sheets, or cheating (Meijer, 1996; Thiessen, 2004).

Person-fit indices can be classified into three categories: deviation-based, covariance-based, and likelihood-based. Deviation-based indices, such as Wright and Stone's (1979) Outfit Mean Square index, sum the squared standardized differences between an examinee's scored response to an item and the expected probability of that correct response. A large difference would indicate a disagreement between the model and the examinee.

Covariance-based indices, such as Tatsuoka's (1984) Extended Caution Indices, measure the degree to which an examinee's item responses deviate from the Guttman Perfect Pattern. Specifically, these indices calculate the ratio of the covariance between an examinee's responses and item difficulty estimates to the covariance between the average probability of correct responses across all examinees (estimated via IRT models) and the item difficulty estimates.

Likelihood-based indices, such as Levine and Rubin's (1979) log-likelihood function ℓ_0, evaluate the shape of a likelihood function. Given an examinee's responses to a set of test items with known item parameters (*a, b, c*), the ability level of an examinee (θ_a) can be estimated by maximizing the likelihood function:

$$\ell_0 = \ln\left[L(\mathbf{u} \mid \theta; \mathbf{a}, \mathbf{b}, \mathbf{c})\right] = \ln\left[\prod_{i=1}^{n} P_i^{u_i} Q_i^{(1-u_i)}\right] = \sum_{i=1}^{n}\left[u_i \ln(P_i) + (1-u_i)\ln(Q_i)\right]$$

The value of θ_a that maximizes this function represents the examinee's most likely ability level given the observed item responses and the estimated item parameters. Examinees whose item responses conform to the IRT model produce likelihood functions with relatively high maximum values. Examinees whose responses deviate from what is predicted by the IRT model produce low maximum values of the likelihood function (Davey, et. al, 2003, p.6). Thus, the relative magnitude of ℓ_0 can be used to identify examinees whose responses are aberrant.

Aberrant Response Indices to Detect Examinee Cheating

While person-fit indices detect many forms of aberrance, some indices were developed specifically to detect examinee cheating. The first set of these indices is referred to as *optimal person-fit statistics.* Levine and Drasgow's (1988) $\lambda(x)$ index was developed to test the null hypothesis of normal examinee responses (based on a chosen IRT model) against the alternative hypothesis that an examinee's responses are consistent with a specific aberrant response model. Thus, if a researcher can specify and estimate an alternative hypothesis of examinee cheating, optimal person-fit statistics can provide the maximum detection rate for aberrance (Krabatsos, 2003, p. 282).

Unfortunately, studies have shown that aberrant response indices are ineffective at detecting cheating (Chason & Maller, 1996; Iwamoto, Nungester, & Luecht, 1996). Demonstrating this ineffectiveness, test security firm *Caveon* analyzed a simulated data set using their six aberrant response indices. With the data set simulated so that examinees cheat on 10% of the test items, the firm's six aberrant response indices were only able to detect 41 (1.2%) of the 3,283 simulated cheaters and none of the five simulated cheating schools (Impara, et al., 2005).

Wollack argues that IRT-based aberrant response indices are inadequate in detecting cheating (specifically, student copying) because the statistical significance of these indices "does not depend on the similarity between the suspected copier's responses and those of a neighboring examinee" (p. 307). He stated that previous attempts to detect cheating, based on Classical Test Theory (CTT), are also inadequate. Wollack argues that since CTT item statistics are dependent on the sample of examinees tested, the expected similarity between a pair of examinees depends "largely on the performance of the other examinees in the sample, rather than only on the two examinees of interest" (p. 307). In order to overcome the apparent limitations of CTT-based indices and IRT-based aberrant response indices, Wollack developed his ω-index to specifically detect copying.

The ω-index is similar to Frary's g2 index in that it attempts to estimate the probabilities associated with each possible item response. It differs from the g2 index in that it uses Bock's Nominal Model to model these probabilities. The ω-index also attempts to control for Type I errors by examining the seating chart used in test administration and analyzing only pairs of examinees who sit physically close enough to make copying possible.

