MULTILEVEL DETECTION OF POSSIBLE TEST TAMPERING THROUGH ERASURE ANALYSIS

by

SHANSHAN QIN

(Under the Direction of Allan Cohen and Seock-Ho Kim)

ABSTRACT

Cheating on high-stakes standardized tests, especially by educators to meet accountability requirement, becomes a widespread serious problem receiving more and more attention. In the limited literature of statistical detection of this test security violation, erasures analysis has the potential to serve as useful data forensic tool. This dissertation developed and compared several methods of erasure analysis for tests with dichotomously scored multiple-choice items in order to identify potential tampering at individual and groups levels. A large-scale grade 8 reading test data set was used to explore characteristics and interrelationships among existing and proposed detection methods. Based this real data set, simulated data sets were generated, containing erasures due to random answer changes, misalignment, speededness, and tampering. Type I error and power rates of different methods were evaluated across simulation settings different in strategies of making illegal wrong-to-right erasures and in numbers of involved examinees and groups.

INDEX WORDS: Erasure analysis, test security, educator tampering, data forensics, Item response theory, mixed modeling

MULTILEVEL DETECTION OF POSSIBLE TEST TAMPERING THROUGH ERASURE ANALYSIS

by

SHANSHAN QIN

BA, Renmin University of China, 2007

Master, University of Georgia, 2011

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2016

MULTILEVEL DETECTION OF POSSIBLE TEST TAMPERING THROUGH ERASURE ANALYSIS

by

SHANSHAN QIN

Major Professor:	Allan S. Cohen
	Seock-Ho Kim
Committee:	Gary J. Lautenschlager
	Zhenqui Lu

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2016

DEDICATION

This is dedicated to my daughter Arya, my husband Yujun, and my parents on the other side of the world.

ACKNOWLEDGEMENTS

There were so many people who have made this dissertation possible and because of whom my doctoral study experience has been one that I will cherish forever. It is difficult to overstate my gratitude to Dr. Allan Cohen and Dr. Steve Cramer, who provided me the graduate assistantship to support my entire study and academic travels to the fascinating world of psychometric society. They stimulated and nourished my interest in this dissertation project, and I would have been lost without them. Throughout my data analysis and dissertation-writing period, Dr. Cohen provided spring-like encouragement, countless detailed advice, and magnificent patience.

It has been my good fortune to be a student of Dr. Seock-Ho Kim, Dr. Gary J. Lautenschlager, and Dr. Zhenqui Lu. They introduced me to the world of educational measurement and applied statistics. Their knowledge, experience and insights made me feel this career so exciting. They exemplified lively to me the objectivity, integrity, and persistence in academic research. I am also indebted to all other professors and all graduate students in Quantitative Methodology for constructing a stimulating and fun environment in which to learn and grow.

My daughter Arya is a source of my joy and a motion of my work. My husband Yujun, together with his colleagues in natural science, are also my role models in academia. Dedicating this dissertation to them is a small token of my love and appreciation.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

STATEMENT OF PROBLEM

1.1 Definitions and Existing Problem

Cheating is defined in *Merriam-Webster's Dictionary* (Merriam-Webster's, 2012) as "depriving of something valuable by the use of deceit or fraud" and "violating rules dishonestly". Test cheating can be defined as any action that violates the rules for administering or taking a test. The Atlanta Public Schools Cheating Scandal of 2011, described as one of the largest in United States history (Frysh, 2011; Johnson, 2011; Resmovits, 2011), has thrust cheating on high-stakes standardized tests, especially by educators to meet accountability requirement, into the national spotlight.

Standardized tests are defined as "any test that's administered, scored, and interpreted in a standard, predetermined manner" (Popham, 2005). The No Child Left Behind Act (NCLB, 2002) requires that states use standardized assessments to measure Adequate Yearly Progress, or AYP. The original goal of NCLB was that states were required to have 100 percent of their students achieve proficient status on state assessments by the 2013-14 school year.

One outcome of the NCLB is that schools or districts that fail to make AYP, may be closed, reconstituted, or taken over by state governments. In 2007, the U.S. Government Accountability Office (2007) estimated that 27% of Title I schools under NCLB restructuring replaced all or most staff. This translates into about 1.35% of all Title I schools and less than 1% of the nation's schools. In some cases, individual teachers can secure merit pay depending on their students' test-scores. For example, at an Atlanta school with the highest pass rate, every

employee got a $2,000 bonus (Frysh, 2011). Test performance is also heavily linked to state competition for federal funding. For example, the U.S. Congress created a $99 million federal grant program called the *Teacher Incentive Plan* (TIP) in 2006 and raised its funding to $600 million in 2010 (U.S. Department of Education, 2012). One major target of TIP is to reform teacher and principal compensation systems so that teachers and principals are rewarded for increases in student achievement.

In such a climate, test security is more likely to be breached. This commonly occurs in areas such as examinees' cheating, non-examinees' cheating or test tampering, exposure of test material, inappropriate test administration and preparation (Fremer & Ferrara, 2013). Thiessen (2007) reviewed public reports on this issue and estimated that more than 25% of educators tampered with high-stakes tests by manipulating answer sheets, test administration, test reports or teaching-learning process. A nationwide survey among 23,000 American high school students revealed that 51 percent of students admitted cheating on a test at least once during the past year (Josephson Institute of Ethics, 2012). USA Today in March 2011 identified 1,610 examples of anomalies in which public school classes boasted what analysts regard as statistically rare gains on statewide tests in six states and the District of Columbia between 2009-10 (Toppo, Amos, Gillum, & Upton, 2011). In addition to the Atlanta scandal, widespread test tampering also has been reported in other parts of the country as well. Tampering has been reported in Baltimore (Mathews, 2012), El Paso (Fernandez, 2012), Philadelphia (Herold, 2012), Pittsburgh and other districts of Pennsylvania (Chute & Niederberger, 2012), as well as in Toledo and Columbus, Ohio, and in St. Louis (Bock, 2012). In Los Angeles principles of six charter schools were ordered by their director to break the seals on state tests and help students prepare for the exams with actual test questions (Blume, 2011). Over a dozen educators at two elementary schools in

New York were accused of providing improper test assistance to examinees, including coaching and violating test protocol (Baker, 2013). Amrein-Beardsley, Berliner, and Rideau (2010) collected the responses of 3,085 educators in Arizona to their email surveys on high stakes test cheating. About 10% of those respondents reported knowing of colleagues who erased and changed test answers of students, and over 17% were aware of colleagues who actually gave answers to students.

It is clear that the violation of test security is a widespread problem. This behavior threatens the psychometric integrity of test scores as accurate measures of the effectiveness of teaching, curricula, and educational policies. It also penalizes students who, because of inflated scores, are misplaced into classes for which that they are unprepared, or deprives them of resources intended to increase their achievement. Educators who cheat fail as role models to students and destroy the reputation of their educational systems (Mathews, 2012). Unfortunately, security breach doesn't receive much scrutiny in testing programs. Camara, in an interview with the Atlanta Journal and Constitution (AJC) noted that $760 million a year is spent on testing required by NCLB) but "…the one area where we haven't devoted the same energy is standardizing the administration of the test to deter cheating" (AJC; Pell, 2012, para 12). Pell also reported that of the 46 state education departments that responded to the AJC survey on test security, 41 states allowed teachers to monitor tests for their own students, 21 states did not look for an improbable number of changes from wrong to right answers in 2012, and 24 states did not conduct an analysis looking for improbable test improvements in 2012. Of the 25 states using independent proctors in 2012, most sent fewer than 20 monitors to oversee testing in hundreds of schools.

1.2 Motivation and Purpose of this Study

Test security in psychometric research area generally falls into the following categories: post hoc detection of answer sharing and answer copying during exams, test tampering, and preknowledge of exam content. Among these topics, answer copying received somewhat more attention, although there is still a limited, albeit growing literature compared with other aspects of cheating on educational tests (e.g., Angoff, 1974; Belov & Armstrong, 2010; Frary, 1993; Marianti et al., 2014; McLeod, Lewis, & Thissen, 2002; Wesolowsky, 2000; Wollack, 1997; Wollack & Cohen, 1998; Wollack &Maynes, 2011; Sotaridona, Van der Linden, & Meijer, 2006; Zhang, Searcy, & Horn, 2011).

The current interest in educator cheating mainly focuses on detecting tampering on tests composed of multiple-choice items. This tampering might be in the form of erasing students' original wrong responses and replacing them with correct answers, or guiding students to make the answer changes. Modern optical scanners can be programmed to record the mark densities of all choices in answers sheets. This information can be used to help differentiate noise, erasures, and final answers (Cohen & Wollack, 2006). This information can potentially be used to tell whether an erasure is from a right to wrong (RW), wrong to another wrong (WW), or wrong to right (WR) answer. Erasure counts could vary by program, since optical scanning sensitivity settings often vary.

Statistical detection of test tampering is not necessarily based on erasure analysis. Measures of person-fit of final item responses have been used to flag aberrant test takers (e.g., Birenbaum, 1985; Cronbach, 1946; Guttman, 1944; Huang, 2012; Karabatsos, 2003; Li & Olejnik, 1997; Sijtsma & Meijer, 2001; Spearman, 1910; Thurstone, 1927), and then aggregated to group levels. This approach has been questioned, because such aberrances may be explained

by other reasons than misconduct, such as fatigue, carelessness, unpreparedness, and partial mastery of exam content (Fremer & Ferrara, 2013).

Cheating alters the measurement quality for affected examinees. In this case, it may be appealing to catch the breach with differential item functioning (DIF) analysis, which is able to identify groups with different propensities for correctly endorsing items (Pine, 1977). One caveat to the use of DIF analysis in tampering detection is that an act like test tampering could be very random such that it may not create symmetric impact on item parameters. The effect of tampering may not necessarily be as universal as might those of gender, ethnicity, or special pedagogy. Which items to be tampered with can vary depending on who is doing the cheating and on the original performance of individual examinees. If an item is only tampered for a few examinees, whether the item parameter estimation would be distorted is questionable. To date, it does not appear that the efficacy of DIF analysis has been studied for use in detecting cheating subjects.

Score changes over time have also been studied as indicators of cheating (e.g., Benton & Hacker, 2004; Jacob & Levitt, 2003; Kao, Woo, & Gorham, 2013; Toppo et al., 2011; Skorupski & Egan, 2012). This approach usually needs testing data of the same examinees or groups from at least three administrations.

Compared with the approaches mentioned above, erasure analysis is based on both initial or pre-erasure responses and final or post-erasure responses. Erasures have the potential to serve as useful indicators of cheating. Unusually large numbers of WR erasures within classrooms and schools are used in some states as indicators of potential test tampering (Herold, 2012; Mathews, 2012; Pell, 2012). Other states calculate the proportion of WR changes to total erasures per student (Otterman, 2011). Three, five, and even eight standard deviations at minimum away from

means are commonly used criteria for indicators of test tampering (Primoli, Liassou, Bishop, & Nhouyvanisvong, 2011). Those analyses were reported to be successful for detecting several high-profile incidences of tampering in public schools (Almasy, 2015; Lattanzio, 2014; Otterman, 2011).

However, practical methods for analyzing erasures as possible indices of tampering have not been well studied. Although some alternatives do exist (e.g., Van der Linden, & Jeon, 2012; Wollack, Cohen, & Eckerly, 2013), there is still a lack of research of these methods. Questions remain such as how detection accuracy might change for different types and numbers of erasures and for different contexts (e.g., low stakes vs high stakes testing, paper-pencil vs computerized testing), how are detection methods related, and how should individual level results be aggregated to group levels. Thus, the main purposes of this dissertation are (1) to develop new erasure detection methods for a test with dichotomously scored MC items in order to identify potential tampering at individual and groups levels, (2) to compare the Type I error and power rates of current and proposed methods across simulation settings different in strategies of making illegal erasures and in numbers of involved examinees and groups, and (3) to explore characteristics and interrelationships among current and proposed detection indices in empirical testing data.

CHAPTER 2

LITERATURE REVIEW

A review is provided in this chapter describing the different existing methods used for detection of test tampering, including both non-erasure and erasure analysis.

2.1 Person-fit Measures on Final Item Responses

Person-fit statistics calculated from the final item responses of individuals have been used to detect test response patterns distinguished from those of test norms (Birenbaum, 1985; Cronbach, 1946; Guttman, 1944; Huang, 2012; Karabatsos, 2003; Li & Olejnik, 1997; Sijtsma & Meijer, 2001; Spearman, 1910; Thurstone, 1927). Meijer and Sijtsma (2001) reported over 40 statistics available to test person-fit. At least five factors can explain an examinee's spuriously high or spuriously low score: cheating, carelessness, lucky guessing, creative responding, and random responding (Meijer, 1996a, 1996b).

Research showed that Guttman-based personal-fit indices, like Within-Ability-Concern and Beyond-Ability-Surprise Indices (Wc&Bs), Sato Caution Index (SCI), and Modified Caution Index (MCI), generally performed better than the IRT-based counterparts, including Norm Conformity Index (NCI), Extended Caution Indices (ECI4z), and Outfit mean square (OUTFITz) and l_z (Huang, 2012; Karabatsos, 2003). In Huang's study (2012), the marginal power of most Guttman-based personal-fit indices could pass .90 but with no report on their Type I error rates. The common assumption of these Guttman-based indices is that, a normal test taker should correctly answer most items that are either easier than or matching his or her ability level, and

should fail most items with difficulty levels higher than that ability level (Tatsuoka & Linn, 1983). For a test with *k* dichotomously scored items, to calculate a typical Guttman-based index, like MCI, items are first sorted in ascending order by classical item difficulty level, which are defined as the percentages of persons who get the item right. Then, MCI is defined as

$$\text{MCI} = \frac{w_t - b_t}{H_t - L_t}, \tag{1}$$

where *t* is the total raw score of an examinee, w_t is the sum of item difficulties for failed items among the first *t* items after sorting, b_t is the sum of item difficulties for correctly answered items among the remaining *k*-*t* items, H_t is the sum of difficulty levels of difficulty levels of the first *t* items, and L_t is the sum of difficulty levels of the last *t* items (D'Costa, 1993a). The emphasis only on items with unexpected changes instead of all items may explain the higher power of Guttman-based personal-fit indices over IRT-based person-fit indices.

Thus far, no research has been reported on group-level analysis of these personal-fit indices. At this point, in the absence of erasures, use of aberrance as the primary evidence for test fraud is not recommended (Maynes, 2013).

2.2 Score Changes Over Time

Tampering is intended to increase test scores. Looking for unexpected or unexplained score changes is a major focus of efforts to detect tampering. Benton and Hacker (2004) and Toppo et al.(2011) examined standardized tests of six states and the District of Columbia to determine whether test results from the previous lower grade could help detect aberrant score changes in the next higher grade. For every school, the average test scores of one grade in one year would be predicted on the basis of the average test scores of the same grade level of that school in the previous year, or on the basis of the average test scores of the same group of students, when they

were at a lower grade in the previous year. Outliers would be the ones with a standardized residual greater than 3.0. While the administrators of school districts acknowledged that some oddities revealed from the regression analysis need further investigation, educators also raised questions on that methodology (Smith, 2005; Toppo *et al,* 2011). One concern is whether or not examinee groups from different years are comparable. A school may have more high-achieving students enrolled in the current year, and students may not move to higher grades between the two test administrations. There are always reasonable argument for the boost of test performance, such as inspired teaching, curriculum changes tailored to the tests, extras tutoring, hard work and dedication (Smith, 2005; Toppo *et al,* 2011). From a researchers view, the homoscedasticity assumption of regression analysis is violated due to the bounded nature of test scores. This results in variances smaller at the upper and lower ends of data (Maynes, 2013). Also, regression towards the mean could punish high-performing schools for having large and positive but legitimate residuals (Maynes, 2013).

A more appropriate analysis should keep track of scores of single students. This can be seen in the approach using hierarchical growth models proposed by Skorupski and Egan (2012), which can also address the problem of heteroscedasticity and regression to the mean. The Skorupski and Egan model for an individual's score at time point *t* nested within student *i* within Group g (indicated by "G") is,

$$Y_{igt} = \beta_0 + \beta_{1g}(G) + \beta_{2t}(T) + \beta_{3gt}(GT) + \varepsilon_{igt}. \quad (2)$$

Y_{igt} is the vertically linked score of student *i.* G and T are the indicators for group and time respectively. β_0 is the grand mean, and all other effects are centered around it. β_{1g} is the mean effect for Group g, representing the group performance. β_{2t} is the main effect for time *t* ,representing the average growth rate. β_{3gt} is the interaction effect between group and time,

representing the unique change for Group g at time *t.* It is assumed that tampering would cause an unusually large value of β_{3gt}. Skorupski and Egan standardized this into Cohen's δ effect-size statistics (Cohen, 1988) to estimate how many standard deviations a group's β_{3gt} is away from a baseline value. To measure the chance of each group having a true β_{3gt} greater than a baseline value, Poster Probability of Cheating (PPoC) was used in the way similar to that of Posterior Probability of Passing by Wainer, Wang, Skorupski, and Bradlow (2005). A spurious group would be the one with δ greater than .5 and also a PPoC larger than .75. Those cutoff values were selected based on analysis of real data taken from Skorupski and Egan (2011). Skorupski and Egan (2012) simulated testing data at three time points in each of 50 replications, each of which had a sample size of 4,650 nested within 60 groups.

With a Type I error rates of 0.04 at α =.05, the combination of δ and PPoC had .07 in power for detecting cheating at time 1 (baseline), .71 for cheating at time 2, and 1 for cheating at time 3. Results indicated that since δ and PPoC both measures deviance from baseline values, if a measurement time point is contaminated bv aberrance or cheating already and set a high baseline value, it could be hard to find any other time point showing large positive deviance from the baseline, and the power of the method would be weaken. However, if cheating or aberrance occurs at a time point after an established baseline, the methodology could be very effective at flagging potentially cheating groups. The number of time points in the growth models depends out the judgement of investigators. The longer the time interval between measurement points, the more historical events can account for score changes. Also, due to Federal privacy laws, school systems might refuse to release student-level data (Smith, 2005).

2.3 The Marginal Distribution of Erasures

Whether one flags individuals on the basis of the proportion of WR erasures to total erasures, or judges based on the frequency of total or WR erasures, what's used is the marginal distribution of erasures. For group-level detection, those indices are further weighted by within-group samples size, such as the number of examinees in a school. Aberrant subjects are ones with values of indices some predefined number of standard deviations away from the measures of central tendency of those distributions. As mentioned in the previous chapter, how many standard deviations it should be varies a lot in practice, in part because of differences between tests (Data Recognition Corporation, 2009; Primoli et al., 2011; State of Louisiana, 2010). The justification for the use of criteria was not always clear or not adequately justified. Also, researchers found significant positive relationships between ability estimates and the proportion of WR erasures to total erasures in four large-scale, K-12 achievement testing programs (Primoli et al., 2011). To a certain extent, this may provide some evidence that the number of WR erasures may be a function of ability rather than educator tampering.

2.4 Two Educator-Tampering Indicators by Jacob and Levitt

Jacob and Levitt (2003) developed two indicators for classroom-level cheating: Unexpected Test Score Fluctuations (UTSF), and Suspicious Answer Strings (SAS). Both indicators are based on the sums of squares of multiple flagging statistics. This formulation provides a way to summarize multiple single evidences. A classroom needs to have multiple statistics at high levels to yield large values of UTSF and SAS hence the bar for a signal to be alarming is higher than the one using only a single index. (i.e., multiple statistics at high levels) for a signal to be alarming (Maynes, 2013).

The logic of UTSF is that test score gains that result from cheating do not represent real gains in knowledge and would not be expected to continue on future exams. This, in turn, would result in large fluctuation in measures of score gains. For classroom *c* at time *t* on the test of subject *b*, the UTSF described by Jacob and Levitt can be given as follows

$$SCORE_{cbt} = (rank\_gain_{c,b,t})^2 + (rank\_gain_{c,b,t+1})^2. \quad (3)$$

For the students at time *t*, compute their score gains from time *t*-1 to time *t* and from time *t*-1 to time *t*+1, and then get the averages of these two gains (i.e., $gain_{c,b,t}$ and $gain_{c,b,t+1}$) in every classroom. Rank $gain_{c,b,t}$ and $gain_{c,b,t+1}$ separately for all classrooms. A suspicious classroom would have a very high $rank\_gain_{c,b,t}$ but very low $rank\_gain_{c,b,t+1}$ and end up with a high $SCORE_{cbt}$.

The SAS described by Jacob and Levitt refers to evidence that indicates changing of the answers of consecutive questions. The SAS is composed of the squared ranks of four aberrance measures: (1) the probability of the least likely blocks of identical answers in every classroom, (2) the average of the variance of response patterns in every classroom , (3) the variance of the variance of response patterns in every classroom, and (4) the difference between the responses of students in every classroom and those of other students in the population conditioned on their test scores (see appendix A in Jacob & Levitt, 2003). Student's final response strings were predicted based on their previous and subsequent achievements, responses of other examinees with the same raw scores, and demographic information.

No distribution theory was applied in deriving the critical values for UTSF and SAS. Rather, the empirical distributions were calculated by sorting the two indices in ascending order among all classrooms in the sample. Jacob and Levitt used the 80th, 90th, and 95th percentiles as cut-off values. Using the Chicago public schools' standardized test data sets, they estimated that

4 to 5 percent of elementary school classrooms involved in serious teacher or administrator cheating.

Jacob and Levitt conducted two simulation studies to observe the method's power and Type I error rates. In one simulation, students were randomly assigned to hypothetical classrooms so that no classroom participated in cheating. The method didn't flag any classroom. In the other simulation, selected classrooms' answers were altered to mimic teacher cheating. Three factors were manipulated: percentages of victim examinees in a classroom (25%, 50%, or 100%), numbers of tampered items (3 or 6 items), and cheating strategies (cheating on the same blocks of items or random item). The Type I error rates and power of their simulation studies were not completely reported. The paper only showed that the highest power was close to .6 when when six same successive questions were changed for half of a class, with a Type I error rate of 2%.The power was weakened further when reducing the numbers of tampered items and examinees, or choosing items randomly. .

2.5 The Two-Stage Modeling Method by Van der Linden and Jeon

Van der Linden and Jeon (2012) describe a two-stage IRT analysis on both erased responses and final responses that does not target on particular patterns of answer changes. Van der Linden and Jeon assumed regular test reponses were composed of two stages with enough time to allow answers to all items:

1) A first stage in which examinee j produces the initial response $U_{ij}^{(1)}$ to item i. (j = $1, \ldots, N$ examinees).
2) A final stage in which examinee j reviews the initial response and creates the final response $U_{ij}^{(2)}$. If no change or erasure happen, $U_{ij}^{(1)} = U_{ij}^{(2)}$. If a WW erasure occurs

or an incorrect response is confirmed, $U_{ij}^{(1)} = 0$ and $U_{ij}^{(2)} = 0$. If a RW erasure occurs, $U_{ij}^{(1)} = 1$ and $U_{ij}^{(2)} = 0$. If a WR erasure occurs, $U_{ij}^{(1)} = 0$ and $U_{ij}^{(2)} = 1$.