In applying the ω-index, each examinee is analyzed as a possible copier. If the test is administered in a classroom with a rectangular seating chart, the examinees sitting to the copier's left, right, front left, front center, or front right are analyzed as potential sources for copying. Assuming item responses are locally independent, as is necessary to use IRT models, the expected number of identical responses and its associated variance can be modeled with the binomial distribution. Thus, an index similar to the g2 can be compared to the standard normal distribution to evaluate the statistical significance of the degree of similarity between two examinees' item responses:

$$\omega = \frac{C - \sum_i \hat{P}\left(k_{ia} = k_{ib} \mid \theta_a, K_b, \xi\right)}{\sqrt{\sum_i \hat{P}\left(k_{ia} = k_{ib} \mid \theta_a, K_b, \xi\right)\left[1 - \sum_i \hat{P}\left(k_{ia} = k_{ib} \mid \theta_a, K_b, \xi\right)\right]}},$$

where C is the number of identical answers for a pair of examinees and $\hat{P}\left(k_{ia} = k_{ib}\right)$ is the probability that examinee *a* would choose the identical response of examinee *b* given the ability θ_a of examinee *a*, the item responses K_b of examinee *b*, and the matrix of item parameters, ξ.

Wollack (1997) compared his index to Frary's g2 index on two data sets with three types of simulated copiers. The first type of copying was random copying, in which randomly selected items were simulated as being copied from another examinee. The second type was difficulty-weighted copying, in which the more difficult items were simulated as being copied. The third type was random-strings copying, in which strings of 4 consecutive items were simulated as being copied. Between 10-40% of items were simulated as being copied by 5% of examinees for each type of copying. Wollack ensured that in this simulation, simulated copiers copied answers from examinees with higher ability estimates sitting close to them, as would be expected in reality. Based on these simulations, Wollack concludes that the ω-index is more effective in both detecting copying and controlling Type 1 error rates than the g2 index when a seating chart is available (Wollack, 1997, p.312).

Adjacent Seating Methods

Similar to Wollack's ω-index, Kvam (1996), Roberts (1987), and the National Board of Medical Examiners (Cizek, 1999) developed methods that require knowledge about which examinees sat adjacent to others. In applying Kvam's method, examinees in a classroom are randomly administered two forms of the exam. Minor changes are made to the questions so that the two multiple-choice forms have different answers. After administering the exams, maximum likelihood methods are used to estimate the probability that an examinee copies an item response from an adjacently-seated source given that the examinee does not formulate an answer to the question (Kvam, 1996, p.239).

The *score-difference method* developed by Roberts (1987) also relies on two test forms randomly administered to examinees. After administering the test forms, each answer sheet is scored using the answer keys from both test forms. The difference between the scores obtained from the two answer keys provides an indication of possible cheating. While conceptually simple, Roberts concludes that his method is "seriously flawed and has little to recommend" (p. 77).

In administering its exams, the National Board of Medical Examiners (NBME) uses two methods to detect potential cheating. The first method, *adjacent-nonadjacent analysis*, requires the test to consist of at least two parallel parts. Each part must be administered in separate testing sessions in which test takers are randomly assigned to seat locations. Under these conditions, it is expected that the response patterns from any two examinees would have the same degree of similarity regardless of where the examinees are seated. If two examinees are seated adjacently during at least one test session and it is discovered that their responses are more similar in those sessions than would be expected if they were seated apart, then potential copying has been detected.

The NBME method uses a simple 2x2 chi-square test for independence to detect potential cheaters. The test statistic is calculated from the following table using the given formula, and then evaluated for significance:

	Number of items answered incorrectly		
Seating Location	Identical Responses	Different Responses	**Total**
Adjacent	a	b	(a+b)
Nonadjacent	c	d	(c+d)
Total	(a+c)	(b+d)	(a+b+c+d)

$$\chi_1^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

The NBME method uses a simple 2x2 chi-square test for independence to detect potential cheaters. The test statistic is calculated from the following table using the given formula, and then evaluated for significance:

Methods to Detect Educator Cheating

The previously discussed methods all attempt to detect cheating (usually in the form of copying) by students on tests. Only two serious attempts have been made to detect potential educator cheating on tests. The first method, erasure analysis, is currently used by several states in auditing their testing programs. The second method, developed by Jacob and Levitt (2003), was used to detect cheating educators in Chicago Public Schools.