For dichotomously scored items, the 3-parameter logistic model was used to fit the initial responses,

$$\Pr\left(U_{ij}^{(1)} = 1\right) = c_i + (1 - c_i)\frac{\exp\left\{a_i\left(\theta_j^{(1)} - b_i\right)\right\}}{1 + \exp\left\{a_i\left(\theta_j^{(1)} - b_i\right)\right\}}, \tag{4}$$

where $\theta_j^{(1)}$ is the ability of examinee j, and a_i, b_i, and c_i are item parameters for discrimination, difficulty, and lower asymptote, respectively, for item *i*. It is assumed that, prior to the operational use of the items, their item parameters will be estimated with satisfactory precision along with acceptable model fit. In this method, the final responses depend on the initial ones, therefore, their response functions are assumed to be conditional probability functions with the same person ability as for the initial responses but different item parameters for discrimination and difficulty,

$$\Pr\left(U_{ij}^{(2)} = 1 | U_{ij}^{(1)} = 1\right) = \frac{\exp\left\{a_{1i}\left(\theta_j^{(1)} - b_{2i}\right)\right\}}{1 + \exp\left\{a_{1i}\left(\theta_j^{(1)} - b_{2i}\right)\right\}}, \tag{5}$$

and

$$\Pr\left(U_{ij}^{(2)} = 1 | U_{ij}^{(1)} = 0\right) = \frac{\exp\left\{a_{0i}\left(\theta_j^{(1)} - b_{0i}\right)\right\}}{1 + \exp\left\{a_{0i}\left(\theta_j^{(1)} - b_{0i}\right)\right\}}. \tag{6}$$

This method assumes there is no guessing parameter at this stage, because answer changing takes extra time and thinking and most people probably would not randomly guess on an item twice. It is possible to add guessing parameters to the models, but the sparseness of erasures could render its estimation very low accuracy. The focus of this detection method is Equation 6, which

describes the probability of making a WR erasure on item *i*. The combined probability of making a WW erasure or confirming a wrong initial response at item *i* is $1 - \Pr\left(U_{ij}^{(2)} = 1 | U_{ij}^{(1)} = 0\right)$.

The estimation of Equation 6 is based on the subset of the final responses for which $U_{ij}^{(1)} = 0$. The data in this subset are potentially sparse, leading to unstable or even unbounded slope estimates. To estimate this model, Van der Linden and Jeon (2012) used a Bayesian approach with weakly informative priors on the parameters. Equation 6(?) can be re-written as

$$\Pr\left(U_{ij}^{(2)} | U_{ij}^{(1)} = 0\right) = \text{logit}^{-1}\left(b_{oi}^{*} + a_{oi}^{*}\theta_{j}^{(1)}\right), \quad (7)$$

where $b_{oi} = -a_{oi}^{*}b_{oi}^{*}$. Following the suggestion of Gelman et al. (2008) for logistic regression models, Van der Linden and Jeon used a Cauchy distribution for the prior for the slope parameters $a_{oi}^{*}$with location 0 and scale 2.5, and a Cauchy distribution with location 0 and scale 10 for the prior for the intercept parameter b_{oi}^{*}. The two priors were assumed to be independent.

The detection index that Van der Linden and Jeon proposed is the probability of observing at least E_j, the total number of WR erasures for examinee *j*, among all items that examinee *j* incorrectly answered at the first stage. Let I_j denote the number of those items. E_j is considered to be the result of I_j independent Bernoulli trials, each with a different probability $P_{ij} \equiv \Pr\left(U_{ij}^{(2)} = 1 | U_{ij}^{(1)} = 0\right)$ given by Equation 6. The complement of P_{ij} is Q_{ij}. The distribution of E_j is known as the generalized or compound binomial distribution. If $E_j = x \leq I_j$, there would be $\binom{I_j}{x}$ combinations of WR erasures and non-WR-erasure responses. Let C_{lj} be one combination, $C_{lj} = \{d_{l1j}, d_{l2j}, \ldots, d_{lI_jj}\}$, where $d_{lmj} = 1$ indicates the presence of a WR erasure on the m^{th} item, and the sum of elements of C_{lj} is x, then

$$\Pr\left(C_{lj}\right) = {P_{1j}}^{d_{l1j}}{Q_{1j}}^{(1-d_{l1j})}{P_{2j}}^{d_{l2j}}{Q_{1j}}^{(1-d_{l2j})} \ldots {P_{I_jj}}^{d_{lI_jj}}{Q_{1j}}^{(1-d_{lI_jj})} \quad (8)$$

For instance, assuming that examinee *j* didn't correctly answer items 1, 4 and 8 at the first stage and then made two WR erasures on the first and forth items at the second stage. This would be flagged as an irregular case if the probability of making not less than two WR erasures among three items incorrectly answered at first is smaller than a critical value α. That is

$$\alpha \leq \Pr(E_j \geq 2) = \Pr(E_j = 3) + \Pr(E_j = 2), \quad (9)$$

where $\Pr(E_j = 3)$=$\prod_{i\in\{1,4,8\}} P_{ij}$, and $\Pr(E_j = 2)$=$P_{1j}P_{4j}Q_{8j} + P_{1j}Q_{4j}P_{8j} + Q_{1j}P_{4j}P_{8j}$.

Van der Linden and Jeon applied this method to an empirical data set from a large-scale Grade 3 math assessment that consisted of the responses of 2,555 students to 65 items. The items were part of a larger set that had been pre-calibrated prior to their administration to students in this data set. At $\alpha = .05,\ .01,$ and .001, 2.62%, 1.29%, and 0.47% of examinees were flagged, though no suspicious patterns, such as blocks of adjacent items with large positive residuals or other communalities between flagged students, were found. A simulation study evaluating the efficacy of this two-stage IRT analysis was not provided.

One assumption of this method is that test takers have enough time to answer the items and review each of their responses. If this assumption is violated, it will be unclear whether an unchanged incorrect response, which usually would be scored as $\left(U_{ij}^{(1)} = 0, U_{ij}^{(2)} = 0\right)$, is actually $\left(U_{ij}^{(1)} = 0, U_{ij}^{(2)} = \text{missing}\right)$. This miss-scoring would generally lead to an underestimation of Equation 6 for the items that were not reviewed. In such a case, even a few erasures would seem unlikely under the null hypothesis of no tampering. This, in turn, would lead to Type I errors. Those high-achieving examinees who work quickly and are able to review more items would be penalized in this case. Van der Linden and Jeon conducted a simulation study, and showed that the value e_j^*, satisfying $\Pr(E_j \geq e_j^*) = \alpha$ increased, when more items

were scored as $\left(U_{ij}^{(1)} = 0, U_{ij}^{(2)} = \text{missing}\right)$ for the majority of examinees, resulting in the loss of power. However, the lack of being able to review on one item didn't affect other items. The estimated P_{ij} for an item that examinee *j* reviewed was found to be almost the same as when other items were not reviewed by most examinees. This was attributed to the finding by Gelman et al. (2008) that the particular priors used could enable the estimation of subsets of parameters with extreme numbers of missing data while not affecting any of the others.

2.6 The Erasure Detection Index

As the total number of items increases, the computation of the previous generalized binomial distribution would get more and more demanding. Wollack et al. (2013) introduced another IRT-based approach to modeling erasure data which utilized a normal approximation to the generalized binomial distribution. Abilities and item parameters are estimated using only the subset of responses for which no evidence of tampering exists, for example, the subset in which responses to erased item are treated as missing values. Let $I_{E,j}$ denote the set of items for which examinee *j* produced erasures, then $\hat{\theta}_{j[i \notin I_{E,j}]}$ refers to the estimated ability based on only all non-erased item i's not in $I_{E,j}$. The *observed* score on erased items, $X_{j,I_{E,j}}$, is computed as the sum of raw right/wrong scores across those items. Its expected value, $E\left(X_{j,I_{E,j}}\right)$, is computed as

$$E\left(X_{j,I_{E,j}}\right) = \sum_{i \in I_{E,j}} P\left(x_{ij} = 1 \mid \hat{\theta}_{j[i \notin I_{E,j}]}\right) . \quad (10)$$

Any item response model appropriate for the data may be used to estimate $P(x_{ij} = 1)$. The standard error for $X_{j,I_{E,j}}$ is given by

$$SE\left(X_{j,I_{E,j}}\right) = \sqrt{\sum_{i \in I_{E,j}} P\left(x_{ij} = 1 \mid \hat{\theta}_{j[i \notin I_{E,j}]}\right)\left[1 - P\left(x_{ij} = 1 \mid \hat{\theta}_{j[i \notin I_{E,j}]}\right)\right]}. \quad (11)$$

The erasure detection index (EDI) is defined as,

$$EDI = \frac{X_{j,I_{E,j}} - E\left(X_{j,I_{E,j}}\right) + C}{SE\left(X_{j,I_{E,j}}\right)}. \quad (12)$$

EDI measures how far the observed value of WR erasures deviates from the expected. In the cheating-free circumstance, it is assumed to follow a normal distribution with mean 0 and variance 1. *C* in Equation 8 represents a correction for continuity. Controls for inflated false positive rates have been used with other indexes of this structure for low ability examinees when the number of involved items are small (Chen & Wollack, in preparation; Van der Linden & Sotaridona, 2006). $C = -.5$ was used by Wollack et al.(2013).

To evaluate the performance of EDI, Wollack et al. (2013) simulated erasures due to random changes, misalignment, speededness, fixed-number tampering, and score-based tampering for a sample of 250,000 examinees to 50 five-choice items. A random erasure in their simulation refers to the situation in which an examinee accidently fills in a wrong answer location on the answer sheet, then changes it to the intended answer, or when a student initially answers an item one way, but upon reconsideration, changes that response. A student may bubble in the answer to item *i* in the position for item *i + 1* (or *i – 1*) on the answer sheet, and repeat the same error for a string of consecutive items, then recognize and change the wrong markings, resulting in misalignment erasures. Speededness erasures or string-end erasures occur when an examinee is able to revise his/her original responses to items at the end of the test which he/she randomly answered at first in anticipation of insufficient time. In the fixed tampering conditions, a person other than the examinee changes the wrong answers for selected examinees to a specific set of items to correct ones. The particular number of tampered items in score-based tampering can change, when an administrator or teacher, for example, makes just enough WR

erasures to move the level of an examinee's performance to just above a passing standard. The target distribution of the total number of stimulated erasures was binomial (50,.02), which means that 60% of the examinees have at least one erasure, in the range of 50% to 70 % (Primoli, et al. , 2011; Qualls, 2001).

The simulations by Wollack et al. (2013) excluded all erased items from the estimation of ability for tampered examinees, reducing both the bias and root mean square errors (RMSEs) on θ as defined by the following:

$$Bias = \frac{\sum_{j=1}^{N}\left[\left(\hat{\theta}_{j[i\notin I_{E,j}]} - \theta_j\right)\cdot\left(1 - I(T_j)\right)\right]}{\sum_{j=1}^{N}[1 - I(T_j)]}, \tag{13}$$

and

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}\left[\left(\hat{\theta}_{j[i\notin I_{E,j}]} - \theta_j\right)^2\cdot\left(1 - I(T_j)\right)\right]}{\sum_{j=1}^{N}[1 - I(T_j)]}}. \tag{14}$$

$I(T_j)$ in the above two equations is an indicator function, which equals 0 when the test responses of examinee j were tampered with, and 1 otherwise. Results suggested that, for non-tampered examinees, bias and RMSEs were negligible, even when simulees had as many as 15 benign erasures, which are erasures due to random changes, misalignment, and speededness.

Results from Wollack et al. were estimated at α =.00001, .0001, .0005, .001, .005, .01, and .05, across ability quintiles. Most Type I error rates were well controlled after continuity correction, except in the first quintile. Also, false positive cases tended to occur more often for string-end erasures in the lower quintiles. Among the three types of fixed-number tampering considered (i.e., 5-item, 10-item, and 15-item erasures), at α =.005, .01, .05, the average power of EDI with continuity correction across ability quintiles exceeded .5 in general. The smallest

average power, .225, was found in the combination of 5-item tampering and $\alpha = .0005$. Results suggested in general that the lower the quintile, the higher the power. Still, among all fixed-number tampering conditions, at $\alpha = .001, .005$, the average power across the first, second and third ability quintiles exceeded .9 in general, except for the 5-item tampering, of which the average power was .68. The scored-based tampering was only simulated among the first, second and third ability quartiles. After continuity correction, Wollack et al. reported average power was .57 and .75, respectively, for $\alpha = .001$ and .005. These results were for individual-level detection. Group-level detection wasn't discussed.

2.7 Other Statistical Methods for Detecting Test Tampering

Tampering resulting in some identical item responses in a group of examinees could also be detected by similarity-based techniques for detecting test collusion. Such techniques include, for example, cluster analysis (Wollack & Maynes, 2011) and factor analysis (Zhang, Searcy, & Horn, 2011). However, those analyses are not suitable for tests for which the set of items can vary with examinees, like CAT. Even for tests where the same test forms are administrated to large numbers of examinees, similarity-based techniques could lose power, if incorrect responses of examinees in a group cannot be matched.

UTSF and SAS (Jacob & Levitt, 2003) are examples of combining multiple flagging statistics to detect tampering. The Data Recognition Corporation (2009) summed flags of several group-level statistics, including high numbers of wrong-to-right erasures, and improbable yearly changes in scale scores, test participation rates, and the percentages of proficient and advanced students. In that study, it was arbitrarily decided that three or more flags should trigger the request for further investigation. Maynes (2009b) used a test to assess whether all values of the

15 indices used for flagging schools were not larger than population or expected values. The test compares the following probability of school j with a chosen α level,

$$P_j = 1 - [\max(1 - p_1, \ldots, 1 - p_{15})]^{15}, \quad (14)$$

where p_1, …, p_{15} represent the 15 indices that measured the probabilities of test responses, response time, response similarity, the percentages of maximum scores, and the percentage of retest attempts that violated test policy. Formulas of these indices were not provided in Maynes' paper. P_j less than α support the act of flagging out school j. Maynes didn't report the power and Type I error rates of this method.

2.8 A Proposed Index for Individual-level Detection: Modified EDI

Qin and Cohen (2013) conducted a study on EDI using an empirical date set containing test responses to 45 four-choice items. Some limitations concerning calculations in that study were found. First, the inclusion of cases with only a few erasures made the distribution of EDI highly negatively-skewed. This can be seen in Figure 1, which contains all examinees' EDIs, and in Figure 2, which contains the EDIs of examinees with more than one erasure. EDI in Figure 2 ranged from -8.53 to 3.34, instead of starting from -36.2953 as in Figure 1. It appears to be distributed more symmetrically with just a single mode. Second, the inclusion of WW erasures and RW erasures could potentially conceal cases with unusually large numbers of WR erasures. For example, one examinee had 22 erasures, consisting of 8 WR erasures, 13 WW erasures, and 1 RW erasure. Although this examinee's WR erasures might appear to be suspicious, the EDI in the original form was only 0.48, i.e., not significant. When only the 8 WR erasures were used in calculating this examinee's EDI, however, the value was 3.78 and significant at $p < .05$. Another issue in the calculation of EDI is that extremely large positive values of EDI can occur for

examinees with only one erasure when their probability of selecting an option is very low. Even with a very accurate method, it's difficult to defend flagging one examinee who has made only a single erasure. Therefore, in the current study, EDI will be calculated based on WR erasures only for examinees with more than one WR erasure, denoted as EDI_WTR.

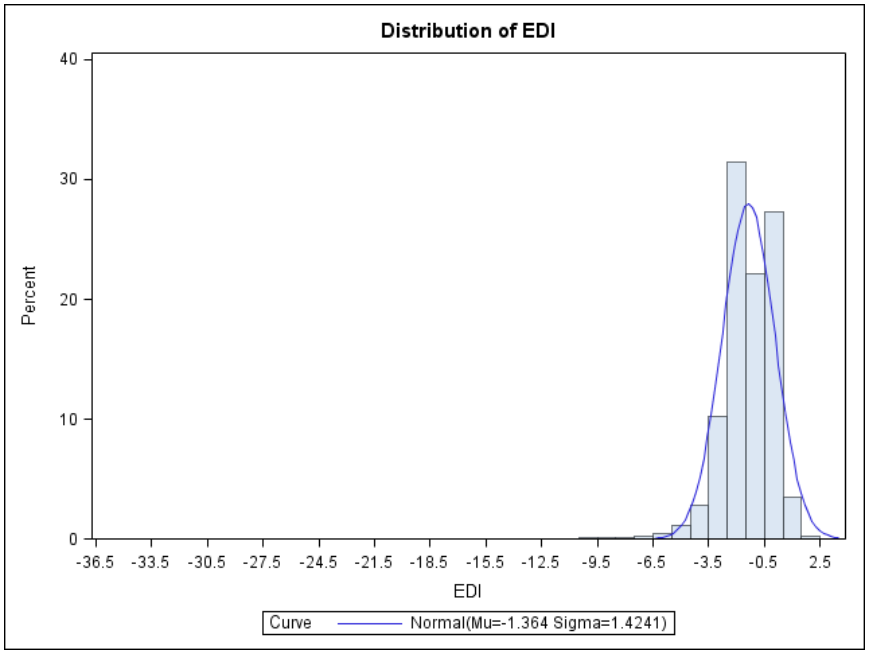

Figure 1: The histogram of EDI from all empirical examinees

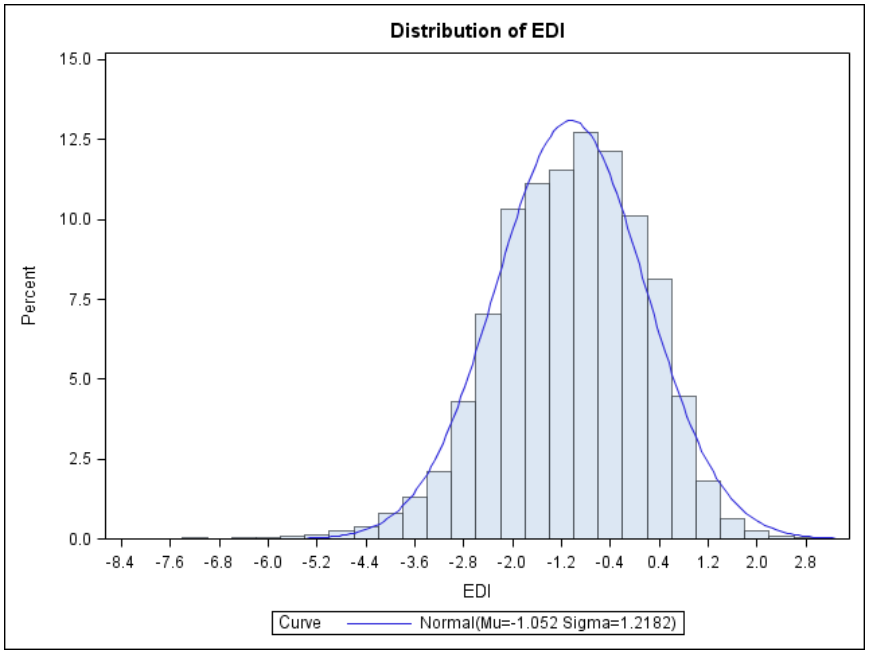

Figure 2: The histogram of EDI from empirical examinees with more than one erasure

2.9 A Proposed Index for Individual-level Detection: The increment on estimated ability due to erasures

Instead of looking at the score change over time, in this dissertation, we will exam the potential of using unusually large score gains following erasures as an indicator of test irregularity. Shorter time intervals between two scores may prevent historical events (such as remedy lessons, enhanced test preparation, or more confidence due to the experience of having taken the test previously) to affect large score changes. As mentioned above, IRT ability estimates may not be sensitive to certain answer changes from wrong to right ifmaximum likelihood estimation is used as this uses weights of each item by all of its item parameters and also considers whether it is answered correctly (Baker & Kim, 2004). Consequently, the power of using the increment based on estimated ability following erasures as an indicator of tampering

may be limited. However, as there is relatively little published research exploring this type of index, this approach also will be explored here.

Let $\hat{\theta}_{j[i\notin I_{E,j}]}$ be the estimated ability of examinee j based on only non-erased items as in EDI. $\hat{\theta}_{j[post]}$ is the estimated ability based on the final responses including any with erasures. The standard errors for these estimates are $SE_{\theta_j[i\notin I_{E,j}]}$ and $SE_{\theta_j[post]}$ respectively, formulated as,

$$SE_{\theta_j[i\notin I_{E,j}]} = I_{\hat{\theta}_{j[i\notin I_{E,j}]}}^{-1/2}, \quad (15)$$

And,

$$SE_{\theta_j[post]} = I_{\hat{\theta}_{j[post]}}^{-1/2}. \quad (16)$$

$I_{\hat{\theta}_{j[i\notin I_{E,j}]}}$ and $I_{\hat{\theta}_{j[post]}}$ are information functions of the two estimated abilities (See Equation 11 in Bock, 1996). The difference between the two estimated abilities would be

$$Diff_{\theta_j} = \hat{\theta}_{j[post]} - \hat{\theta}_{j[i\notin I_{E,j}]}. \quad (17)$$

This difference is distributed as

$$Diff_{\theta_j} \sim N\left(0, SE^2_{\theta_j[i\notin I_{E,j}]} + SE^2_{\theta_j[i\notin I_{E,j}]} - 2R_{\theta_j} SE_{\theta_j[i\notin I_{E,j}]} SE_{\theta_j[post]}\right), \quad (18)$$

where R_{θ_j} is the correlation between the two ability estimates for examinee j. Given a significance level, α, if $Diff_{\theta_j}$ is a positive extreme value in the above distribution, examinee j will be flagged.

R_{θ_j} can be accurately estimated only when examinee j retakes the test multiple times. In a situation where there is one test administration available, R_{θ_j} can be replaced with the correlation between the two ability estimates in group m that examinee j belongs to, denoted as R_{θ_m}. However, if tampering occurs within a group, R_{θ_m} is likely to be smaller than when tampering does not occur. The smaller that correlation is, the more conservative the test becomes. Further,

if the group size is small, the estimation of R_{θ_m} may not be accurate. Another substitute for R_{θ_j} could be a statistic from the distribution of R_{θ_m} over all groups, like the mean, mode, or a particular percentile.

2.10 Proposed Group Level Detection Methods

No matter which individual detection method is applied, if examinees can be flagged one by one, it is possible to calculate the percentages of flagged examinees within each group, and then to identify outliers among all groups. One method will be based on Tukey's definition (1977, p.43). In that method, observations will be taken as outliers which are at least 1.5 interquartile ranges (IQRs) greater than the third quartile.