Some erasure analysis methods attempt to detect cheating by identifying answer sheets with an unusual number or pattern of erasures. It is assumed that a large number of erased answers might indicate an educator who is manipulating answer sheets. Other erasure analysis methods only look at wrong-to-right erasures – items in which an incorrect answer was erased and changed to a correct answer.

The Louisiana Educational Assessment Program has written and implemented erasure analysis procedures to audit its state testing program. These procedures require the test scoring contractor to scan every answer sheet for wrong-to-right erasures and compute the mean and standard deviation for each subject at each grade level. The scoring contractor must then flag examinees whose wrong-to-right erasures exceed the state mean by more than four standard deviations as potential cheaters. The number of erasures for each potential cheater, along with the proportion of wrong-to-right erasures found on their answer sheet, are then reported to the State Superintendent of Education and the scores from the potential cheaters are voided from official records (Louisiana Educational Assessment Program, 2003).

One problem with using erasure analysis to flag potential cheaters is determining how many erasures indicate potential cheating. To address this problem, Qualls (2001) examined 4,553 answer sheets from the 1995-96 administration of the *Iowa Tests of Basic Skills* to determine the typical erasure behavior of examinees in a low-stakes testing environment. Qualls found that more than 90% of examinees changed 3 or fewer answers per test and only 2% of examinees changed more than 6 responses. While the erasure behavior depended on the test subject, less than 7% of item responses were erased, on average. Furthermore, the results indicate that it would be rare for an examinee to erase and change more than 15% of items on a single test. Qualls found that the first or second items on the test were most frequently erased and that students with one erasure had about a 50% chance of changing an incorrect response to a correct response. In a focus on wrong-to-right erasures, Qualls found that examinees gained an average of between 0.167 and 0.494 points per erasure on the tests. Perhaps these results could be used to develop new and improve upon current erasure analysis methods for high-stakes tests.

Jacob & Levitt Methods to Detect Educator Cheating

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:

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

- 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.

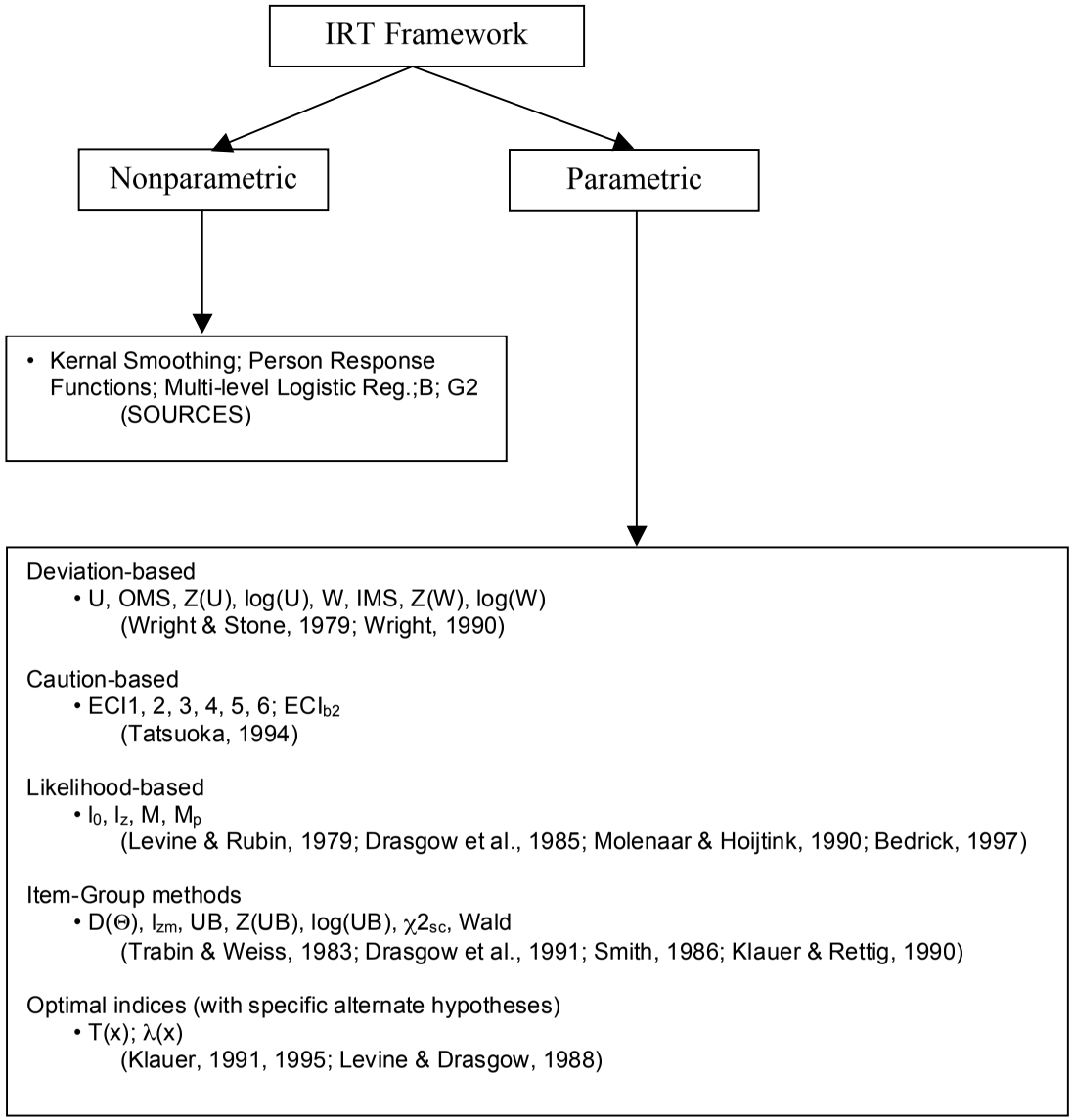
Putting It Together: Detecting Potential Cheating Classrooms