Another group-level approach is based on use of mixed models where a group's effects on the values of individual detection indices are treated as random. In a large-scale exam, the numbers of classes, schools, and even districts are usually far larger than 20. This is generally considered too large and diverse for them to be regarded as uniform categories as in general linear models (Snijders & Bosker, 2012). Further, a testing program may have multiple administrations in a year, and there might be only part of a class, school, or district taking the same administration. In such a scenario, it would be appropriate to view the units at every level as samples from a corresponding population. In addition, in an effort to rule out alternative explanations to tampering and to increase the detection rate, it may be useful to control for covariates at different levels. Such covariates could include, for example, examinees' perceptions on answer changes (Prinsell et al., 1994; Skinner, 1983), gender (Al-Hamly & Coombe, 2005; Bath, 1967; Copeland, 1972; Geiger, 1991a), the use of certain new curricula in some classrooms, schools or district, and examinee sample sizes. For nested data, using a general

linear model could mistakenly assume that the effects of covariates at a macro level are the same as the effects on a micro ecological level (Robinson, 1950). A mixed model allows macro- and micro-effects to vary. It is also a less restrictive model to start with and should lead to more accurate estimates of standard errors for significance tests. Another benefit of mixed models is improved estimation. If group effects are treated as fixed as in general liner models, best linear unbiased estimators (BLUEs) will be used to obtain group means (Casella & Berger, 2001, p.544). This ignores the impact of the group size on the reliability of the estimator. If group effects are random, however, then best linear unbiased predictors (BLUPs will be used for their estimation (Robinson, 1991).

Considering a school in a district, there are two sources of information about the school, information from students and information from the district. BLUPs offer a better estimate of a school-level value by using a weighted average that combines information from the two sources. The empirical BLUP (EBLUP; Verbeke & Molenberghs, 2000) for the random effect of school k in district m would be,

$$u_{km} = a_{km}\mu_{km} + (1 - a_{km})\mu_m \tag{19}$$

where

a_{km}: a weighting factor for school k, where $a_{km} = \tau^2_{random}/(\tau^2_{random} + \sigma^2_{km})$
σ^2_{km}: the variance of the random factor for school k, calculated by σ^2_{error}/n_{km}
n_{km} = the examinee size of school k
τ^2_{random}: The variance of the random effects associated with the random factor
μ_m: The mean of the response values in district m
μ_{km}: The mean of the response values in school k

If the schools in the same district vary a lot (i.e., τ^2_{random} is large), more emphasis should be put on within-school information, whereas, if n_{km} is small, u_{km} will shrink towards the district mean which is more reliable.

In the current study, test data will be simulated with three levels: persons, schools, and districts. Each examinee's values on EDI, EDI_WTR, and $Diff_\theta$ can be used as the level-1 dependent variable, Y_{jkm}, in a three-level mixed model. The unconditional form of this model can be written as

Level 1: the person level

$$Y_{jkm} = \beta_{0km} + e_{jkm}, \tag{20}$$

Level 2: the school level

$$\beta_{0km} = \delta_{00m} + u_{0km}, \tag{21}$$

Level 3: the district level

$$\delta_{00m} = \gamma_{000} + V_{00m}, \tag{22}$$

Where,

$$\begin{bmatrix} u_{0km} \\ V_{00m} \end{bmatrix} \backsim N_2(0, G = \begin{bmatrix} \tau_1^2 & \tau_{12} \\ \tau_{21} & \tau_2^2 \end{bmatrix}), e_{zij} \backsim N(0, \sigma_e^2).$$

Y_{jkm} will be any of EDI, EDI_WTR, and $Diff_\theta$ from examinee j at school k of district m. γ_{000} is the fixed grand mean of Y_{jkm} i\ m,n all examinees. u_{0km} and V_{00m} are random effects of school and districts on Y_{jkm}, respectively, and they are assumed to follow a bivariate normal distribution with mean 0 and covariance matrix G. $e_{jkm}$is the individual residual and assumed to be normal with mean 0 and variance σ_e^2. If the BLUP of the random intercept u_{0km} or $V_{00m}$is significantly greater than 0, school k or district m will be flagged.

In the Tukey approach, examinees will be flagged by an individual-level detection at a given α-level. A school will be a suspect-school, if its within-school percentage of flagged examines is an outlier among all schools. That is, if the percentage of flagged examines for the school is at least 1.5 IQRs greater than the third quartile of all the schools in the sample. Tukey's

method makes no distributional assumptions nor does it depend on α as a criterion for evaluating Type I error rates. Previous research has shown that, for standard normal null distributions and sample sizes greater than 500, the right-tailed Type I error rate of Tukey's method can be less than .354% (Dawson, 2011; Seo, 2006). For positively skewed null distributions and sample sizes greater than 500, the right-tailed Type I error rates will be higher than that of a standard normal distribution. For example, in Seo's study(2006), if the right-tailed Type I error rate by Tukey's method is 1.512% in 500 samples from a lognormal distribution with mean 0 and variance .04, it would increase to 7.682%, if the variance is 1. Therefore, to create a criterion for the Type I error rate resulted from Tukey's method in the current study, it is necessary to know how big such a Type I error rate is, denoted as ω, in the null distribution, i.e., the distribution without tampering cases. This is then compared with the Type I error rate in the distribution with tampering cases. If the latter is close to ω (e.g., in the range of (.5 ω, 1.5 ω), it would be considered as acceptable and the test would be referred to as robust in validity (Zumbo & Jennings, 2002). Such a null distribution needs to be simulated for every instance in which this Tukey method is used. The within-school percentages of examinees various with the individual detection indices and α levels, so the null distributions of those percentages across all schools will vary with detection methods.

The within-school percentages of flagged examines will be also used as dependent variables in the general linear models where district is the predictor. When the omnibus F tests of the models are significant at $\alpha = .05$, the flagged districts will be the ones with significant coefficients at $\alpha = .05$.

2.11 Literature Summary

This chapter describes the major current approaches to detection of test tampering. This discussion is not exhaustive and there are other approaches that researchers and practitioners have been trying in the field of test security. Novel thinking and procedurals keep being presented in journals, and conferences (e.g. Belov, 2014; Cavalcanti et al., 2012; Clark III et al., 2013; Richmond, 2015). However, except for the analysis on marginal distributions of erasures, there is no widely accepted and applied approach yet in use. The lacking of training in interpreting the more complicated methods and their results is one reason (Maynes, 2013). Another import reason is probably the lack of comparative research evaluating multiple tampering detection methods across different settings. It is in this area that this dissertation aims to contribute. As Maynes (2013) pointed out, statistical detection of test security threats is still in its infancy, and a solid method should call for (1) clear reasons of why an anomaly merits an investigation, (2) greater understanding into the nature of the potential security beach, and (3) clear presentation of the ways of quantifying anomalies and the credible evidences that link statistical observation with the security risks.

CHAPTER 3

EMPIRICAL STUDY

In this chapter, individual and group tampering detection using marginal distributions of erasures (MD), the two-stage modeling method by Van der Linden and Jeon (VJ), EDI, EDI_WTR, and $Diff_\theta$ were compared using a real data set.

3.1 Empirical Data

An empirical data set from an administration of a statewide Grade 8 reading test administered in a large southeastern state in 2012 was used to motivate the modification of the tampering detection indices. This test was designed to measure the proficiency level for vocabulary, text identification, comprehension, analysis, critics, and application of literary selection and daily informational text. The data set contains the test responses of 35,280 examinees from 1,134 schools in 75 districts. Forty-five four-choice items were created for six passages. The initial and final responses to each item were recorded in the data set. The percentage of students who attained proficient status on this test was 56%.

The data set contained a symbol indicating uncertainty regarding whether an examinee had made multiple initial responses or uncertain final responses to each item. In this study, this was treated as a missing value and excluded, when counting erasures. Fewer than 125 examinees had such missing data and in no case did this occur for more than two items for any one examinee.

3.2 Methods

Individuals with more than one total or WR erasure were flagged if they had unusual values on MD, VJ, EDI with C=-.5, EDI_WTR with C=0, and $Diff_\theta$. For MD, those values were at least1.5 IQR (i.e., interquartile range) above the third quartiles of either total or WR erasures (Tukey, 1977). For VJ, EDI, EDI_WTR, and $Diff_\theta$, unusual cases were detected at α=.001, .01 and .05. When calculating the standard error of $Diff_\theta$, as defined in Equation 18, 242 schools had only one examinees and were not able to obtain R_{θ_m}, the within-school correlation between $\hat{\theta}_{j[post]}$ and $\hat{\theta}_{j[i\notin I_{E,j}]}$. These school's missing R_{θ_m}'s were replaced with a value estimated from other schools, denoted as $R_{\theta_{fix}}$. A smaller value of R_{θ_m} or $R_{\theta_{fix}}$would render a more conservative test for $Diff_\theta$. So, to control Type I error rates, the current study ranked all available R_{θ_m} and picked the first percentile, .96, as $R_{\theta_{fix}}$.

Beside the single-index individual detection, the occurrences of joint induvial flagging by the different methods were studied in this Chapter. Also, the relationships among $Diff_\theta$, EDI, EDI_WTR, different types of erasures, and examinees' estimated ability prior to erasing original responses were explored. To study the relationship between the pre-erasure ability estimates and different detection indices, pre-erasure ability estimates were ranked and divided into four groups of equal sizes. Since the study focus is to explore the group-index bivariate relationships, one-way analysis of variance (ANOVA) was applied to the data, instead of multivariate ANOVA (Huberty & Morris, 1989).

Schools became suspects if the within-school percentages of examinees flagged by MD, VJ, EDI, EDI_WTR, or $Diff_\theta$, separately, were at least1.5 IQR above the third quartiles in corresponding distributions. District was used in one-way analysis of variance (ANOVA) to predict those within-school percentages. If it turned out significant, a district that had unusually

high positive effects on those percentages would be singled out (i.e., flagged) as suspect at α=.05.

EDI, EDI_WTR, and $Diff_\theta$ were also entered into mixed models as level-1 dependent variables. Three-level mixed models (Equation 20, 21, and 22) were used to identify potential tampering districts. Preliminary analysis for this study showed that the three-level models could miss some schools having unusually high positive effect on EDI, EDI_WTR, and $Diff_\theta$; hence school level detection were performed in the following two-level models.

Level 1: the person level

$$Y_{jkm} = \beta_{0km} + e_{jkm}, \tag{23}$$

Level 2: the school level

$$\beta_{0km} = \delta_{000} + u_{0km}, \tag{24}$$

Where,

$$u_{0km} \backsim N_2(0, \tau_1^2),\ e_{zij} \backsim N(0, \sigma_e^2).$$

Y_{jkm} will be any of EDI, EDI_WTR, and $Diff_\theta$ from examinee *j* at school k of district m. $e_{jkm}$is the individual residual and assumed to be normal with mean 0 and variance σ_e^2. δ_{000} is the fixed grand mean of Y_{jkm} for all examinees. u_{0km} is the random effect of school k on Y_{jkm} and assumed to be normal with mean 0 and variance τ_1^2. At α=.05, and .01, schools and districts were flagged if they had significant positive random effects on EDI, EDI_WTR, or $Diff_\theta$.

3.3 Software

The R package "arm" (Gelman & Su, 2015) was used for calculating the conditional probabilities in VJ, the van der Linden and Jeon method. Estimation of item and person

parameters for all item response models was done using MULTILOG 7.0 (Thissen, 2003). SAS 9.3 software (SAS Institute, Cary NC) was used for the remainder of the analyses.

3.4.1 Results of Single-Index Individual Detection

The frequencies of different types of erasures are reported in Table 1.1 and 1.2, along with their medians, and robust measure of scales (i.e., IQRs and Q_n) . As can be seen in Table 1.1, answer changing was prevalent in the data set, where 91.53% of the 35,280 examinees had at least one erasure and 2.59% of them made at least five erasures. The majority of changes were WR (i.e., wrong to right changes) rather than other types of erasures: 57.26% of the total examinees had at least one WR erasures, and .44% of them made more than at least five WR erasures.

The proportions of total erasures and WR erasures above Tukey's 1.5 IQR index were 1,833 (5.2%) and 1,187 (3.3%), respectively. Of the 10,364 examinees with more than one erasure, EDI flagged 0, 2, 11, and 61 cases at $\alpha = .0001, .001, .01, .05$, respectively. Of the 4,437 examinees with more than one WR erasure, EDI_WTR flagged 14, 44, 185, and 719, respectively, at these same α levels and VJ flagged 30, 93, 289, and 818 cases. At $\alpha = .0001, .001, .01, .05$, $Diff_\theta$ identified 579, 1273, 1626, and 2470 examinees, respectively, among examinees with more than one erasure. Those numbers decreased to 558, 918, 1527, and 2264, respectively, for examinees with more than one WR erasure.

MD of total erasures and $Diff_\theta$ flagged between 5% and 7% of all examinees. These were the largest numbers of individuals so flagged as suspicious. Overall, the remaining methods, after controlling for item characteristics and examinee's ability, flagged somewhat lower numbers, ranging from approximately between 0% and 3%. EDI resulted in the smallest

Table 1.1: *Frequencies of Different Types of Erasures in the Real Data Set*

		Erasure Types							
		WR		WW		RW		Total Erasures	
		Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
Erasure Counts	1	15,765	44.69	8,598	24.37	9,030	25.60	21,927	62.15
	2	3,250	9.21	1,095	3.10	1,055	2.99	6,364	18.04
	3	772	2.19	235	.67	208	.59	2,167	6.14
	4	256	.73	60	.17	48	.14	921	2.61
	5	83	.24	28	.08	16	.05	415	1.18
	6	46	.13	4	.01	6	.02	224	.63
	7-15	30	.07	10	.03	2	.01	273	.78
	Total	20,202	57.26	10,030	28.43	10,365	29.38	32,910	91.53

Table 1.2: *Descriptive Statistics of Different Types of Erasures in the Real Data Set*

	Erasure Types			
	WR	WW	RW	Total Erasures
Median	1	0	0	1
IQR	1	1	1	1
Qn	2.22	2.22	0	2.22

number of flagged cases across all α levels. Numbers of flagged erasures by EDI_WTR's were the second smallest at each α level.

3.4.2 Results of Multi-index Individual Detection and Interrelationships among Different Indices

The occurrences of joint flagging by the different indices were studied in this section. Table 2.1 shows the overlap between MD of total erasures and every alternative method (VJ, EDI, $Diff_\theta$). Table 2.2 shows the overlap between MD of WR erasures and VJ, EDI_WTR, and $Diff_\theta$. All overlapping increased as α decreased.

VJ seemed to have the greatest agreement with MD of total erasures. The majority of examinees flagged by VJ also had large numbers of total and WR erasures. At $\alpha = .05$, 73.72% of the examinees identified by VJ had the numbers of total erasures at least 1.5 IQR above the third quartile. Considering only WR erasures, the agreement between VJ and MD increased to 100% at all three α levels. However, there were cases, for example, with 3 WR erasures and 5 total erasures, which all method flagged while VJ didn't. When an examinee had more than 3 WR erasures, he or she were always flagged out by VJ at $\alpha = .05$. At $\alpha = .01$, VJ never miss an examinee with more than 5 WR erasures. When the number of WR erasures increased above 6, VJ at $\alpha = .001$ selected all such cases as suspects.

EDI and EDI_WTR had less overlap of detections with MD. At $\alpha = .05$, .01, and .001, 44.26%, 54.55%, and 100% of the cases detected by EDI, respectively, had total erasures at or

above 1.5 IQR. For EDI_WTR, 48.96%, 65.41%, and 77.27% of flagged examinees had WR erasures at least 1.5 IQR above the third quartile for all three α levels. The disagreement with MD existed on examinees with small numbers of erasures, and also on the ones with large numbers of erasures. For example, both EDI and EDI_WTR flagged some examinees having only 2 WR erasures while MD and VJ didn't. At $\alpha = .05$ EDI ignored a case having 15 WR erasures and 20 total erasures while all other methods chose, and EDI_WTR ignored some cases with 7 to 9 WR erasures while MD, VJ, and $Diff_\theta$ didn't.

Table 2.1: *Percentages of Flagged Cases for Examinees with More Than One Erasure*

	Outliers on total erasures by 1.5 IQR
EDI $\boldsymbol{\alpha}$ =.001	100.00%
EDI $\boldsymbol{\alpha}$ =.01	54.55%
EDI $\boldsymbol{\alpha}$ =.05	44.26%
VJ $\boldsymbol{\alpha}$ =.001	100.00%
VJ $\boldsymbol{\alpha}$ =.01	100.00%
VJ $\boldsymbol{\alpha}$ =.05	73.72%
$Diff_\theta$ $\boldsymbol{\alpha}$ =.001	28.00%
$Diff_\theta$ $\boldsymbol{\alpha}$ =.01	20.48%
$Diff_\theta$ $\boldsymbol{\alpha}$ =.05	17.57%

Table 2.2: *Percentages of Flagged Cases for Examinees with More Than One WR Erasure*

	Being outliers on WR erasures by 1.5 IQR
EDI_WTR $\boldsymbol{\alpha}$ =.001	77.27%
EDI_WTR $\boldsymbol{\alpha}$ =.01	65.41%
EDI_WTR $\boldsymbol{\alpha}$ =.05	48.96%
VJ $\boldsymbol{\alpha}$ =.001	100.00%
VJ $\boldsymbol{\alpha}$ =.01	100.00%
VJ $\boldsymbol{\alpha}$ =.05	100.00%
$Diff_\theta$ $\boldsymbol{\alpha}$ =.001	45.32%
$Diff_\theta$ $\boldsymbol{\alpha}$ =.01	36.15%
$Diff_\theta$ $\boldsymbol{\alpha}$ =.05	30.43%

Like EDI and EDI_WTR, $Diff_\theta$ also appeared to ignore some cases with 6 even 15 WR erasures and selected many more cases with 2 WR erasures, while MD and VJ took opposite action, resulting in a much smaller overlap between $Diff_\theta$ and MD. Compare with VJ and EDI, The agreement percentages of $Diff_\theta$ with MD of total erasures were the smallest, and they were 17.57%, 20.48%, and 28% separately at At $\alpha = .05$, .01, and .001. The agreement between $Diff_\theta$ and MD of WR erasures increased a bit, and were 30.43%, 36.15%, and 45.32% for all three α levels, which are still the smallest, compared with those of VJ and EDI_WTR.

Correlations between detection indices, numbers of different types of erasures, and pre-erasure ability estimates are shown in Tables 3.1 and 3.2. All correlation coefficients were significant at $\alpha = .05$, p <.0001.

Using Cohen's (1988) guideline for interpreting Pearson correlations, EDI was strongly related to WR erasures (r_{EDI_WR}= .639) but had a moderate negative relationship with RW erasures (r_{EDI_RW}= -.454), and a small negative relationship with WW erasures (r_{EDI_WW}= -.264). As a result, the overall linear relationship between EDI and total numbers of erasures barely existed (r_{EDI_RW}=-.072).

EDI_WTR had a small linear relationship with RW erasures (r_{EDIwtr_Rw}=.237), and moderately correlated with WR (r_{EDIwtr_WR}=.375) and WW erasures (r_{EDIwtr_WW}=.364). As a result, the linear relationship between EDI_WTR and total numbers of erasures almost fell on the strong side (r_{EDIwtr_T}=.471).

Table 3.1: *Pearson Correlations between EDI, $Diff_\theta$, Pre-Erasure Estimated Ability and Different Types of Erasures among Examinees with at Least One Erasure, N=10,364*

	EDI	Diff	$\hat{\theta}_{[i \notin I_E]}$	WR	WW	RW	Total Erasures
EDI	1	.859	-.237	.639	-.263	-.454	-.072
Diff	-	1.00	-.196	.487	-.310	-.501	-.101
$\hat{\theta}_{[i \notin I_E]}$	-	-	1.00	.184	-.356	-.095	-.133
WR	-	-	-	1.00	-.180	-.168	.566
WW	-	-	-	-	1.00	-.019	.493
RW	-	-	-	-	-	1.00	.451

Table 3.2: *Pearson Correlations between EDI_WTR, $Diff_\theta$, Pre-Erasure Estimated Ability and Different Types of Erasures among Examinees with at Least One WR Erasure, N=4437*

	EDI_WTR	Diff	$\hat{\theta}_{[i \notin I_E]}$	WR	WW	RW	Total Erasures
EDI_WTR	1.000	.266	-.743	.375	.364	.240	.471
Diff	-	1.000	-.140	.171	-.334	-.524	-.302
$\hat{\theta}_{[i \notin I_E]}$	-	-	1.000	-.031	-.325	-.184	-.252
WR	-	-	-	1.000	.211	-191	.697
WW	-	-	-	-	1.000	.306	.719
RW	-	-	-	-	-	1.000	.693

The Pearson coefficient between EDI and $Diff_\theta$ was very high, r_{EDI_Diff}= .859. The scatterplot of this relationship is shown in Figure 3 for examinees having more than one total erasure. The colors of points in the scatterplot reflect whether a case was an outlier based on MD of total erasures; that is, whether an examinee's total erasures were at least 1.5 IQR above the third quartile. When EDI was greater than 0, outliers based on MD of total erasures tended to have larger values of $Diff_\theta$ than non-outliers. The marginal histograms of EDI and $Diff_\theta$ are also reported in Figure 3. These were negatively skewed due to the inclusion of WW and RW erasures.