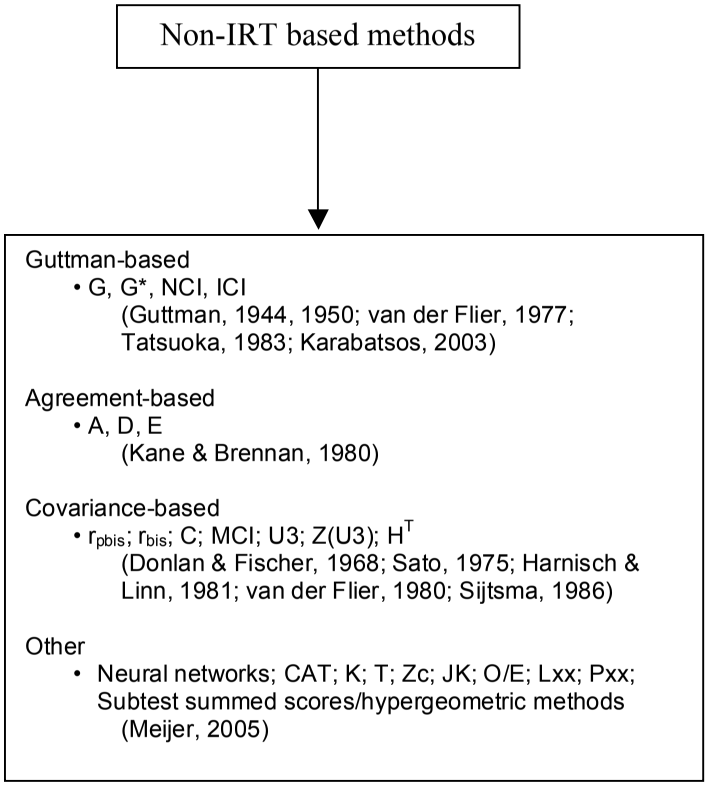
Jacob & Levitt argued that taken individually, the above two indices do not detect teachers who manipulate answersheets. After all, there are always going to be (innocent) classrooms with unexpected score fluctuations and there are going to be (innocent) classrooms with improbable answer patterns. The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

(Thiessen, 2006)

FIGURE ABERRANT: Aberrant Response Detection Methods & Indices





See (Thiessen, 2004) and (Karabatsos, 2003) for formulas and discussion of these indices