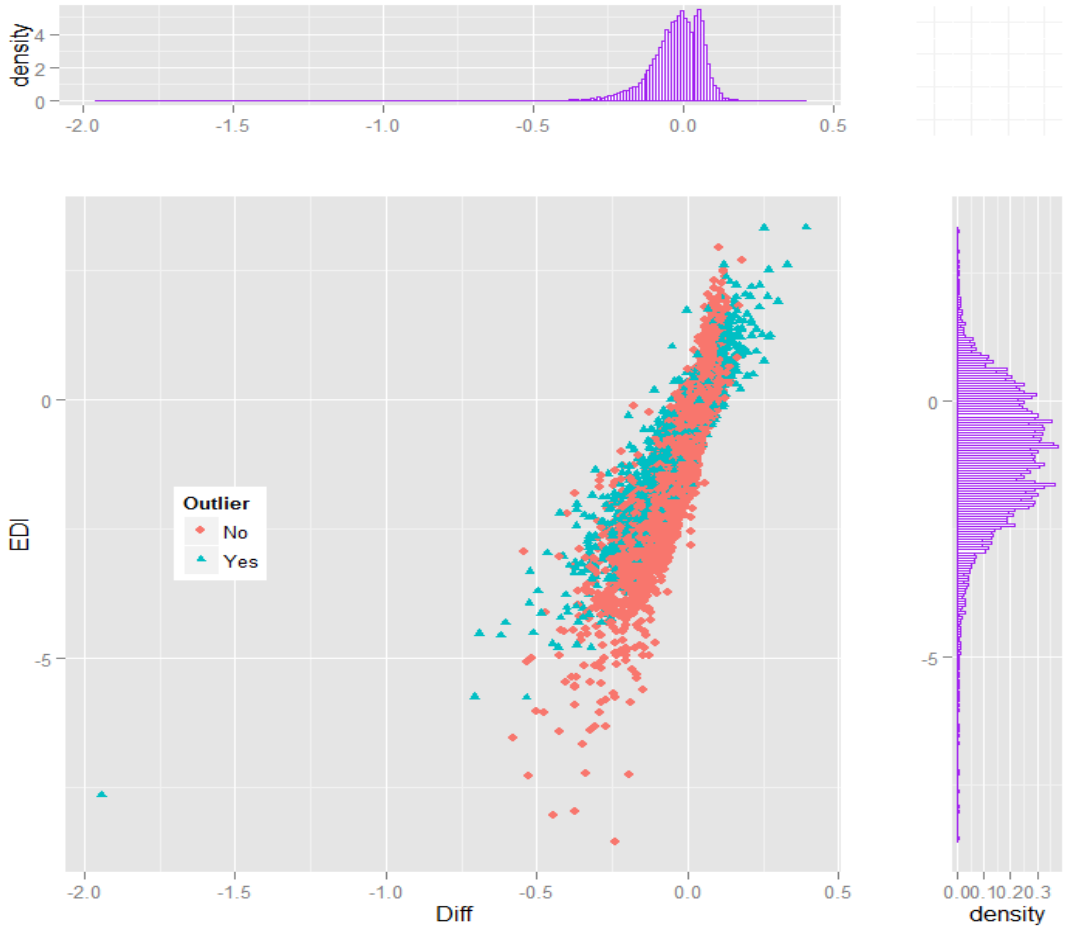

Figure 3: The scatter plot of EDI and $Diff_\theta$ with their marginal histograms from empirical examinees with more than one erasure in the real data set

The correlation between EDI_WTR and $Diff_\theta$ was small, r_{EDIwtr_Diff}=.266. A correlation of this magnitude tends not to show much of a discernible linear trend. This is evident in the scatterplot (see Figure 4) for examinees having more than one WR erasure. This was expected, as EDI_WTR was affected by only WR erasures while $Diff_\theta$ included all three types of erasures. When EDI_WTR was greater than 0, outliers based on MD of WR erasures tended to

have larger values of $Diff_\theta$ than non-outliers. $Diff_\theta$'s marginal histogram was more symmetric for examinees having more than one WR erasure. Due to the exclusion of RW and WW erasures, EDI_WTR resulted in more extremely positive values than EDI. There were 333 examinees with EDI_WTR values greater than 2, while only 24 examinees had EDI values greater than 2. As a result, EDI_WTR created more flags than EDI.

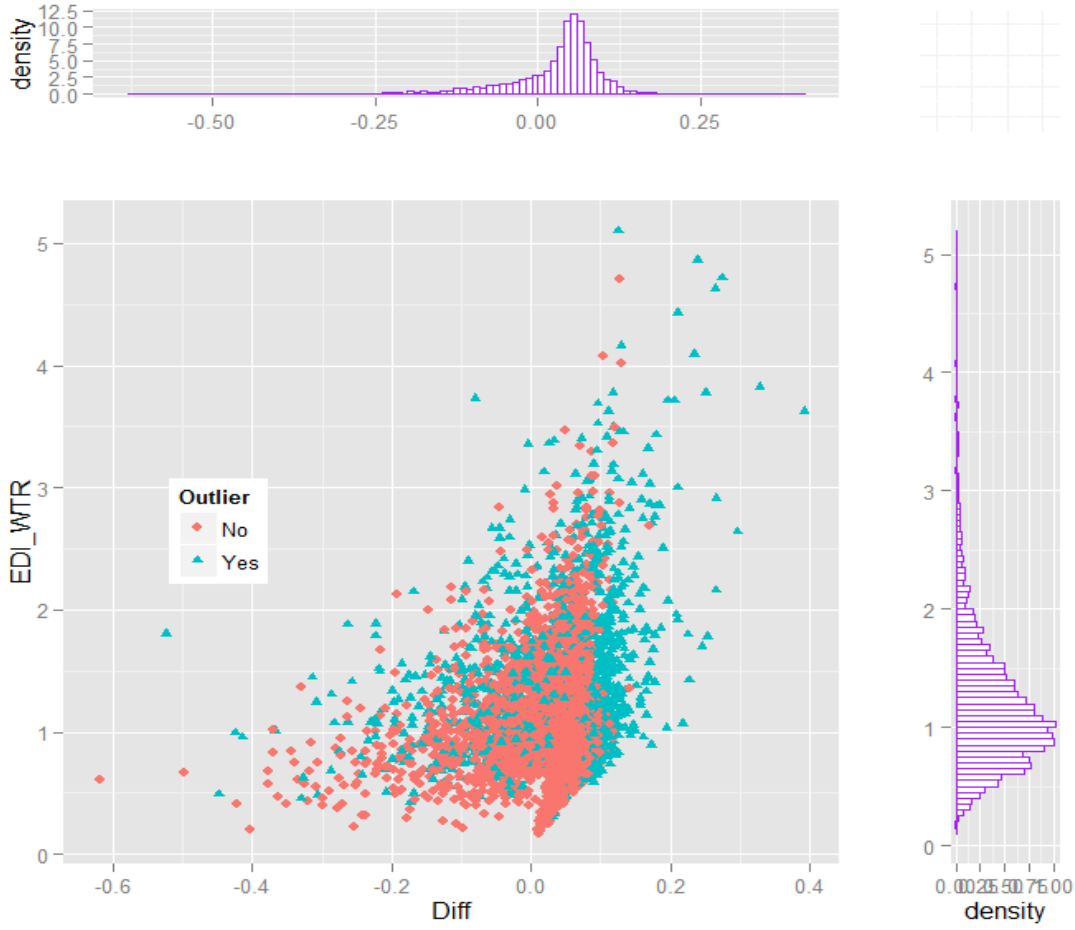

Figure 4: The scatter plot of EDI_WTR and $Diff_\theta$ with their marginal histograms from empirical examinees with more than one WR erasure in the real data set

One-way ANOVA was done between the groups of pre-erasure ability estimates and each of EDI, EDI_WTR, $Diff_\theta$, and the numbers of total and WR erasures. The omnibus F-tests in all ANOVAs were significant (see Table 4.1 to 4.5), $F_{EDI}(3, 10360) = 183.13$, $p_{EDI} < .0001$, $F_{Diff}(3, 10360) = 119.65$, $p_{Diff} < .0001$, $F_{EDI\_WTR}(3, 4430) = 1491.86$, $p_{EDI\_WTR} < .0001$, $F_{total\_era}(3, 35276) = 87.37$, $p_{total\_era} < .0001$, and $F_{wr\_era}(3, 35276) = 151.61$, $p_{wr\_era} < .0001$. The different results for each of the indices indicate that not all pre-erasure ability groups had the same means on each index.

Table 4.1: *ANOVA Table for EDI of Pre-Erasure Ability Groups*

Source	***df***	*Mean Square*	*F*	***p***	$\boldsymbol{r^2}$
Ability Groups	3	258.147	183.13	<.0001	.050
Error	10360	1.410			
Corrected Total	10363				

Table 4.2: *ANOVA Table for $Diff_\theta$ of Pre-Erasure Ability Groups*

Source	***df***	***Mean Square***	***F***	***p***	$\boldsymbol{r^2}$
Ability Groups	3	.964	119.65	<.0001	.033
Error	10360	.008			
Corrected Total	10363				

Table 4.3: *ANOVA Table for EDI_WTR of Pre-Erasure Ability Groups*

Source	***df***	***Mean Square***	***F***	***p***	$\boldsymbol{r^2}$
Ability Groups	3	229.713	1491.86	<.0001	.050
Error	4433	.154			
Corrected Total	4436				

Table 4.4: *ANOVA Table for Total Erasures of Pre-Erasure Ability Groups*

Source	***df***	***Mean Square***	***F***	***p***	r^2
Ability Groups	3	123.41	87.37	<.0001	.007
Error	35276	1.41			
Corrected Total	35279				

Table 4.5: *ANOVA Table for WR Erasures of Pre-Erasure Ability Groups*

Source	***df***	***Mean Square***	***F***	***p***	r^2
Ability Groups	3	109.462	151.61	<.0001	.013
Error	35276	.722			
Corrected Total	35279				

Boxplots by ability group for EDI, EDI_WTR, and $Diff_\theta$ are presented in Figures 5.1 to 5.3. The means of each index increased as the pre-erasure ability level decreased. On average the lowest group of pre-erasure ability estimates (labeled as Q1 in the boxplots in Figures 5.1) had an EDI of -.6923. This was .7369 greater than that of the highest group (labeled as Q4), $t(10360)= 22.15$, $p<.0001$. EDI_WTR in Q1 had a mean of 1.8716 (see Figure 5.2), which was 1.10 greater than that of Q4, $t(4433)=64.63$, $p<.0001$. Q1's mean on $Diff_\theta$ was -0.0035 (see Figure 5.3), .74 larger than that of Q4 the highest group. Across the three methods, the lowest group of pre-erasure ability estimates had more examinees with larger values on each of the detection indices.

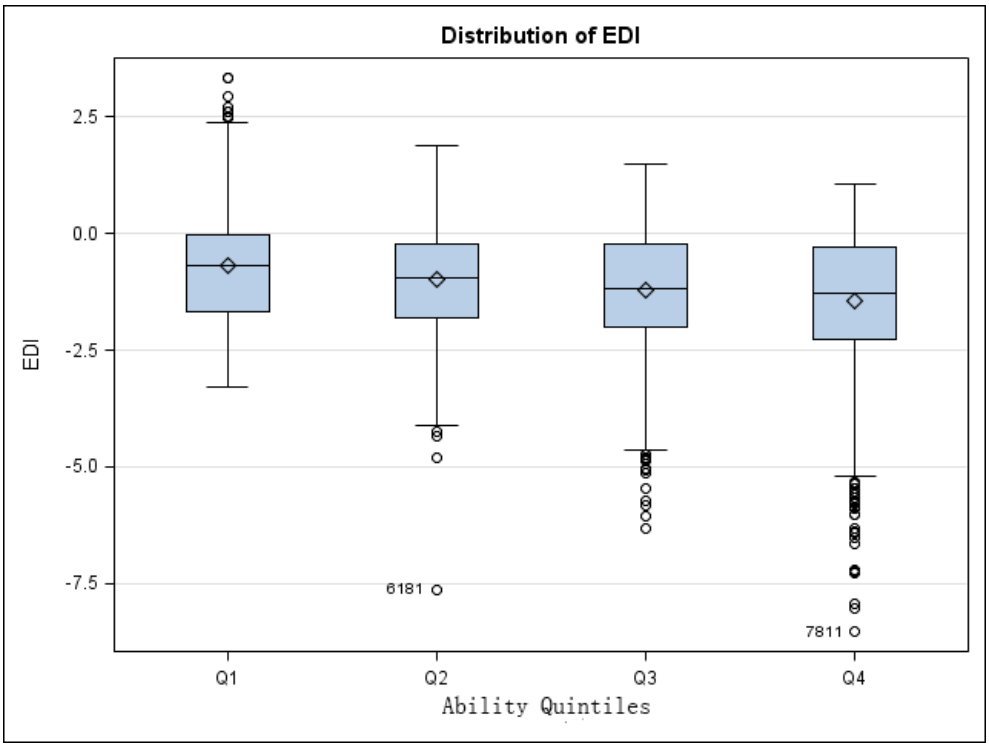

Figure 5.1: The boxplot of EDI by pre-erasure ability groups

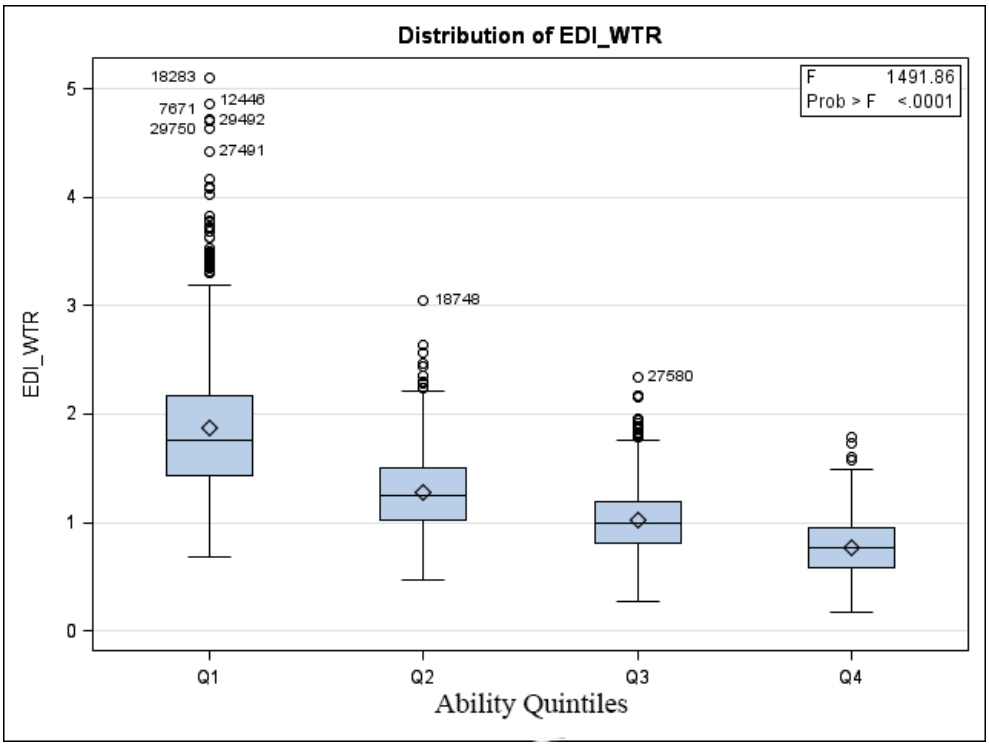

Figure 5.2: The boxplot of EDI_WTR by pre-erasure ability groups

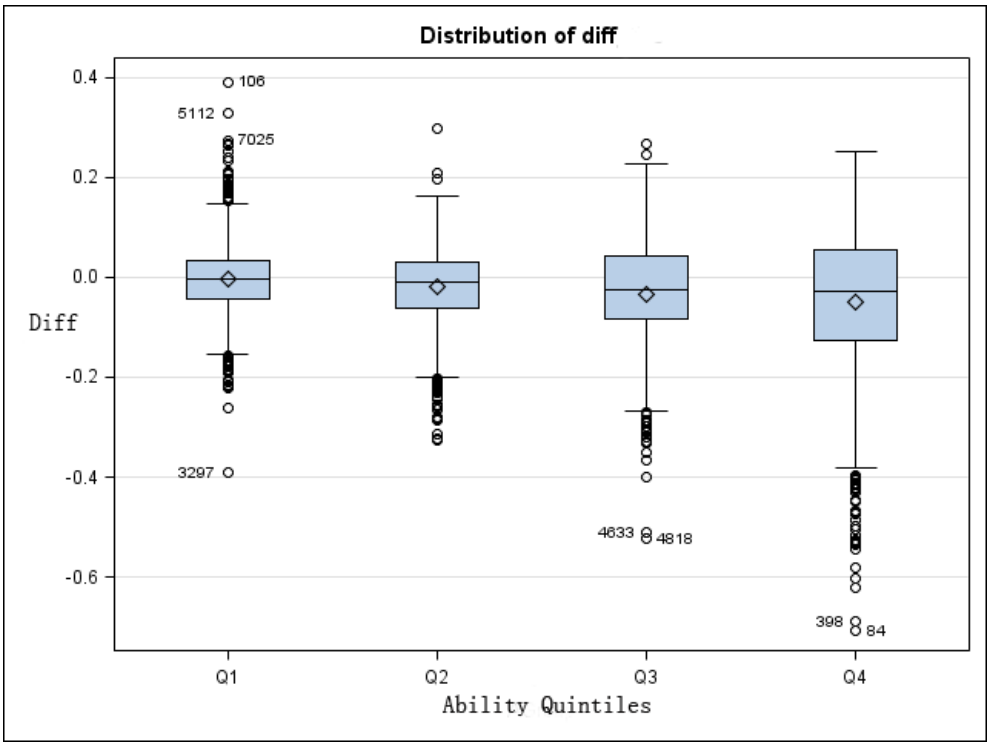

Figure 5.3: The boxplot of $Diff_\theta$ by pre-erasure ability groups

Contingency tables of numbers of total and WR erasures by ability groups are reported in Tables 5.1 and 5.2 rather than boxplots, because the latter were less informative. The highly skewed distributions of the erasures would compress the boxes too much to make group differences obvious. Further, because of such skewness, the differences on the group means of total and WR erasures were less than 1, and thus could not reflect the true group distinction.

Table 5.1: *Contingency Table of Total Erasures and Pre-Erasure Ability Groups*

		Pre-erasure ability groups				
		Group 1	Group 2	Group 3	Group 4	Total
Total Erasures	0	748	730	719	792	2989
	1-3	7419	7611	7706	7722	30458
	4-6	525	411	342	282	1560
	7-29	137	70	48	18	273
	Total	8829	8822	8815	8814	35280

Table 5.2: *Contingency Table of WR Erasures and Pre-Erasure Ability Groups*

		Pre-erasure ability groups				
		Group 1	Group 2	Group 3	Group 4	Total
WR Erasures	0	4748	3952	3471	2907	15078
	1-3	3984	4774	5229	5800	19787
	4-6	85	90	108	102	385
	7-15	12	6	7	5	30
	Total	8829	8822	8815	8814	35280

Chi-square tests were conducted to analyze group differences in these contingency tables. Results of the analyses indicated more examinees in the lower pre-erasure ability groups made larger amounts (≥4) of total erasures. A $\chi^2_{Total}(9, N = 35{,}280) = 207.68$, $p_{Total} < .0001$, supported existence of group differences. Overall, the number of examinees with no WR erasures decreased as ability levels increased. This could imply that high achievers were more able to correct their mistake, but it also may be evidence that illegitimate WR erasures boosted scores.

The largest proportion, 40%, of examinees making more than 6 WR erasures was from the lowest pre-erasure ability group. The chi-square test indicated that there was group difference on the numbers of WR erasures, $\chi^2_{WR}(9, N = 35280) = 847.66$, $p_{WR} < .0001$. This result agrees with what might be expected when tampering has occurred; that is, individuals with lower ability would be expected to have more WR erasures, possibly suggesting that tampering was present in these responses. At the very least, those individuals could be considered likely suspects on which to follow up.

Chi-square tests were also conducted between the pre-erasure ability groups and the flagged status by VJ at all three α levels (see Table 6). These were $\chi^2_{VJ001}(3, N = 35280) = 4.72$, $P_{VJ001} = .1938$, $\chi^2_{VJ01}(3, N = 35280) = 20.67$, $P_{VJ01} < .0001$, and $\chi^2_{VJ05}(3, N = 35280) =$

163.44, P_{VJ05} < .0001. At each α level, the higher ability groups had more cases flagged by VJ. In part, that can be attributed to the larger numbers of examinees having at least one WR erasures in the higher ability groups.

Table 6: *Contingency Table of VJ and Pre-Erasure Ability Groups*

		Pre-erasure ability groups				
		Group 1	Group 2	Group 3	Group 4	Row Total
VJ	0	8814	8879	8879	8785	35187
$\boldsymbol{\alpha}$ =.001	1	15	23	26	29	93
VJ	0	8782	8753	8743	8713	34991
$\boldsymbol{\alpha}$ =.01	1	47	69	72	101	289
VJ	0	8726	8681	8576	8479	34462
$\boldsymbol{\alpha}$ =.05	1	103	141	239	335	818

3.4.3 Results of Group (i.e., School-Level and District-Level) Detection

At the school level, when looking at their percentages of flagged examinees at least 1.5 IQR above corresponding third quartiles, EDI tended to identify fewer schools than EDI_WTR, VJ, or $Diff_\theta$ (see Table 7). In part, this was because EDI had the smallest number of individual detections.

Table 7: *Numbers of School Outliers and Percentages of Flagged Examinees within Schools*

	School outliers		**School outliers**
EDI $\boldsymbol{\alpha}$ =.001	2	**EDI_WTR $\boldsymbol{\alpha}$ =.001**	40
EDI $\boldsymbol{\alpha}$ =.01	11	**EDI_WTR $\boldsymbol{\alpha}$ =.001**	141
EDI $\boldsymbol{\alpha}$ =.05	56	**EDI_WTR $\boldsymbol{\alpha}$ =.05**	100
Diff $\boldsymbol{\alpha}$ =.001	80	**VJ $\boldsymbol{\alpha}$ =.001**	86
Diff $\boldsymbol{\alpha}$ =.01	67	**VJ $\boldsymbol{\alpha}$ =.01**	204
Diff $\boldsymbol{\alpha}$ =.05	30	**VJ $\boldsymbol{\alpha}$ =.05**	69

Although $Diff_\theta$ at α = .01 and .05 flagged more examinees than VJ and EDI_WTR, the numbers of schools flagged by $Diff_\theta$ were smaller than from either of the other two methods.

This might be due in part to the fact that the percentages flagged by $Diff_\theta$ were relatively large and their IQRs were 11.36% and 20.91% at $\alpha = .01$ and .05, respectively. For example, a school had 23% of examinees flagged by $Diff_\theta$ and only 7.69% flagged by VJ at $\alpha = .05$. Twenty-three percent was still not an extreme value for $Diff_\theta$, while 7.69% was already at least 1.5 IQR above the third quartile for VJ.

Only one school was flagged simultaneously by at least three of the methods. In that school, there were two out of nine examinees flagged, one of whom made 13 WR erasures on the 45 items and the other who made four erasures all of which were from wrong to right.

District was not a significant predictor of any percentage of flagged examinees within schools, except for the percentage of examinees flagged by VJ at $\alpha = .05$ (see Table 8). This was likely a false alarm, because only one district had an unusually large mean of within-school flagged percentages, but that mean did not reflect the majority. In that district, there are three schools and two flagged examinees. The schools had examinee sizes of 15, 5, and 1 and their flagged percentages were 0%, 20%, and 100%, respectively. So, the district mean of within-school flagged percentages was biasedly lifted to a large value due to the school with one examinee but 100%. This suggests that any modeling that would aggregate the flagged percentages from schools of very different sizes is probably not a good idea.

Table 8: *ANOVA Table for the Flagged Percentages by VJ*

Source	***df***	***Mean Square***	***F***	***p***	r^2
District	73	.0082	4.79	<.0001	.2479
Error	1060	.0017			
Corrected Total	1122				

Two-level mixed models were applied to EDI, EDI_WTR, and $Diff_\theta$ (Equations 23 and 24) to study school effects. In these models, schools with less than three examinees were excluded to obtain more reliable estimates. For the same reason, when studying district effects in three level mixed models (Equations 20, 21, and 22), schools with less than three examinees and districts with less than three schools were excluded.

For examinees having at least one erasure, neither results for the three-level nor the two-level mixed models for EDI were significant at $\alpha = .05$ (see covariance parameter estimates in Tables 9.1 and 9.2), where $Z_{District}(51, 9844) = 1.48$, $p_{District} = .0694$, $Z_{School} = 1.33$, $p_{School} = .0919$. This means that the variances of random effects of districts and schools were not different from 0, hence there was no random effect. Therefore, the three-level model was reduced to an ANOVA of districts (see Table 9.3), and the two-level mixed model was reduced to an ANOVA of schools (see Table 9.4).

Table 9.1: *Covariance Parameter Estimates Table for EDI in Three-Level Models*

Covariance Parameter	**Subject**	**Estimate (SE)**	***Z***	***p***
Intercept	District	.0033 (.0022)	1.48	.0694
Intercept	School	.0052 (.0052)	1.00	.1576
Residual		1.4814 (.0216)	68.59	<.0001

Table 9.2: *Covariance Parameter Estimates Table for EDI in Two-Level Models*

Covariance Parameter	**Subject**	**Estimate (SE)**	***Z***	***p***
Intercept	School	.0071 (.0052)	1.33	.0919
Residual		1.4816(.0214)	69.09	<.0001

Table 9.3: *ANOVA Table for EDI by District*

Source	***df***	***Mean Square***	***F***	***p***	r^2
District	51	2.0993	1.41	<.0283	.0073
Error	9844	1.4864			
Corrected Total	9895				

Table 9.4: *ANOVA Table for EDI by School*

Source	***df***	***Mean Square***	***F***	***p***	r^2
School	723	1.6399	1.11	<.0249	.0792
Error	9340	1.4769			
Corrected Total	10063				

The ANOVA models for EDI (see Table 9.3) showed significant fixed effects for districts and schools, where $F_{District}(51, 9844) = 1.41$, $p_{District} = .0283$ and $F_{School}(723, 9340) = 1.11$, $p_{School} = .0249$. This indicates some districts and some schools had different mean EDIs. Further, there were significant negative effects for all flagged schools and districts on EDI at $\alpha = .05$. Since only WR erasures would lead to a positive increase on EDI, a negative effect on EDI could mean that either these schools and districts didn't tamper with the test, or their tampering didn't create enough WR erasures to make their means for EDI higher than the rest of the schools and districts.

Mixed models didn't work for $Diff_\theta$ either (see Tables 10.1 to 10.3). The three-level models in Table 10.1 seemed to have random effects of districts ($Z_{3\_District} = 1.67$, $p_{3\_District} = .0473$), but it appeared to occur with non-significant random effects of schools ($Z_{3\_School} = .31$, $p_{3\_School} = .3780$). This means that schools should be treated as fixed effects in the three-level models. In other words, that u_{0km} in Equation 21 should be removed. The new three-level models were reported in Table 10.3.1 and 10.3.2, which still didn't show random effects of

districts ($Z_{District}$= 0.00, $p_{District}$ = .5). So, the three-level models reduced to ANOVA of districts. The two-level models in Table 10.2 also didn't support random effects of schools with Z_{2_School}= .31, and p_{2_School} = .3780. The three-level models reduced to ANOVA of schools.

Table 10.1: *Covariance Parameter Estimates Table for $Diff_\theta$ in Three-Level Models*

Covariance Parameter	**Subject**	**Estimate (SE)**	***Z***	***p***
Intercept	District	.000025 (.000015)	1.67	.0473
Intercept	School	.000009(.000029)	.31	.3780
Residual		1.4814 (.0216)	68.41	<.0001

Table 10.2: *Covariance Parameter Estimates Table for $Diff_\theta$ in Two-Level Models*

Covariance Parameter	**Subject**	**Estimate (SE)**	***Z***	***p***
Intercept	School	.0071 (.0052)	1.33	.0919
Residual		1.4816(.0214)	69.09	<.0001

Table 10.3.1: *Covariance Parameter Estimates Table for $Diff_\theta$ in Three-Level Models with Fixed School Effects*

Covariance Parameter	**Subject**	**Estimate (SE)**	***Z***	***p***
Intercept	District	.000001 (145.35)	.00	.0919
Residual		.008299(.00012)	67.79	<.0001

Table 10.3.2: Type 3 Tests of Fixed Effects of Schools *for $Diff_\theta$ in Three-Level Models with Fixed School Effects*

Effect	**Numerator df**	**Denominator df**	***F***	***p***
School	703	9132	1.24	<.0001

In the ANOVA of districts (see Table 11.1) and ANOVA of schools (see Table 11.2), significant fixed effects of districts and schools were observed on $Diff_\theta$, where $F_{District}$(51, 9844) = 1.63, $P_{District}$ = .0031 and F_{School}(723, 9340) = 1.23, P_{School} < .0001. Similar to the

results of EDI, all flagged schools and districts had significant negative effects on $Diff_\theta$ at $\alpha = .05$, indicating that either these schools and districts didn't tamper with the test, or their tampering didn't result in enough score gain to make their means for $Diff_\theta$ higher than the rest of the schools and districts.

Table 11.1: *ANOVA Table for $Diff_\theta$ by District*

Source	***df***	***Mean Square***	***F***	***p***	r^2
District	51	.0137	1.63	.0031	.0084
Error	9844	.0084			
Corrected Total	9895				

Table 11.2: *ANOVA Table for $Diff_\theta$ by School*

Source	***df***	***Mean Square***	***F***	***p***	r^2
School	723	.0102	1.23	<.0001	.0870
Error	9340	.0083			
Corrected Total	10063				

Three-level mixed models analyses of EDI_WTR revealed significant school and district effects, Z_{School}= 5.66, p_{School} < .0001, and $Z_{District}$= 2.08, $p_{District}$ = .0186 (see also Table 12.1). These indicate that the variances of random effects of schools and districts were not 0, hence the random effects exist. The two-level model analysis of EDI_WTR also supported the presence of random effects of schools, where Z_{School}= 6.05, P_{School} < .0001(see Table 12.2).

Table 12.1: *Covariance Parameter Estimates Table for EDI_WTR in Three-Level Models*

Covariance Parameter	**Subject**	**Estimate (SE)**	***Z***	***p***
Intercept	District	.0054 (.0044)	2.08	.0186
Intercept	School	.0239 (.0063)	5.66	<.0001
Residual		.2794 (.0107)	41.70	<.0001

.1

Table 12.2: *Covariance Parameter Estimates Table for EDI_WTR in Two-Level Models*

Covariance Parameter	**Subject**	**Estimate (SE)**	***Z***	***p***
Intercept	School	.0283 (.0043)	6.05	<.0001
Residual		.2789 (.0066)	42.13	<.0001

EDI_WTR ‘s three-level mixed models flagged four districts and 20 schools, as the BLUPs of their random effects on EDI_WTR were significantly larger than 0. The two-level mixed models flagged 27 schools. Table 13 presents a comparison of flagged schools by the two models. The three-level models identified some districts and schools within the districts as including possible tampering. More schools from those districts were flagged in the two-level models, however, thus suggesting that the two-level model should be used to perform school-level detection.

Table 13: *Comparison between Part of Flagged Subjects by Model 1 and Model 2 for EDI_WTR*

Three-level models					Two-level models			
District	**School**	**Random intercept estimate (SE)**	**df**	**t (p>t)**	**School**	**Random intercept estimate (SE)**	**df**	**T (p>t)**
A3		.1186 (.0417)	3978	2.85 (.0022)				
A3	**A3641**	.3075 (.1587)	3978	1.94 (.0263)	**A3641**	.3830 (.1715)	4061	2.23 (.0127)
A3	**A3605**	.2669 (.1471)	3978	1.81 (.0349)	**A3605**	.3378 (.1567)	4061	2.15 (.0156)
					A3663	.2518 (.1333)	4061	1.89 (.0589)
					A3602	.2480 (.1146)	4061	2.16 (.0305)

3.5 Summary of the Empirical Study

This chapter presented a study using real data from a statewide testing program in which MD, VJ, EDI, EDI_WTR, and $Diff_\theta$ were used detect potential tampering of students' answers. The detection indices were used to detect this tampering for individual students, for schools and for districts.

The results suggested that answer changing was a common behavior in this data set. The major type of answer changing was from wrong to right (i.e., WR). The real dataset also showed that examines with higher ability were more likely to make WR erasures, however, much larger frequencies of WR erasures were observed among lower scoring students. These lower ability examinees tended to have higher values of EDI, EDI_WTR, and $Diff_\theta$.

MD of total erasures, $Diff_\theta$, and EDI all looked for tampering that leads to all three types of erasures. The first two indices had the largest numbers of individual flagging. For MD of the total erasures, one reason could be that the method does not exclude the erasures caused by randomness in test responses conditioned on the characteristics of items and persons. For the $Diff_\theta$ index, it is possibly overly sensitive and thus might flag a lot of examinees with fewer answer changes. EDI appeared to be the most conservative detection method, in part, due to its way of summing up three types of erasures.

The numbers of suspects detected by MD of WR erasures, VJ, and EDI_WTR were in the middle range. VJ, and EDI_WTR, which control for examinee ability and item characteristics, had fewer flagged cases. This appears to agree with results from Primoli, Liassou, Bishop, and Nhouyvanisvong (2011) that a large number of erasures could be the function of individual abilities.

$Diff_\theta$, EDI, and EDI_WTR ignored some examinees who had large numbers of WR erasures but who were flagged by VJ and MD. These three indices also flagged some examinees who had relatively small numbers of WR erasures but who were not flagged by VJ and MD.

Group-level detection using percentages of within-school flagged examinees didn't raise concerns about security breaches for any district. But at the district-level, averaging out the flagged percentages from schools of different sizes appeared to be problematic. At the school-level, the percentages of flagged examinees $Diff_\theta$ and EDI yield fewer outliers than the other methods.

When schools and districts were used to predict $Diff_\theta$ and EDI, EDI_WTR using either mixed models or ANOVA, only EDI_WTR identified some schools and districts that appeared to involve tampering. These appeared to be because they had sufficient numbers of WR erasures to be detected by EDI_WTR. This suggests that EDI_WTR in mixed models may be useful for detect tampering groups.

In the next chapter, all detection indices will be studied in simulated data sets with known distributions and known numbers of potentially fraudulent erasures.

CHAPTER 4

SIMULATION STUDY

The empirical study in the previous chapter presented cross comparison among multiple tampering detection tools. These tools included marginal distribution (MD) of total and WR erasures, the method by van der Linden and Jeon (VJ), Erasure Detection Index of total erasures (EDI), Erasure Detection Index of WR erasures (EDI_WTR), and the difference between the ability estimated from responses with erasures and ability estimated without erasure items ($Diff_\theta$). With each of these methods, there is the potential for Type I errors to intrude such that even though an individual or school is detected by multiple methods, it may not be either a true victim or a true violator. One step to determining the accuracy of each of these methods is to examine the Type I error rates and power in real data sets with known cases of tampering. Unfortunately, such data are not always readily available. A useful alternative is to develop simulations of different types of tampering and then study the Type I error rates and power for each of these detection methods. Therefore, in this chapter, multiple data sets with known tampered cases were created to simulate similar data to those in the real data example in the previous chapter. The simulated data sets differed in tampering strategies and the numbers of examinees' detected as having been tampered with were examined for MD, VJ, EDI, EDI_WTR, and $Diff_\theta$ for individual and group level detection.

4.1.1 Simulated Data

The estimated item and ability parameters of 45 four-choice items and 35,280 individuals from the real data set analyzed in Chapter 3 were used as the true item parameters and individual abilities in these simulations. The grouping structure of examinees was also basically unchanged to simulate the multilevel ability distribution of examinees in the real data example as these distributions differ by school and district. In the real data set, some schools had less than 3 examinees, and some districts had less than 3 schools. In order to obtain larger sample sizes for schools and districts, the simulation merged some small schools in the same district and some small districts with closer ID numbers. The resulting simulated data set consisted of 905 schools in 54 districts. The numbers of schools within districts ranged from 3 to 125, and the number of examinees within schools ranged from 4 to 234.

All item and individual parameters were the ones estimated from the real data described in Chapter 3 without erased responses using nominal response models (NRMs; Bock, 1972). These parameter estimates were then fixed in the NRM was used to generate the initial item responses. The initial data set was labeled Set 0. As in Wollack et al. (2013), this simulation created erasures due to misalignment, speededness, random changes, fixed-number tampering, and score-based tampering. These were added to Set 0 to create the additional data simulated data sets to be analyzed in this study.

Wollack et al. (2013) first sorted 250,000 examinees by ability, and then chose 200 examinees from each quintile of the ability distribution to have misalignment erasures and another 200 examinees to have speededness erasures. In this study, 0.4% of examinees were randomly selected in each ability quintile to have misalignment erasures, and another 0.4% were

randomly selected to have speededness erasures. The remainder of the 35,000 examinees were simulated to have random erasures.

This study differed from Wollack et al. in that erasures due to random were not created by selecting one out of the remaining three answer choices with equal probabilities. Rather, this study created the randomness of answer changing as the result of random sampling based on the conditional probability given the item response model. That is, it was assumed that an examinee would be more likely to change an answer to a choice which had a relatively high conditional probability of being selected.

Three tampering strategies were simulated: tampering of 5 items tampering, tampering of 10 items, and score-based tampering. All tampering strategies created WR erasures. The extent of fixed-number tampering varied at two levels: 5 and 10 items. Since examinees were randomly chosen, there was the potential for some examinees to have fewer than 5 or 10 wrong answers. In such cases, the numbers of tampered items would be less than the designated level. Score-based tampering generally created enough WR erasures with a little extra cushion so that examinees whose true (i.e., simulated) abilities were lower than the proficiency level to pass the exam. The maximum number of score-based tampered items was chosen to be 10 in this simulation as having more than 10 WR erasures in a 45-item test would be quite rare. There was only .01% of examinees in the real date set with more than 10 erasures.

To evaluate the group detection performance of different indices, ten districts were randomly selected to have tampering schools. Three levels of school involvement in tampering were simulated: 25%, 50%, and 100% of schools in the ten districts. In each tampering schools, there were three levels of examinee involvement: 25%, 50%, and 100% of examinees.

To summarize, across the three factors, there were 3 strategies × 3 school levels × 3 examinee levels of tampering involvement = 27 conditions. In addition, a condition was simulated that included answer sheets that were free of erasures due to tampering yielding a total of 28 simulation conditions. Each condition was replicated 10 times, resulting in 280 simulated data sets. Every data set also included erasures due to random changes, misalignment, and speededness (as described below). Ten data sets had no tampering, and 140 data sets had only fixed-number tampering. The remaining140 data sets had only score-based tampering

4.1.2 Set 1: Misalignment Erasure Simulation.

a.1 The following procedures were used for creating the misalignment erasures starting with the data simulated for Set 0. Randomly sample 0.4% (i.e., 28) of the examinees from each ability quintile.

a.2 As in Wollack et al. (2013), draw a random number from a binomial distribution with N = 45 items and an erasure probability = .25 for each selected examinee j. This number, denoted by L_j, is the number of misaligned items that the examinee legitimately erased.

a.3 In Set 0, for each chosen examinee j, randomly select an item to be the first misaligned item. Call this item M_j.

a.4 From Item M_j to item $M_j + L_j - 1$, change the response to each item to be the response for the next item.

4.1.3 Set 2: Speededness Erasure Simulation.

Wollack et al. (2013) selected the last one-fifth of the items on a 50-item test to simulate speededness erasures. Following their approach in the current study, speededness erasures were generated from Item 37 to Item 45 for the selected examinees.

b.1 Exclude the examinees simulated with misalignment erasures, and then randomly sample another 28 = 0.4% of the examinees from each quintile.

b.2 For each chosen examinee, beginning with Item 37, randomly generate a response such that the probability of selecting each answer is the reciprocal of the number of alternatives. In this example, it would be .25. The differences between the new responses and the ones in Set 0 become speededness erasures. This simulates an examinee who randomly fills in the last 9 items because of the assumed concern that he or she will run out of time.

4.1.4 Set 3: Random Erasures Simulation.

The distribution of the total numbers of simulated random erasures was designed to approximate the probabilities observed in the empirical data set (see Table 2.1), where Pr($X=0$) = .0847, Pr($X=1$) = .6215, Pr($X=2$) = .1804, and Pr($X=3$) =.0614. These were used as the probabilities of generating corresponding numbers of erasures.

c.1 Exclude the previously 280 examinees chosen to have misalignment and speededness and generate random numbers, R_j, from a uniform (0,1) distribution. Assign this number to each of the remaining examinees.

c.2 For every examinee with $R_j > 1 - \Pr(X = 0)$, randomly pick 29 items. This was the maximum number of total erasures in the real data set. These items were labeled as $ER_1, ER_2, \ldots, ER_{29}$.

c.3 Use the NRM to generate a new response to item ER_1. If $R_j > 1 - [\Pr(X = 0) + \Pr(X = 1)]$, use NRM to generate a new response to item ER_2. Additional new responses were generated in a similar way. That is, the NRM was used to generate a new response to item ER_k, where $k = \{3, ..., 29\}$, when $R_j > 1 - [\Pr(X = 0) + \Pr(X = 1) + \cdots + \Pr(X = (k-1))]$. The differences between a new response and the original one in Set 0 became the random erasures.

c.4 Check the frequencies of simulated total erasures. Make sure that they are close to the ones reported in Table 2.1. Since there is some randomness in the process of generating responses via NRM, it could not be guaranteed that a new response to item ER_k would differ from the old response. As a result, it was possible that there might be considerable differences in the frequencies between the simulated erasures and the real ones. Adjust those thresholds for R_j to decrease such possible differences

4.1.5 Simulation of Fixed-numbers of Fraudulent Erasures in Set 3.

In a tampering school, a percentage (25%, 50%, or 100%) of all examinees was chosen. For every chosen examinee in a tampering school, fixed numbers of fraudulent WR erasures were generated by the following steps:

d.1 Randomly select *W* items from all incorrectly answered items by this student in Set 3. *W* is the number of tampered items (*W*=5, 10). Replace responses to the *W* items with the correct ones.

d.2 If the total number of incorrectly answered items is less than *W*, change all to correct.

4.1.6 Simulation of Score-based Fraudulent Erasures in Set 3.

Based on the responses in Set 3, data sets with score-based fraudulent WR erasures were generated by the following steps:

e.1 Randomly choose 25%, 50%, or 100% of examinees who did get 30 items right, which are the minimum correct answers required for passing the current exam. For every selected examinee j, compute the total number of correctly answered items, Y_j.

e.2 Randomly select a random integer number Z_j^* from 0 to 5. This will identify the cushion, that is, the number of erasures beyond the proficient threshold that someone tampering with the answer sheet would erase, just to make sure the student's score comfortably met the passing standard.

e.3 $X_j = (30 - Y_j) + Z_j^*$, where X_j is the number of fraudulent WR erasures and

e.4 Randomly select X_j items from all incorrectly answered items by this student in Set 3. Replace response to the selected items with the correct answers.

e.5 If the total number of incorrectly answered items is less than X_j, make all of them correct.

e.6 If X_j is greater than 10, only correct 10 items.

4.2 Methods

At each of the three alpha levels, α=.05, .01, and 001, individuals were simulated to have tampered answer sheets by their values on MD of WR erasures, VJ, EDI (C=0), EDI_WTR (C=0) and 0.5, and $Diff_\theta$. Different from the previous empirical study, individual detection by MD of WR erasures no long used the 1.5IQR rule. Instead, the means and standard deviations for MD of WR erasures were used to establish normal distributions in order to generate p values of the observed numbers of WR erasures. Hence, the performance of MD can be evaluated on different α levels and compared with that of other methods. .

Schools became suspect if the within-school percentages of examinees flagged by any of MD, VJ, EDI, EDI_WTR, and $Diff_\theta$ separately were at least1.5 IQR above the third quartiles (Tukey, 1977) in corresponding distributions. Tukey's method of identifying outliers does not rely on rely on a chosen α level. In order to perform useful detection, Type I error rates for outliers should be controlled at or below acceptable level. Therefore, this study investigated the Type I error rates of detection indices and to determine if they were affected by the number of items, examinees, or schools in the sample or for which tampering was detected. If there exists an index, of which the school-level Type I error rate barely varies or constantly reduces when more items, examines, and schools are engaged in tampering, the school-level Type I error rate from no-tampering data sets could then be viewed as the expectation or maximum of the ones from tampered data sets. For a testing program, it is easier to find out such a Type I error rate from test administrations known for no tampering.

District was used in general linear models to predict the within-school percentages. A significant district variable would indicate a positive effect on those percentages.

EDI, EDI_WTR, and $Diff_\theta$ were also entered into mixed models as level-1 dependent variables. The three-level mixed models in Equations 20, 21, and 22 were used to identify potential tampering districts. Preliminary analysis for this study showed that the three-level models could potentially miss some schools that presented unusually high means of EDI, EDI_WTR, and $Diff_{\theta.}$ Therefore, school level detections were performed in the following two-level models.

Level 1: the person level

$$Y_{jkm} = \beta_{0km} + e_{jkm}, \tag{23}$$

Level 2: the school level

$$\beta_{0km} = \delta_{000} + u_{0km}, \tag{24}$$

Where

$$u_{0km} \backsim N(0, \tau_1^2),\ e_{zij} \backsim N(0, \sigma_e^2).$$

Y_{jkm} will be any of EDI, EDI_WTR, and $Diff_\theta$ from examinee j at school k of district m. e_{jkm} is the individual residual and assumed to be normal with mean 0 and variance σ_e^2. δ_{000} is the fixed grand mean of Y_{jkm} for all examinees. u_{0km} is the random effect of school k on Y_{jkm} and assumed to be normal with mean 0 and variance τ_1^2. At α=.05, and .01, schools and districts were flagged if they had significant positive random effects on EDI, EDI_WTR, or $Diff_\theta$.

In order to determine if Type I errors were controlled during the detection, Bradley's (1978) criterion was used. Bradley suggested a conservative and a liberal criterion for identifying boundary conditions for determining whether Type I error rates were acceptable. The conservative criterion is $.9|\alpha| \leq \tau \leq 1.1|\alpha|$ and the liberal criterion is $.5|\alpha| \leq \tau \leq 1.5|\alpha|$, where τ denotes a Type I error rate. A result lower than the lower limit of the criterion or higher than the upper limit of the criterion is taken to indicate loss of Type I error control. In this study,

Bradley's liberal criterion was used. Therefore, when the Type I error rate for a method is larger than |1.5 α| (i.e., absolute value of 1.5α), its power is not reported in the table.

4.3 Software

The simulated data sets were generated in using code written in the computer software R 3.1.3 (R Core Team 2015). The R package "arm" was used for calculating the conditional probabilities in the Van der Linden and Jeon method (2012). Estimation for item and person parameters for all item response models was done with MULTILOG 7.0 (Thissen, 2003). SAS 9.3 software (SAS Institute, Cary NC) was used for the remainder of the analyses.

4.4 Results: Recovery of Generating Parameters

Being a simulation study, it is necessary to demonstrate that the simulated data sets are close to the generating data. This requires small difference between the generating parameters and their estimated values from the simulated data after being placed into the same scales. Using one data set under the no-tampering condition, the bias and root mean square error (RMSE) on item or person parameters were calculated as follows:

$$Bias = \frac{\sum_{j=1}^{N}(\hat{\psi}_j - \psi_j)}{N}, \tag{25}$$

and

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\psi}_j - \psi_j)^2}{N}}. \tag{26}$$

ψ represented the generating discrimination, intercept, or ability parameters, and $\hat{\psi}$ represented its estimates based on the no-tampering data. N was 180 for item parameters, and 35,280 for person parameters.

The bias for discrimination was negligible, only .00006. The RMSE for discrimination was .0649, which is also small, considering that the generating discrimination parameters ranged from -1.33 to 1.71 with standard deviation of .68. Also, the correlation between generating slope parameters and estimated values was high at .9992. These all support the suggestion that the recovery of the generating values was good for discrimination.

A similar conclusion also was reached for intercepts. The bias for intercepts was -.000111. Given that the generating intercept parameters ranged from -2.46 to 2.93 with standard deviation of 1.10, The RMSE was .0217, very small. The correlation was also high at .9998.

The generating abilities ranged from -2.403 to 2.378 and the average standard error of estimate of abilities was .3125. Bias for ability in this simulation was .0427. RMSE for ability was .3077. Both are still acceptable, considering that considering that many responses were tampered and the departure wouldn't affect the interpretation on Type I error and power. The correlation between generating values and ability estimates was good at .9359.

4.5.1 Distribution of Erasures in Fixed-number Tampering

This section presents the distribution of erasures in fixed-number of tampering. Table 14.1 reports the percentages of examinees having different numbers of total erasures observed in the simulated data under the no-tampering and fixed-number tampering conditions. These conditions are labeled as "(0)" to "(18)". In each condition, the means of the results from the ten replications in the simulation study are presented. Standard errors for these means are given in parenthesis in Table 14.1. The magnitudes of these coefficients indicate that, as the number of schools and examinees in the tampering condition increased, the percentages of examinees with simulated numbers of fraudulent erasures also increased as might be expected.

Since the simulation conditions varied with the 3×3=9 combinations of different levels of tampering schools (i.e., 25%, 50%, 100%) and tampered examinees (i.e., 25%, 50%, 100%), Spearman correlations were calculated to represent the association between the percentages of examinees with the target numbers of erasures and the levels of tampering schools or examinees (See Table 14.2). For 5-item tampering, the Spearman correlation between levels of tampered examinees and the percentage of examinees with 5 erasures was .8893 when 25% of schools in tampering districts were selected. It was .8916 when the school percentage was 50%, and .8255 when it was 100%. For 10-item tampering, the Spearman correlation coefficients between levels of tampered examinees and the percentage of examinees with 10 erasures were .9150, .8867, and .8207 at the three levels of tampering schools, respectively. The high correlations indicated that the simulation successfully increased the numbers of examinees having target numbers of erasures.

Conditioned on the three levels of tampered examinees, the Spearman correlations between the levels of percentage of tampering schools and the percentage of examinees with 5 erasures were .8729, .8894, and .8703, respectively. For 10-item tampering, the correlation coefficients between the level of tampering schools and the percentage of examinees with 10 erasures were.8820, .8773, and .8773, respectively. The high correlation indicated that the numbers of examinees with target numbers of erasures increased as planned by having more schools involved in tampering. As an example, Table 15.1 and 15.2 show frequencies of both WR and total erasures for tampered and non-tampered examinees from one data set in which 25% of examinees in 25% of schools from 10 randomly chosen districts were victims of 5-item tampering. As expected, the tampered examinees had greater percentages of having 5 WR erasures and total erasures than the untampered one.

Table 14.1: *Percentages of Examinees Having Different Numbers of Total Erasures in the Simulated Data Sets*

		Simulation Conditions						
			Tampering					
			25% schools in 10 districts					
			5 items			**10 items**		
		(0) No Tempering	**(1) 25% examinees**	**(2) 50% examinees**	**(3) 100% examinees**	**(4) 25% examinees**	**(5) 50% examinees**	**(6) 100% examinees**
		Percent (SE)	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**
Total Erasure Counts	0	15.45 (.76)	15.28 (.75)	15.12 (.73)	14.76 (.7)	15.28 (.75)	15.12 (.73)	14.76 (.7)
	1	58.53 (.75)	57.92 (.77)	57.25 (.79)	55.93 (.85)	57.92 (.77)	57.25 (.79)	55.93 (.85)
	2	16.34 (.06)	16.17 (.06)	16.00 (.06)	15.65 (.07)	16.17 (.06)	16.00 (.06)	15.65 (.07)
	3	5.02 (.02)	4.99 (.03)	4.95 (.03)	4.87 (.03)	4.99 (.03)	4.95 (.03)	4.87 (.03)
	4	2.44 (.03)	2.44 (.03)	2.44 (.03)	2.45 (.03)	2.44 (.03)	2.44 (.03)	2.45 (.03)
	5	1.22 (.02)	1.51 (.03)	1.79 (.03)	2.38 (.11)	1.25 (.02)	1.27 (.02)	1.32 (.02)
	$\geqslant$6	0.99 (.01)	1.69 (.06)	2.46 (.12)	3.97 (.25)	1.96 (.08)	2.98 (.17)	5.02 (.35)

Table 14.1 Continued

Simulation Conditions								
Tampering								
50% schools in 10 districts						**100% schools in 10 districts**		
5 items			**10 items**			**5 items**		
(7) 25% examinees	**(8) 50% examinees**	**(9) 100% examinees**	**(10) 25% examinees**	**(11) 50% examinees**	**(12) 100% examinees**	**(13) 25% examinees**	**(14) 50% examinees**	**(15) 100% examinees**
Percent (SE)	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**
14.72 (.64)	14.38 (.61)	13.73 (.57)	14.72 (.64)	14.38 (.61)	13.73 (.57)	14.34 (.47)	13.67 (.52)	12.34 (.64)
57.77 (.69)	56.52 (.75)	53.99 (.92)	57.77 (.69)	56.52 (.75)	53.99 (.92)	56.56 (.50)	53.93 (.82)	48.69 (1.59)
16.04 (.08)	15.70 (.09)	15.04 (.14)	16.04 (.08)	15.70 (.09)	15.04 (.14)	15.72 (.12)	15.02 (.23)	13.62 (.46)
4.88 (.03)	4.81 (.03)	4.68 (.03)	4.88 (.03)	4.81 (.03)	4.68 (.03)	4.75 (.03)	4.62 (.04)	4.32 (.09)
2.45 (.03)	2.47 (.03)	2.47 (.03)	2.45 (.03)	2.47 (.03)	2.47 (.03)	2.45 (.03)	2.48 (.03)	2.48 (.04)
1.76 (.05)	2.29 (.11)	3.35 (.22)	1.30 (.02)	1.34 (.02)	1.43 (.03)	2.33 (.16)	3.47 (.34)	5.71 (.68)
2.38 (.15)	3.84 (.3)	6.75 (.6)	2.85 (.19)	4.78 (.39)	8.66 (.79)	3.84 (.47)	6.81 (.96)	12.84 (1.92)

Table 14.1 Continued

Simulation Conditions		
Tampering		
10 items		
(16) 25% examinees	**(17) 50% examinees**	**(18) 100% examinees**
Percent (SE)	**Percent (SE)**	**Percent (SE)**
14.34 (.47)	13.67 (.52)	12.34 (.64)
56.56 (.50)	53.93 (.82)	48.69 (1.59)
15.72 (.12)	15.02 (.23)	13.62 (.46)
4.75 (.03)	4.62 (.04)	4.32 (.09)
2.45 (.03)	2.48 (.03)	2.48 (.04)
1.35 (.02)	1.44 (.03)	1.62 (.06)
4.82 (.62)	8.84 (1.27)	16.93 (2.55)

Table 14.2: *Spearman Correlations between the Percentages of Examinees having Target Numbers of Erasures and the Levels of Tampering Involvement*

	Examinees with 5 erasures	**Examinees with 10 erasures**
Tampered examinees in 25% of schools in 10 districts	.8893	.9150
Tampered examinees in 50% of schools in 10 districts	.8916	.8867
Tampered examinees in 100% of schools in 10 districts	.8255	.8207
Tampering schools with 25% of examinees being victims	.8729	.8820
Tampering schools with 50% of examinees being victims	.8894	.8773
Tampering schools with 100% of examinees being victims	.8703	.8773

Table 15.1: *Frequencies of WR Erasures for Untampered and Tampered Examinees*

WR Erasure Count	**Untampered Examinees**		**Tampered Examinees**	
	Frequency	***Percent***	***Frequency***	***Percent***
0	20,004	57.43	0	0
1	12,414	35.64	1	.22
2	1,825	5.24	5	1.11
3	389	1.12	7	1.55
4	98	.28	19	4.21
5	41	.12	285	63.19
6	21	.06	120	26.61
7	17	.05	12	2.88
8	16	.05	1	.22
9	4	.01	0	0

Table 15.2: *Frequencies of Total Erasures for Untampered and Tampered Examinees*

Total Erasure Count	**Untampered Examinees**		**Tampered Examinees**	
	Frequency	***Percent***	***Frequency***	***Percent***
0	4,849	13.92	0	0
1	20,928	60.09	0	0
2	5,638	16.19	4	.89
3	1,761	5.06	7	1.55
4	855	2.45	13	2.88
5	434	1.25	110	24.39
6	176	.51	237	52.55
7	53	.15	51	11.31
8	54	.16	21	4.66
9	43	.12	6	1.33
10	21	.06	1	.22
11-19	17	.04	1	.22

4.5.2 Type I Error Rates and Power of Single-Index Individual Detection in Fixed-number Tampering

This section presents the Type I error rates and power for individual detection by MD of WR erasures, EDI (C=0), VJ, EDI_WTR (C=0 and .5), and $Diff_\theta$ in fixed-number tampering. Type I error rates were evaluated at three α levels, .05, .01 and .001 (see Table 16). Error rates greater than |1.5α| or lower than |.5α| are considered to reflect lack of Type I error control, and they were bolded in Table 16.

In general, MD did not control Type I error rates well. Further, these rates were consistently lower than |.5α| at α =.05 and .01. At α =.001, the rates moved from greater than |1.5α| to lower than |.5α|. As the numbers of tampered items, examinees, and schools increased, the Type I error rates for MD tended to decrease.

$Diff_\theta$ and EDI (C=0) did not control Type I error either, and the rates were never lower than |.5α| under any condition or α level. Adopting C=-.5 as in Wollack et al. (2013) may help obtain acceptable Type I error rates of EDI, but doing that will also reduce power. Given the fact that EDI (C=0) already produced the smallest power among all methods in this study, the results from EDI (C=.5) were not pursued

EDI_WTR (C=0)'s Type I error was only controlled at α =.01, and the observed error rates were less than |α| but greater than |.5α|. At α =.05, none of these error rates was greater than |.5α|. When α =.001, the majority of Type I error rates were larger than |1.5 α|, indicating loss of Type I error control. The Type I error rates did not show any monotonic trend as the numbers of tampered items, examinees, or schools increased.

EDI_WTR (C=.5) maintained control of Type I error at α =.05, and the observed error rates were greater than |α| and less than |1.5 α|. However, at α =.01 and .001, all Type I error

rates of EDI_WTR (C=0) were inflated and thus uncontrolled. Like EDI_WTR (C=0), the rates didn't seem to be linearly related to the number of tampered items, examinees, or schools.

VJ performed best in controlling Type I error, and at α =.05 and .01, and most of the rates were between $|\alpha|$ and $|1.5\ \alpha|$. When α =.001, they were all greater than $|1.5\ \alpha|$. Like MD, the more items, examinees, and schools were involved in tampering, the less VJ made Type I errors.

Power of all methods was evaluated at three α levels, .05, .01 and .001. Only when Type error rates were controlled, the corresponding power values were interpreted and reported in Table 16.2. Since MD is the most common used method in practice, its power was reported just for information purpose as long as the Type I error rates were less than $|1.5\ \alpha|$. But none of MD, $Diff_\theta$, or EDI (C=0) would get interpretation on power.

For fixed-number tampering, VJ always obtained the highest power among methods with controlled Type I error rates. At α =.05 and .01, Power for this index was always greater than .94 and barely varied with for different numbers of tampering schools and or victim examinees.

EDI_WTR (C=.5) only controlled Type I error rates at α =.05. At this level, the power ranged from .7853 and .9081, which were the second highest and tended to decline when more examinees and schools were involved in tampering.

As expected, power of EDI_WTR (C=0) was much smaller than EDI_WTR (C=.5). At α =.01, the power ranged from .2631 to .4252 and went up when more items were contaminated. At α =.001, for the few conditions where existed controlled Type I error, the power ranged from .0833 to .2785.

Table 16.1: *Type I Error Rates of Individual Detection for Fixed-Number Tampering by Six Methods at α =.05, .01, and .001*

Methods	α	Simulation Conditions						
			Tampering					
			25% schools in 10 districts					
			5 items			10 items		
		(0) No Tempering	(1) 25% examinees	(2) 50% examinees	(3) 100% examinees	(4) 25% examinees	(5) 50% examinees	(6) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	.05	**.0165** (.0001)	**.0132** (.0016)	**.0079** (.0015)	**.0049** (.0005)	**.0051** (.0004)	**.0028** (.0004)	**.0012** (.0001)
	.01	**.0165** (.0001)	.0053 (.0003)	**.0048** (.0005)	**.0026** (.0002)	**.0025** (.0002)	**.0014** (.0002)	**.0004** (.0001)
	.001	**.0056** (.0001)	**.0035** (.0005)	**.0025** (.0002)	.0013 (.0001)	.0014 (.0002)	**.0004** (.0001)	**.0000** (.0000)
EDI (C=0)	.05	**.0867** (.0003)	**.0868** (.0003)	**.0868** (.0003)	**.0868** (.0003)	**.0867** (.0003)	**.0849** (.0020)	**.0868** (.0003)
	.01	**.0809** (.0003)	**.0809** (.0003)	**.0810** (.0003)	**.0810** (.0003)	**.0809** (.0003)	**.0810** (.0003)	**.0810** (.0003)
	.001	**.0791** (.0004)	**.0791** (.0003)	**.0793** (.0003)	**.0792** (.0003)	**.0791** (.0003)	**.0793** (.0003)	**.0792** (.0003)
EDI_WTR (C=0)	.05	**.0195** (.0003)	**.0194** (.0003)	**.0195** (.0003)	**.0196** (.0003)	**.0195** (.0003)	**.0183** (.0014)	**.0196** (.0003)
	.01	.0065 (.0002)	.0062 (.0005)	.0066 (.0002)	.0066 (.0002)	.0066 (.0002)	.0062 (.0005)	.0067 (.0002)
	.001	**.0017** (.0001)	**.0016** (.0002)	**.0017** (.0001)	**.0017** (.0001)	.0015 (.0001)	.0016 (.0002)	**.0018** (.0001)

Table 16.1 Continued

Methods	α	Simulation Conditions						
			Tampering					
			25% schools in 10 districts					
			5 items			10 items		
		(0) No Tempering	(1) 25% examinees	(2) 50% examinees	(3) 100% examinees	(4) 25% examinees	(5) 50% examinees	(6) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
EDI_WTR (C=0.5)	.05	.0631 (.0003)	.0630 (.0003)	.0630 (.0003)	.0631 (.0003)	.0630 (.0003)	.0630 (.0003)	.0588 (.0044)
	.01	**.0196** (.0002)	**.0196** (.0002)	**.0197** (.0003)	**.0183** (.0014)	**.0196** (.0003)	**.0197** (.0002)	**.0185** (.0014)
	.001	**.0061** (.0002)	**.0061** (.0002)	**.0061** (.0002)	**.0058** (.0002)	**.0061** (.0002)	**.0062** (.0002)	**.0059** (.0003)
$Diff_\theta$	.05	**.1094** (.0033)	**.1073** (.0032)	**.1081** (.0032)	**.1103** (.0034)	**.1066** (.0032)	**.1080** (.0033)	**.1114** (.0034)
	.01	**.0505** (.0013)	**.0493** (.0012)	**.0498** (.0012)	**.0508** (.0013)	**.0490** (.0012)	**.0497** (.0012)	**.0512** (.0013)
	.001	**.0240** (.0007)	**.0234** (.0007)	**.0237** (.0006)	**.0242** (.0007)	**.0232** (.0006)	**.0236** (.0006)	**.0244** (.0007)
VJ	.05	.0410 (.0003)	.0388 (.0003)	.0367 (.0005)	.0326 (.0007)	.0371 (.0004)	.0334 (.0004)	.0268 (.0011)
	.01	**.0172** (.0003)	**.0157** (.0003)	.0143 (.0003)	.0121 (.0004)	.0146 (.0003)	.0126 (.0004)	.0097 (.0004)
	.001	**.0058** (.0001)	**.0054** (.0001)	**.0050** (.0001)	**.0044** (.0001)	**.0050** (.0001)	**.0045** (.0001)	**.0036** (.0001)

Table 16.1 Continued

Methods	α	Simulation Conditions					
		Tampering					
		50% schools in 10 districts					
		5 items			10 items		
		(10) 25% examinees	(11) 50% examinees	(12) 100% examinees	(13) 25% examinees	(14) 50% examinees	(15) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	.05	**.0102** (.0018)	**.0048** (.0005)	**.0026** (.0002)	**.0026** (.0004)	**.0014** (.0002)	**.0004** (.0001)
	.01	**.0048** (.0005)	**.0025** (.0002)	**.0015** (.0002)	**.0013** (.0003)	**.0005** (.0002)	**.0000** (.0000)
	.001	**.0025** (.0002)	**.0020** (.0006)	.0006 (.0001)	**.0004** (.0001)	**.0001** (.0001)	**.0000** (.0000)
EDI (C=0)	.05	**.0862** (.0003)	**.0862** (.0004)	**.0863** (.0003)	**.0863** (.0003)	**.0873** (.0012)	**.0864** (.0004)
	.01	**.0805** (.0003)	**.0805** (.0004)	**.0806** (.0003)	**.0806** (.0003)	**.0805** (.0004)	**.0806** (.0003)
	.001	**.0786** (.0003)	**.0786** (.0004)	**.0786** (.0003)	**.0786** (.0003)	**.0786** (.0004)	**.0786** (.0003)
EDI_WTR (C=0)	.05	**.0173** (.0017)	**.0186** (.0012)	**.0200** (.0002)	**.0198** (.0002)	**.0200** (.0002)	**.0188** (.0013)
	.01	.0058 (.0006)	.0064 (.0005)	.0070 (.0002)	.0069 (.0002)	.0070 (.0002)	.0061 (.0007)
	.001	.0014 (.0002)	**.0016** (.0001)	**.0018** (.0001)	**.0018** (.0001)	**.0018** (.0001)	**.0016** (.0002)

Table 16.1 Continued

Methods	α	Simulation Conditions					
		Tampering					
		50% schools in 10 districts					
		5 items			10 items		
		(10) 25% examinees	(11) 50% examinees	(12) 100% examinees	(13) 25% examinees	(14) 50% examinees	(15) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
EDI_WTR (C=0.5)	.05	.0628 (.0004)	.0628 (.0004)	.0629 (.0005)	.0627 (.0001)	.0626 (.0004)	.0586 (.0043)
	.01	**.0199** (.0002)	**.0199** (.0002)	**.0201** (.0002)	**.0199** (.0001)	**.0188** (.0012)	**.0189** (.0013)
	.001	**.0064** (.0002)	**.0065** (.0002)	**.0065** (.0002)	**.0065** (.0001)	**.0064** (.0002)	**.0063** (.0003)
$Diff_\theta$	.05	**.1024** (.0035)	**.1040** (.0036)	**.1080** (.0038)	**.1013** (.0034)	**.1039** (.0035)	**.1100** (.0039)
	.01	**.0476** (.0014)	**.0484** (.0015)	**.0505** (.0017)	**.0472** (.0014)	**.0484** (.0015)	**.0513** (.0017)
	.001	**.0229** (.0006)	**.0232** (.0007)	**.0248** (.0008)	**.0227** (.0006)	**.0233** (.0007)	**.0247** (.0008)
VJ	.05	.0370 (.005)	.0332 (.0008)	.0266 (.0012)	.0340 (.0008)	.0278 (.0012)	**.0190** (.0014)
	.01	.0143 (.0003)	.0124 (.0005)	.0097 (.0005)	.0129 (.0005)	.0103 (.0005)	.0070 (.0004)
	.001	**.0051** (.0001)	**.0045** (.0001)	**.0038** (.0001)	**.0047** (.0001)	**.0039** (.0002)	**.0029** (.0001)

Table 16.1 Continued

Methods	α	Simulation Conditions					
		Tampering					
		100% schools in 10 districts					
		5 items			10 items		
		(19) 25% examinees	(20) 50% examinees	(21) 100% examinees	(22) 25% examinees	(23) 50% examinees	(24) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	.05	**.0058** (.0001)	**.0032** (.0003)	**.0020** (.0002)	**.0017** (.0003)	**.0005** (.0001)	**.0001** (.0001)
	.01	**.0032** (.0004)	**.0017** (.0002)	**.0008** (.0002)	**.0008** (.0002)	**.0001** (.0000)	**.0000** (.0000)
	.001	**.0019** (.0002)	.0007 (.0002)	**.0003** (.0001)	**.0001** (.0001)	**.0000** (.0000)	**.0000** (.0000)
EDI (C=0)	.05	**.0862** (.0004)	**.0863** (.0006)	**.0867** (.0006)	**.0863** (.0004)	**.0864** (.0006)	**.0867** (.0006)
	.01	**.0806** (.0005)	**.0807** (.0006)	**.0811** (.0007)	**.0806** (.0005)	**.0808** (.0006)	**.0809** (.0007)
	.001	**.0786** (.0004)	**.0787** (.0006)	**.0789** (.0006)	**.0786** (.0004)	**.0787** (.0006)	**.0789** (.0006)
EDI_WTR (C=0)	.05	**.0196** (.0002)	**.0180** (.0017)	**.0200** (.0003)	**.0196** (.0002)	**.0181** (.0017)	**.0183** (.0016)
	.01	.0068 (.0002)	.0061 (.0007)	.0071 (.0002)	.0068 (.0002)	.0062 (.0007)	.0063 (.0007)
	.001	**.0016** (.0001)	.0014 (.0002)	**.0017** (.0001)	**.0016** (.0001)	.0015 (.0002)	.0015 (.0002)

Table 16.1 Continued

Methods	α	Simulation Conditions					
		Tampering					
		100% schools in 10 districts					
		5 items			10 items		
		(19) 25% examinees	(20) 50% examinees	(21) 100% examinees	(22) 25% examinees	(23) 50% examinees	(24) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
EDI_WTR (C=0.5)	.05	.0629 (.0003)	.0629 (.0004)	.0627 (.0002)	.0626 (.0002)	.0520 (.0070)	.0572 (.0054)
	.01	**.0198** (.0003)	**.0195** (.0002)	**.0199** (.0002)	**.0196** (.0002)	**.0165** (.0020)	**.0183** (.0017)
	.001	**.0063** (.0002)	**.0062** (.0002)	**.0066** (.0002)	**.0063** (.0002)	**.0052** (.0007)	**.0062** (.0004)
$Diff_\theta$	.05	**.1043** (.0033)	**.1068** (.0038)	**.1139** (.0042)	**.1027** (.0037)	**.1068** (.0038)	**.1170** (.0042)
	.01	**.0488** (.0018)	**.0497** (.0017)	**.0533** (.0019)	**.0478** (.0017)	**.0498** (.0018)	**.0552** (.0020)
	.001	**.0227** (.0011)	**.0231** (.0011)	**.0249** (.0011)	**.0224** (.0011)	**.0232** (.0011)	**.0254** (.0011)
VJ	.05	.0340 (.0006)	.0282 (.0011)	**.0198** (.0011)	.0292 (.0010)	**.0214** (.0012)	**.0113** (.0044)
	.01	.0129 (.0002)	.0102 (.0003)	.0074 (.0003)	.0109 (.0003)	.0076 (.0003)	**.0044** (.0003)
	.001	**.0046** (.0001)	**.0039** (.0001)	**.0033** (.0001)	**.0039** (.0001)	**.0032** (.0001)	**.0023** (.0001)

Table 16.2: *Power of Individual Detection for Fixed-Number Tampering by Six Methods at α =.05, .01, and .001*

Methods	α	Simulation Conditions						
			Tampering					
			25% schools in 10 districts					
			5 items			10 items		
		(0) No Tempering	(1) 25% examinees	(2) 50% examinees	(3) 100% examinees	(4) 25% examinees	(5) 50% examinees	(6) 100% examinees
		Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)
MD	.05	- -	.9830 (.0026)	.9734 (.0027)	.9612 (.0041)	.9648 (.0042)	.9376 (.0055)	.8850 (.0104)
	.01	- -	.9693 (.0031)	.9605 (.0040)	.8085 (.0870)	.9372 (.0058)	.8948 (.0106)	.8147 (.0150)
	.001	- -	- -	- -	.2053 (.0373)	.8977 (.0099)	.8223 (.0115)	.5520 (.0982)
EDI (C=0)	.05	- -	- -	- -	- -	- -	- -	- -
	.01	- -	- -	- -	- -	- -	- -	- -
	.001	- -	- -	- -	- -	- -	- -	- -
EDI_WTR (C=0)	.05	- -	- -	- -	- -			
	.01	- -	.2890 (.0107)	.2815 (.0112)	.2826 (.0072)	.4252 (.0110)	.3999 (.0107)	.4079 (.0090)
	.001	- -	- -	- -	- -	.1985 (.0095)	.2015 (.0112)	- -

Table 16.2 Continued

Methods	α	Simulation Conditions						
		Tampering						
		25% schools in 10 districts						
		5 items			10 items			
		(0) No Tempering	(1) 25% examinees	(2) 50% examinees	(3) 100% examinees	(4) 25% examinees	(5) 50% examinees	(6) 100% examinees
		Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)
EDI_WTR (C=0.5)	.05	- -	.9028 (.0060)	.8970 (.0042)	.8949 (.0029)	.9059 (.0066)	.8914 (.0036)	.8746 (.0046)
	.01	- -	- -	- -	- -	- -	- -	- -
	.001	- -	- -	- -	- -	- -	- -	- -
$Diff_\theta$	.05	- -	- -	- -	- -	- -	- -	- -
	.01	- -	- -	- -	- -	- -	- -	- -
	.001	- -	- -	- -	- -	- -	- -	- -
VJ	.05	- -	.9939 (.0012)	.9943 (.0006)	.9932 (.0007)	.9962 (.0011)	.9964 (.0007)	.9963 (.0005)
	.01	- -	- -	.9770 (.0021)	.9703 (.0026)	.9962 (.0011)	.9964 (.0007)	.9963 (.0006)
	.001	- -	- -	- -	- -	- -	- -	- -

Table 16.2 Continued

Methods	α	Simulation Conditions					
		Tampering					
		50% schools in 10 districts					
		5 items			10 items		
		(7) 25% examinees	(8) 50% examinees	(9) 100% examinees	(10) 25% examinees	(11) 50% examinees	(12) 100% examinees
		Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)
MD	.05	.9725 (.0031)	.9559 (.0039)	.8004 (.0913)	.9292 (.0067)	.8824 (.0100)	.8016 (.0142)
	.01	.9574 (.0041)	.7984 (.0922)	.2660 (.0832)	.8827 (.0112)	.8182 (.0165)	.3479 (.1109)
	.001	- -	- -	.0200 (.0050)	.8213 (.0122)	.5553 (.1012)	.0202 (.0141)
EDI (C=0)	.05	- -	- -	- -	- -	- -	- -
	.01	- -	- -	- -	- -	- -	- -
	.001	- -	- -	- -	- -	- -	- -
EDI_WTR (C=0)	.05	- -	- -	- -	- -	- -	- -
	.01	.2712 (.0056)	.2780 (.0039)	.2762 (.0037)	.4051 (.0058)	.4007 (.0049)	.3952 (.0043)
	.001	.0912 (.0015)	- -	- -	- -	- -	- -

Table 16.2 Continued

Methods	α	Simulation Conditions					
		Tampering					
		50% schools in 10 districts					
		5 items			10 items		
		(7) 25% examinees	(8) 50% examinees	(9) 100% examinees	(10) 25% examinees	(11) 50% examinees	(12) 100% examinees
		Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)
EDI_WTR (C=0.5)	.05	.8938 (.0044)	.8923 (.0032)	.8816 (.0030)	.8875 (.0057)	.8713 (.0042)	.8389 (.0069)
	.01	- -	- -	- -	- -	- -	- -
	.001	- -	- -	- -	- -	- -	- -
$Diff_\theta$	.05	- -	- -	- -	- -	- -	- -
	.01	- -	- -	- -	- -	- -	- -
	.001	- -	- -	- -	- -	- -	- -
VJ	.05	.9949 (.0010)	.9938 (.0007)	.9907 (.0018)	.9961 (.0009)	.9956 (.0006)	- -
	.01	.9776 (.0016)	.9757 (.0020)	.9635 (.0041)	.9961 (.0009)	.9956 (.0006)	.9954 (.0006)
	.001	- -	- -	- -	- -	- -	- -

Table 16.2 Continued

Methods	α	Simulation Conditions					
		Tampering					
		100% schools in 10 districts					
		5 items			10 items		
		(13) 25% examinees	(14) 50% examinees	(15) 100% examinees	(22) 25% examinees	(23) 50% examinees	(24) 100% examinees
		Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)
MD	.05	.9643 (.0017)	.9392 (.0034)	.3261 (.0917)	.8945 (.0100)	.8246 (.0143)	.6012 (.1003)
	.01	.9396 (.0045)	.3249 (.0912)	.0535 (.0283)	.8454 (.0103)	.6014 (.0999)	.1170 (.0912)
	.001	- -	.0535 (.0285)	.0030 (.0009)	.7559 (.0179)	.1425 (.0772)	.0006 (.0003)
EDI (C=0)	.05	- -	- -	- -	- -	- -	- -
	.01	- -	- -	- -	- -	- -	- -
	.001	- -	- -	- -	- -	- -	- -
EDI_WTR (C=0)	.05	- -	- -	- -	- -	- -	- -
	.01	.2631 (.0084)	.2688 (.0094)	.2958 (.0086)	.3909 (.0079)	.3816 (.0105)	.3757 (.0090)
	.001	- -	.0833 (.0034)	- -	- -	.1890 (.0071)	.1871 (.0074)

Table 16.2 Continued

Methods	α	Simulation Conditions					
		Tampering					
		100% schools in 10 districts					
		5 items			10 items		
		(13) 25% examinees	(14) 50% examinees	(15) 100% examinees	(16) 25% examinees	(17) 50% examinees	(18) 100% examinees
		Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)
EDI_WTR (C=0.5)	.05	.8906 (.0032)	.8812 (.0058)	.9081 (.0096)	.8764 (.0051)	.8119 (.0313)	.7853 (.0079)
	.01	- -	- -	- -	- -	- -	- -
	.001	- -	- -	- -	- -	- -	- -
$Diff_\theta$	.05	- -	- -	- -	- -	- -	- -
	.01	- -	- -	- -	- -	- -	- -
	.001	- -	- -	- -	- -	- -	- -
VJ	.05	.9934 (.0005)	.9924 (.0010)	- -	.9960 (.0004)	- -	- -
	.01	.9791 (.0023)	.9719 (.0030)	.9487 (.0062)	.9960 (.0004)	.9962 (.0006)	- -
	.001	- -	- -	- -	- -	- -	- -

4.5.3 Type I Error Rates and Power of Single-Index School and District Detection Based on Flagged percentages in Fixed-number Tampering

This section presents the Type I error rates and power for school and district detection based on the within-school percentages of examinees identified by MD, EDI (C=0), EDI_WTR (C=.5), $Diff_\theta$ and VJ. These percentages were sorted and a school was flagged as an outlier in Tukey's method if that percentage was at least 1.5 IQRs greater than the third quartile of all schools.

The different α levels used during individual detections resulted in different percentages of flagged cases for each school. That could create different results in school and district detection. For example, EDI_WTR (C=.5) at α = .05, and .01, flagged 50% and 8.3% of examinees in a school of size 12 respectively. After sorting the flagged percentages of all schools of which individual detection was done at α = .05, 50% was not 1.5 IQRs greater than the third quartile. But after sorting the flagged percentages resulted from individual detection choosing α = .01, 8.3% turned out to be an outlier based on the 1.5 IQR rule. That led to different judgment on whether that school is suspicious. Therefore, results of school detection on the basis of flagged percentages need to be distinguished by which the α level were used for individual detection in Table 17.1 and 17.2. Only individual detection at α = .05 and .01 were followed. Individual detection at α = .001 only identify small amount of truly tampered examines, so its flagged percentages were not used for school detection.

Type I error rates for school detection are reported in Table 17.1. As previously mentioned, since there is no α used in Tukey's method, to make this approach of school detection sensible for test analysts or stakeholders, the school detection should allow them to understand the chance of making Type I errors. It is desirable to find a detection index for which

the school-level Type I error rates are basically unchanged or at least keep decreasing as tampered cases increased so that users can establish the expectation or maximum of the risk of making false positive mistakes. The error rates that didn't meet the two criteria were bolded in Table 17.1. Type I error rates under tampering conditions were compared to those under the no-tampering condition. Let Ω denote the latter. A Type I error rate under a tampering condition was viewed as unchanged if the value lay between $|.5\Omega|$ and $|1.5\Omega|$ and there was no trend in its changing by conditions.

MD's Type I error rates did not show stability across simulated conditions. As can be seen in Table 17.1, the error rates under no-tampering condition was .0701, and it varied a lot with the tampering strategies and the numbers of tampered examinees and school. There was no trend in its changing.

VJ's school-level Type I error rates also fail to be maintained within acceptable range or provide a maximum under no-tampering condition. Type I error rates for VJ tended to increase with the growing numbers of tampered items or examinees but to decrease when more schools were involved in tampering.

EDI_WTR (C=.5)'s school-level Type I error rates under tampering conditions were almost always smaller than the school-level values under no-tampering condition, when individual detection was performed at α = .05 and .01. The school-level Type I error rates based on individual detection at α = .01 tended to decrease as the number of tampered items, examinees, and schools increased. These results indicated that the school-level Type I error would not be restricted at a fixed level. But, if applying Tukey's method on the flagged percentages of EDI_WTR (C=.5), user might be able to estimate the maximum of Type I error by simulating no-tampering datasets or using data sets known for tampering-free.

EDI (C=0) and $Diff_\theta$ shared similar characteristics with EDI_WTR (C=.5) on Type I error rates. For flagged percentages from individual detection at either α = .05 or .01, school-level Type I error rates under tampering conditions were not controlled and consistently smaller than under the no-tampering condition. Further, they tended to decrease as the numbers of tampered items, examinees, and schools increased, hence, the maximum might be estimable from no-tampering data sets.

Power of school detection based on Tukey's 1.5 IQR rule was reported in Table 17.2. Results from flagged percentages by MD and VJ were not reported since it is hard to estimate the exact or even maximum chance of making Type I errors.

EDI_WTR (C=.5) had more correctly identified schools than $Diff_\theta$, and EDI (C=0), ranging from .4264 to 1. The power of school detections didn't vary much between the two α levels chosen for individual flagging. It also increased as more examinees became the victims of tampering but decreased as the number of tampering schools increased. The impact of numbers of tampered items had not clear trend.

It was not clear whether $Diff_\theta$ worked better than EDI (C=0). They both made more true school detections as tampered items and examinees increased, but their performances were hindered by the numbers of tampering schools. The school-level power of $Diff_\theta$ ranged from .0084 to .9946, and went higher if the flagged percentages were from individual detection made at α =.01. For EDI (C=0), power ranged from .0641 to .8746, and more tampering schools were identified if the flagged percentages were from individual detection made at α =.05.

Table 17.1: *Type I Error Rates of Schools Selected as Tukey's Outliers Based on the Within-School Percentages of Flagged Examinees at α =.05, and .01*

Methods	α at Individual Detection	Simulation Conditions						
			Tampering					
			25% schools in 10 districts					
			5 items			10 items		
		(0) No Tempering	(1) 25% examinees	(2) 50% examinees	(3) 100% examinees	(4) 25% examinees	(5) 50% examinees	(6) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	[.05]	.0701 (.0020)	.0625 (.0066)	**.1492** (.0163)	**.1561** (.0137)	**.1644** (.0108)	.0946 (.0111)	.0402 (.0054)
	[.01]	.0701 (.0020)	**.1737** (.0066)	**.1561** (.0137)	.0825 (.0053)	.0859 (.0053)	.0483 (.0068)	**.0117** (.0038
EDI (C=0)	[.05]	.0440 (.0026)	.0367 (.0023)	.0315 (.0021)	.0299 (.0019)	.0316 (.0021)	.0300 (.0019)	.0298 (.0018)
	[.01]	.0413 (.0016)	.0385 (.0023)	.0382 (.0022)	.0358 (.0024)	.0353 (.0031)	.0333 (.0029)	.0279 (.0021
EDI_WTR (C=0.5)	[.05]	.0329 (.0028)	.0279 (.0021)	.0253 (.0014)	.0254 (.0014)	.0252 (.0014)	.0254 (.0014)	.0254 (.0014)
	[.01]	.0728 (.0024)	.0565 (.0020)	.0570 (.0019)	.0570 (.0019)	.0577 (.0021)	.0572 (.0018)	.0574 (.0018)
$Diff_\theta$	[.05]	.0346 (.0023)	.0312 (.0019)	.0270 (.0019)	.0265 (.0019)	.0298 (.0019)	.0268 (.0018)	.0268 (.0018)
	[.01]	.0531 (.0008)	.0402 (.0027)	.0373 (.0026)	.0383 (.0027)	.0379 (.0024)	.0372 (.0025)	.0382 (.0026)
VJ	[.05]	.0303 (.0023)	**.0239** (.0013)	**.0269** (.0016)	**.0298** (.0020)	**.0273** (.0016)	**.0317** (.0020)	**.0358** (.0017)
	[.01]	.0622 (.0022)	**.0508** (.0024)	**.0540** (.0017)	**.0584** (.0019)	**.0539** (.0018)	**.0571** (.0019)	**.0685** (.0030)

Table 17.1 Continued

Methods	α at Individual Detection	Simulation Conditions					
		Tampering					
		50% schools in 10 districts					
		5 items			10 items		
		(7) 25% examinees	(8) 50% examinees	(9) 100% examinees	(10) 25% examinees	(11) 50% examinees	(12) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	[.05]	.0716 (.0148)	**.1268** (.0194)	.0853 (.0057)	.0837 (.0155)	.0482 (.0085)	**.0106** (.0034)
	[.01]	**.1391** (.0148)	.0853 (.0057)	.0512 (.0078)	.0482 (.0085)	**.0185** (.0062)	**.0009** (.0005)
EDI (C=0)	[.05]	.0272 (.0016)	.0254 (.0018)	.0228 (.0017)	.0260 (.0017)	.0236 (.0016)	.0214 (.0022)
	[.01]	.0417 (.0042)	.0355 (.0037)	.0294 (.0027)	.0393 (.0065)	.0281 (.0031)	.0264 (.0020)
EDI_WTR (C=0.5)	[.05]	.0222 (.0026)	.0221 (.0025)	.0237 (.0024)	.0220 (.0025)	.0223 (.0026)	.0222 (.0026)
	[.01]	.0458 (.0032)	.0443 (.0036)	.0433 (.0025)	.0449 (.0036)	.0432 (.0034)	.0427 (.0035)
$Diff_\theta$	[.05]	.0252 (.0018)	.0203 (.0022)	.0194 (.0024)	.0252 (.0018)	.0197 (.0026)	.0204 (.0022)
	[.01]	.0278 (.0027)	.0273 (.0028)	.0266 (.0029)	.0287 (.0029)	.0274 (.0027)	**.0316** (.0059)
VJ	[.05]	**.0216** (.0024)	**.0238** (.0021)	**.0280** (.0019)	**.0236** (.0024)	**.0266** (.0020)	**.0323** (.0019)
	[.01]	**.0417** (.0030)	**.0448** (.0024)	**.0533** (.0040)	**.0432** (.0027)	**.0484** (.0026)	**.0661** (.0040)

Table 17.1 Continued

Methods	α at Individual Detection	Simulation Conditions					
		Tampering					
		100% schools in 10 districts					
		5 items			10 items		
		(13) 25% examinees	(14) 50% examinees	(15) 100% examinees	(16) 25% examinees	(17) 50% examinees	(18) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	[.05]	.0456 (.0094)	.0718 (.0129)	.0596 (.0070)	.0570 (.0081)	.0177 (.0041)	.0029 (.0018)
	[.01]	.0817 (.0116)	.0596 (.0070)	.0277 (.0060)	.0258 (.0064)	.0025 (.0017)	.0000 (.0000)
EDI (C=0)	[.05]	.0373 (.0096)	.0192 (.0020)	.0134 (.0022)	.0216 (.0019)	.0152 (.0017)	.0130 (.0025)
	[.01]	.0358 (.0035)	.0332 (.0028)	.0252 (.0034)	.0284 (.0029)	.0183 (.0016)	.0144 (.0017)
EDI_WTR (C=0.5)	[.05]	**.0129** (.0027)	**.0126** (.0028)	**.0140** (.0030)	**.0126** (.0028)	**.0143** (.0031)	**.0135** (.0032)
	[.01]	.0298 (.0022)	.0267 (.0032)	.0187 (.0033)	.0298 (.0024)	.0244 (.0033)	.0234 (.0036)
$Diff_\theta$	[.05]	.0247 (.0018)	.0138 (.0025)	.0203 (.0028)	.0247 (.0022)	.0143 (.0028)	.0137 (.0025)
	[.01]	.0260 (.0023)	.0188 (.0026)	**.0312** (.0048)	.0248 (.0030)	.0196 (.0026)	.0197 (.0039)
VJ	[.05]	**.0147** (.0017)	**.0175** (.0029)	**.0203** (.0028)	**.0175** (.029)	**.0183** (.0029)	**.0273** (.0020)
	[.01]	**.0248** (.0048)	**.0287** (.0037)	**.0312** (.0048)	**.0298** (.0038)	**.0300** (.0047)	**.0407** (.0065)

Table 17.2: *Power of School-Level Detection as Tukey's Outliers Based on the Within-School Percentages of Flagged Examinees at* α *=.05 and .01*

Methods	α at Individual Detection	Simulation Conditions						
			Tampering					
			25% schools in 10 districts					
			5 items			10 items		
		(0) No Tempering	(1) 25% examinees	(2) 50% examinees	(3) 100% examinees	(4) 25% examinees	(5) 50% examinees	(6) 100% examinees
		Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)
MD	[.05]	- -	- -	- -	- -	- -	- -	- -
	[.01]	- -	- -	- -	- -	- -	- -	- -
EDI (C=0)	[.05]	- -	.1499 (.0185)	.3627 (.0168)	.7713 (.0146)	.1950 (.0211)	.5936 (.0214)	.9406 (.0092)
	[.01]	- -	.0639 (.0129)	.1351 (.0224)	.2994 (.0178)	.1029 (.0167)	.3112 (.0290)	.6479 (.0193)
EDI_WTR (C=0.5)	[.05]	- -	.7610 (.0258)	.9950 (.0035)	1 (.0000)	.7627 (.0257)	.9892 (.0058)	1 (.0000)
	[.01]	- -	.7373 (.0202)	.9366 (.0144)	.9946 (.0028)	.7991 (.0121)	.9758 (.0080)	.9946 (.0028)
$Diff_\theta$	[.05]	- -	.0203 (.0075)	.3753 (.0374)	.9844 (.0103)	.0084 (.0047)	.3770 (.0354)	.9568 (.0109)
	[.01]	- -	.1117 (.0271)	.7958 (.0288)	.9734 (.0070)	.0780 (.0252)	.8357 (.0179)	.9568 (.0114)
VJ	[.05]	- -	- -	- -	- -	- -	- -	- -
	[.01]	- -	- -	- -	- -	- -	- -	- -

Table 17.2 Continued

Methods	α at Individual Detection	Simulation Conditions					
		Tampering					
		50% schools in 10 districts					
		5 items			10 items		
		(7) 25% examinees	(8) 50% examinees	(9) 100% examinees	(10) 25% examinees	(11) 50% examinees	(12) 100% examinees
		Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)
MD	[.05]	- -	- -	- -	- -	- -	- -
	[.01]	- -	- -	- -	- -	- -	- -
EDI (C=0)	[.05]	.0916 (.0106)	.3113 (.0166)	.7020 (.0268)	.1541 (.0119)	.5336 (.0162)	.8746 (.0127)
	[.01]	.0641 (.0083)	.1093 (.0088)	.2226 (.0201)	.1011 (.0143)	.2688 (.0144)	.5856 (.0192)
EDI_WTR (C=0.5)	[.05]	.6930 (.0424)	.9822 (.0058)	1 (.0000)	.6744 (.0425)	.9856 (.0037)	1 (.0000)
	[.01]	.6667 (.0269)	.8929 (.0084)	.9834 (.0042)	.7696 (.0231)	.9361 (.0095)	.9918 (.0045)
$Diff_\theta$	[.05]	.0405 (.0096)	.3698 (.0489)	.9774 (.0056)	.0162 (.00042)	.3763 (.0664)	.9597 (.0056)
	[.01]	.0824 (.0231)	.7445 (.0288)	.9724 (.0073)	.0747 (.0268)	.7741 (.0445)	.9603 (.0065)
VJ	[.05]	- -	- -	- -	- -	- -	- -
	[.01]	- -	- -	- -	- -	- -	- -

Table 17.2 Continued

Methods	α at Individual Detection	Simulation Conditions					
		Tampering					
		100% schools in 10 districts					
		5 items			10 items		
		(19) 25% examinees	(20) 50% examinees	(21) 100% examinees	(22) 25% examinees	(23) 50% examinees	(24) 100% examinees
		Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)
MD	[.05]	- -	- -	- -	- -	- -	- -
	[.01]	- -	- -	- -	- -	- -	- -
EDI (C=0)	[.05]	.0970 (.0122)	.2236 (.0204)	.6810 (.0566)	.1186 (.0079)	.4070 (.0621)	.7874 (.0359)
	[.01]	.0691 (.0113)	.0948 (.0076)	.2291 (.0187)	.0960 (.0135)	.2225 (.0324)	.4546 (.0514)
EDI_WTR (C=0.5)	[.05]	.4504 (.0678)	.9890 (.0039)	1 (.0000)	.4264 (.0653)	.9726 (.0120)	.9977 (.0016)
	[.01]	.5302 (.0349)	.8480 (.0337)	.9803 (.0070)	.6596 (.0575)	.9111 (.0252)	.9809 (.0033)
$Diff_\theta$	[.05]	.0292 (.0039)	.1260 (.0341)	.9525 (.1271)	.0392 (.0226)	.1163 (.0367)	.9274 (.1242)
	[.01]	.0368 (.0072)	.4431 (.0739)	.9453 (.0065)	.0208 (.0062)	.4151 (.0691)	.9256 (.0145)
VJ	[.05]	- -	- -	- -	- -	- -	- -
	[.01]	- -	- -	- -	- -	- -	- -

Type I error rates for district detection based on flagged percentages are summarized in Table 17.3Type I error. General linear models used district as a predictor of the within-school percentage of examines detected as having tampered tests. The decision rule used was that if the intercept for a district was greater than 0 at $\alpha = .05$, the district would be suspect.

For most conditions, most Type I error rates of all methods for district level detection fell below the liberal lower boundary, .025. The error rates decreased as the numbers of tampered items and examinees decreased but increased as the number of tampering schools increased. Results in Table 17.3 appear to suggest that EDI(C=0) and $Diff_\theta$ controlled Type I error rates under more conditions than others. However, this was not the case. The Type I error rates of all methods were often 0 but occasionally increased to greater than .1, resulting very large standard errors which were more than half of the Type I error rates. As a result, power was not interpreted for districts.

Table 17.3: *Type I Error Rates of Detected Districts as Indicators of School Outliers Based on the Within-School Percentages of Flagged Examinees at α =.05*

Methods	α at Individual Detection	Simulation Conditions						
			Tampering					
			25% schools in 10 districts					
			5 items			10 items		
		(0) No Tempering	(1) 25% examinees	(2) 50% examinees	(3) 100% examinees	(4) 25% examinees	(5) 50% examinees	(6) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	[.05]	.0352 (.0131)	**.0023** (.0023)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
EDI (C=0)	[.05]	.0593 (.0361)	.0340 (.0201)	.0272 (.0179)	**.0068** (.0068)	.0295 (.0183)	**.0227** (.0182)	**.0000** (.0000)
EDI_WTR (C=0.5)	[.05]	.0241 (.0083)	**.0136** (.0050)	**.0045** (.0030)	**.0000** (.0000)	**.0136** (.0050)	**.0045** (.0030)	**.0022** (.0022)
$Diff_\theta$	[.05]	.0333 (.0120)	**.0248** (.0115)	.0293 (.00118)	**.0045** (.0030)	.0293 (.0102)	**.0225** (.0076)	**.0000** (.0000)
VJ	[.05]	.0222 (.0086)	**.0022** (.0022)	**.0000** (.0000)	**.0000** (.0000)	**.0022** (.0022)	**.0000** (.0000)	**.0000** (.0000)

Table 17.3 Continued

Methods	α at Individual Detection	Simulation Conditions					
		Tampering					
		50% schools in 10 districts					
		5 items			10 items		
		(7) 25% examinees	(8) 50% examinees	(9) 100% examinees	(10) 25% examinees	(11) 50% examinees	(12) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	[.05]	**.0023** (.0023)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
EDI (C=0)	[.05]	.0388 (.0131)	.0273 (.0121)	**.0091** (.0050)	.0477 (.0215)	.0296 (.0122)	**.0114** (.0051)
EDI_WTR (C=0.5)	[.05]	**.0227** (.0117)	**.0068** (.0049)	**.0000** (.0000)	**.0159** (.0090)	**.0068** (.0049)	**.0000** (.0000)
$Diff_\theta$	[.05]	.0386 (.0206)	.0250 (.0120)	**.0000** (.0000)	.0386 (.0206)	**.0205** (.0120)	**.0000** (.0023)
VJ	[.05]	**.0136** (.0113)	**.0023** (.0023)	**.0000** (.0000)	**.0045** (.0030)	**.0023** (.0023)	**.0000** (.0000)

Table 17.3 Continued

Methods	α at Individual Detection	Simulation Conditions					
		Tampering					
		100% schools in 10 districts					
		5 items			10 items		
		(19) 25% examinees	(20) 50% examinees	(21) 100% examinees	(22) 25% examinees	(23) 50% examinees	(24) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	[.05]	**.0085** (.0042)	**.0028** (.0028)	**.0000** (.0000)	**.0028** (.0028)	**.0000** (.0000)	**.0000** (.0000)
EDI (C=0)	[.05]	.0398 (.0165)	.0313 (.0142)	**.0171** (.0094)	.0426 (.0184)	.0284 (.0127)	**.0114** (.0061)
EDI_WTR (C=0.5)	[.05]	**.0799** (.0424)	.0739 (.0427)	.0710 (.0433)	**.0767** (.0432)	.0710 (.0399)	.0455 (.0254)
$Diff_\theta$	[.05]	.0369 (.0192)	.0256 (.0125)	**.0205** (.0093)	.0313 (.0135)	.0313 (.0154)	**.0199** (.0100)
VJ	[.05]	.0341 (.0121)	.0341 (.0122)	**.0159** (.0076)	.0284 (.0120)	.0483 (.0145)	.0341 (.0166)

4.5.4 Type I Error Rates and Power of Single-Index School and District Detection Based on Mixed Models in Fixed-number Tampering

This section presents Type I error rates and power for school and district detection based on the values of EDI_WTR (C=.5), EDI (C=0), and $Diff_\theta$. For an examinee with at least one WR erasure, the values of EDI_WTR (C=.5), EDI (C=0), and $Diff_\theta$ were treated as the level 1 dependent variable in mixed models analyses.

School level detection was done using a two-level mixed model (Equations 23 and 24). A school might be flagged if its random intercept was significantly greater than 0.

Table 18.1 summarizes Type I error for the two-level mixed models at α =.05, and .01. All Type I error rates of EDI (C=0) and $Diff_\theta$ were below the liberal lower boundary of α, indicating lack of control. EDI_WTR (C=.5) controlled the Type I error when 25% schools in chosen districts participated in 5-item and 10-item tampering. As the number of tampering schools and tampered items increased, Type I error rates declined. When 50% of school were involved in 5-item tampering, EDI_WTR (C=.5) at α =.01 still controlled the error rates across the three percentages of tampered examinees, and at α =.05 it only controlled the error rate at 25% of tampered examinees. Beyond these, there was no more controlled error rate.

Power of EDI_WTR (C=.5) was reported in Table 18.2. At α =.01, in the conditions having controlled Type I error rate, the power ranged from .4527 to .9217. Large α levels, tampered examinee and schools percentages, and tampered item amounts all increase the power. At α =.05, in the conditions having controlled Type I error rates, the power varied between .6004 and .9846.

District level detection was performed with three-level mixed models, in order to remove school effects before looking at district effects. A district might be flagged if its random intercept was significantly greater than 0.

Table 18.3 summarizes Type I error of district detection at α =.05, and .01. Across simulation conditions, Type I error rates of all three indices were very small, almost always 0, similar to the results of the previous district direction based on flagged percentages. Therefore, no power of district detection was interpreted.

Table 18.1: *School-Level Type I Error Rates in Two-Level Mixed Models at α =.05 and .01 for Fixed-Number Tampering*

Methods	α	**Simulation Conditions**						
			Tampering					
			25% schools in 10 districts					
			5 items			**10 items**		
		(0) No Tempering	**(1) 25% examinees**	**(2) 50% examinees**	**(3) 100% examinees**	**(4) 25% examinees**	**(5) 50% examinees**	**(6) 100% examinees**
EDI_WTR (C=0.5)	**.05**	.0338 (.0026)	.0316 (.0014)	.0327 (.0021)	.0320 (.0014)	.0291 (.0015)	.0263 (.0020)	.0250 (.0014)
	.01	.0085 (.0012)	.0087 (.0009)	.0092 (.0013)	.0085 (.0010)	.0065 (.0010)	.0066 (.0011)	.0056 (.0010)
EDI (C=0)	**.05**	**.0000** (.0000)	**.0050** (.0037)	**.0034** (.0007)	**.0065** (.0008)	**.0015** (.0005)	**.0042** (.0010)	**.0062** (.0012)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
$Diff_\theta$	**.05**	**.0000** (.0000)	**.0003** (.0002)	**.0004** (.0003)	**.0005** (.0003)	**.0003** (.0003)	**.0003** (.0003)	**.0004** (.0003)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)

Table 18.1 Continued

Methods	α	Simulation Conditions					
		Tampering					
		50% schools in 10 districts					
		5 items			10 items		
		(7) 25% examinees	(8) 50% examinees	(9) 100% examinees	(10) 25% examinees	(11) 50% examinees	(12) 100% examinees
EDI_WTR (C=0.5)	**.05**	.0260 (.0017)	**.0242** (.0017)	**.0207** (.0019)	**.0147** (.0021)	**.0122** (.0021)	**.0086** (.0016)
	.01	.0058 (.0003)	.0066 (.0006)	.0061 (.0007)	**.0036** (.0004)	**.0028** (.0006)	**.0029** (.0006)
EDI (C=0)	**.05**	**.0007** (.0004)	**.0018** (.0004)	**.0030** (.0009)	**.0009** (.0005)	**.0013** (.0007)	**.0014** (.0007)
	.01	**.0000** (.0000	**.0000** (.0000	**.0004** (.0003)	**.0001** (.0001)	**.0001** (.0001)	**.0001** (.0001)
$Diff_\theta$	**.05**	**.0003** (.0003)	**.0001** (.0001)	**.0003** (.0002)	**.0001** (.0001)	**.0000** (.0000)	**.0000** (.0000)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000	**.0000** (.0000)

Table 18.1 Continued

Methods	α	Simulation Conditions					
		Tampering					
		100% schools in 10 districts					
		5 items			10 items		
		(13) 25% examinees	(14) 50% examinees	(15) 100% examinees	(16) 25% examinees	(17) 50% examinees	(18) 100% examinees
EDI_WTR (C=0.5)	**.05**	**.0167** (.0033)	**.0126** (.0027)	**.0080** (.0020)	**.0085** (.0023)	**.0052** (.0015)	**.0033** (.0012)
	.01	**.0041** (.00012)	**.0024** (.0007)	**.0021** (.0009)	**.0019** (.0005)	**.0010** (.0004)	**.0006** (.0004)
EDI (C=0)	**.05**	**.0011** (.0007)	**.0016** (.0007)	**.0030** (.0017)	**.0006** (.0004)	**.0004** (.0002)	**.0005** (.0004)
	.01	**.0000** (.0000)	**.0002** (.0002)	**.0002** (.0002)	**.0002** (.0002)	**.0000** (.0000)	**.0000** (.0000)
$Diff_\theta$	**.05**	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)

Table 18.2: *School-Level Power in Two-Level Mixed Models at α =.05 and .01 for Fixed-Number Tampering*

Methods	α	**Simulation Conditions**					
		Tampering					
		25% schools in 10 districts					
		5 items			**10 items**		
		(1) 25% examinees	**(2) 50% examinees**	**(3) 100% examinees**	**(4) 25% examinees**	**(5) 50% examinees**	**(6) 100% examinees**
EDI_WTR (C=0.5)	**.05**	.6116 (.0146)	.8812 (.0120)	.9875 (.0044)	.7404 (.0223)	.9316 (.0113)	.9846 (.0044)
	.01	.4527 (.0298)	.7972 (.0230)	.9658 (.0065)	.5753 (.0115)	.8655 (.0158)	.9757 (.0058)
EDI (C=0)	**.05**	- -	- -	- -	- -	- -	- -
	.01	- -	- -	- -	- -	- -	- -
$Diff_\theta$	**.05**	- -	- -	- -	- -	- -	- -
	.01	- -	- -	- -	- -	- -	- -

Table 18.2 Continued

Methods	α	Simulation Conditions		
		Tampering		
		50% schools in 10 districts		
		5 items		
		(7) 25% examinees	(2) 50% examinees	(3) 100% examinees
EDI_WTR (C=0.5)	**.05**	.6004 (.0191)	- -	- -
	.01	.4458 (.0199)	.7446 (.0212)	.9217 (.0098)
EDI (C=0)	**.05**	- -	- -	--
	.01	- -	- -	-
$Diff_\theta$	**.05**	- -	- -	-
	.01	- -	- -	

Table 18.3: *District-Level Type I Error Rates in Three -Level Mixed Models at α =.05 and .01 for fixed-number tampering*

Methods	α	**Simulation Conditions**						
			Tampering					
			25% schools in 10 districts					
			5 items			**10 items**		
		(0) No Tempering	**(1) 25% examinees**	**(2) 50% examinees**	**(3) 100% examinees**	**(4) 25% examinees**	**(5) 50% examinees**	**(6) 100% examinees**
EDI_WTR (C=0.5)	**.05**	**.0185** (.0039)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
EDI (C=0)	**.05**	**.0056** (.0028)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
$Diff_\theta$	**.05**	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)

Table 18.3 Continued

Methods	α	Simulation Conditions					
		Tampering					
		50% schools in 10 districts					
		5 items			10 items		
		(7) 25% examinees	(8) 50% examinees	(9) 100% examinees	(10) 25% examinees	(11) 50% examinees	(12) 100% examinees
EDI_WTR (C=0.5)	**.05**	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
EDI (C=0)	**.05**	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
$Diff_\theta$	**.05**	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)

Table 18.3 Continued

Methods	α	Simulation Conditions					
		Tampering					
		100% schools in 10 districts					
		5 items			10 items		
		(13) 25% examinees	(14) 50% examinees	(15) 100% examinees	(16) 25% examinees	(17) 50% examinees	(18) 100% examinees
EDI_WTR (C=0.5)	**.05**	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
EDI (C=0)	**.05**	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
$Diff_\theta$	**.05**	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)

4.6.1 Distribution of Erasures in Score-based Tampering

This section presents the distribution of erasures in score-based tampering. Table 19.1 reports the percentages of examinees having different numbers of total erasures observed in the simulated data under the no-tampering and score-based tampering conditions. There was little if any difference between the percentages under the no-tampering condition and the percentages under the score-based tampering conditions with 25% of examinees from 25% of schools in 10 randomly chosen districts. The only differences observed were from those that had 0.58% more examinees with more than six total erasures.

For score-based tampering, the Spearman correlations (See Table 19.2) between levels of tampered examinees and the percentage of examinees having at least five erasures were .8498, .8237, and .6821, respectively at three levels of tampering schools (i.e., 25%, 50%, 100%).

At the three levels of tampered examinees, the Spearman correlations between the levels of score-based tampering schools and the percentage of examinees with at least five erasures were .6991 .6907, and .6969, respectively.

The correlations noted above indicate that the changes in the numbers of examinees and schools selected for score-based tampering affected, as intended, the numbers of examinees with large numbers of erasures as planned. These correlations were less than the ones found in fixed-number tampering. In part, this was due to the fact that score-based tampering might actually simply add more tampered examinees having less than five erasures, instead of the examinees having at least five erasures.

Table 19.1: *Percentages of Examinees Having Different Numbers of Total Erasures in the Simulated Data Sets*

		Simulation Conditions						
			Tampering					
			25% schools in 10 districts			**50% schools in 10 districts**		
			Score-based			**Score-based**		
		(0) No Tempering	**(19) 25% examinees**	**(20) 50% examinees**	**(21) 100% examinees**	**(22) 25% examinees**	**(23) 50% examinees**	**(24) 100% examinees**
		Percent (SE)	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**	**Percent (SE)**
Total Erasure Counts	0	15.45 (.76)	15.44 (.76)	15.42 (.76)	15.37 (.76)	15.00 (.66)	15.04 (.64)	14.89 (.65)
	1	58.53 (.75)	58.27 (.76)	57.96 (.77)	57.38 (.80)	58.49 (.66)	57.89 (.68)	56.81 (.75)
	2	16.34 (.06)	16.25 (.06)	16.15 (.06)	15.95 (.06)	16.18 (.08)	15.98 (.07)	15.60 (.10)
	3	5.02 (.02)	5.01 (.02)	4.98 (.02)	4.95 (.03)	4.91 (.03)	4.89 (.03)	4.79 (.03)
	4	2.44 (.03)	2.44 (.03)	2.45 (.03)	2.44 (.03)	2.45 (.03)	2.46 (.03)	2.47 (.03)
	5	1.22 (.02)	1.24 (.02)	1.25 (.02)	1.28 (.02)	1.28 (.02)	1.31 (.02)	1.36 (.02)
	$\geqslant$6	0.99 (.01)	1.37 (.03)	1.79 (.07)	2.64 (.15)	1.71 (.08)	2.43 (.17)	4.09 (.33)

Table 19.1 Continued

Simulation Conditions		
Tampering		
100% schools in 10 districts		
Score-based		
(25) 25% examinees	**(26) 50% examinees**	**(27) 100% examinees**
Percent (SE)	**Percent (SE)**	**Percent (SE)**
14.93 (.44)	14.86 (.45)	14.72 (.46)
58.04 (.45)	56.92 (.51)	54.60 (.77)
15.98 (.08)	15.57 (.15)	14.78 (.28)
4.83 (.03)	4.75 (.03)	4.61 (.04)
2.46 (.02)	2.45 (.02)	2.45 (.03)
1.31 (.02)	1.34 (.02)	1.46 (.03)
2.44 (.25)	4.11 (.53)	7.38 (1.08)

Table 19.2: *Spearman Correlation between the Percentages of Examinees Having Target Numbers of Erasures and the Levels of Tampering Involvement*

	Examinees with at least 5 erasures
Tampered examinees in 25% of schools in 10 districts	.8498
Tampered examinees in 50% of schools in 10 districts	.8237
Tampered examinees in 100% of schools in 10 districts	.6821
Tampering schools with 25% of examinees being victims	.6991
Tampering schools with 50% of examinees being victims	.6907
Tampering schools with 100% of examinees being victims	.6969

4.6.2 Type I error rates and Power of Single-Index Individual Detection in Score-based Tampering

This section presents the Type I error rates and power of individual detection for score-based tampering by MD of WR erasures, EDI (C=0), VJ, EDI_WTR (C=0 and .5), and $Diff_\theta$ in score-based tampering.

Type I error rates (see Table 20.1) were evaluated at three α levels, .05, .01 and .001. When the numbers of tampered items, examinees, and schools increased, all Type I error rates tended to decrease. The exceptions were rates for EDI_WTR(C=0.5) and $Diff_\theta$.

MD lost control of most Type I error rates across all α levels and conditions. These rates were consistently lower than $|.5\alpha|$ at α =.05 and .01. At α =.001, the rates changed from being greater than $|1.5\alpha|$ to lower than $|.5\alpha|$, as more examines, schools, and items were involved in tampering.

Neither EDI (C=0) or $Diff_\theta$ maintained Type I error rates lower than $|1.5\alpha|$ under any condition or α level.

At α =.01, EDI_WTR (C=0) was able to control all Type error rates between $|.5\alpha|$ and $|\alpha|$. But at α =.05, no error rate was greater than $|.5\alpha|$ at α =.05. At α =.001, many were greater than $|1.5\alpha|$, only some were controlled between $|\alpha|$ and $|1.5\alpha|$ especially when 100% of schools in chosen districts were involved in score-based tampering.

EDI_WTR (C=.5) kept all Type I error rates under control at α =.05, greater than $|\alpha|$ and less than $|1.5\alpha|$. However, at α =.01 and .001, all Type I error rates of EDI_WTR (C=0) were inflated and thus uncontrolled. The rates didn't seem to maintain positive or negative relationship with the number of tampered items, examinees, or schools.

VJ had all Type I error rates controlled at α =.01. When α =.05, many were greater than $|1.5\alpha|$, but some were controlled between $|\alpha|$ and $|1.5\alpha|$ especially when 100% of schools in chosen districts were involved in score-based tampering. At α =.001, all Type I error rates passed the upper boundary of α.

Power of individual detection was evaluated at three α levels, .05, .01 and .001, and was reported for those conditions for which Type error rates were controlled (see Table 16.2). MD's power was also reported just for information purpose as long as the Type I error rates were less than $|1.5\alpha|$, but it would not be interpreted since MD didn't' t control Type I error rates .

For score-based tampering, VJ always had the highest power (i.e., for those conditions under which Type I error rates were controlled). At α =.05, power ranged from .9518 to .9812. At α =.01, the power was above .88.

The power of EDI_WTR (C=.5) was as good as VJ at α =.05, ranging from .9694 to .9815.

As expected, power of EDI_WTR (C=0) was smaller than that for EDI_WTR (C=.5). At α =.05, power of EDI_WTR (C=0) ranged from .6946 to .7256. As was observed for both VJ and EDI_WTR (C=.5), there was no clear trend between power and the numbers of tampered items, of tampered examinees, and of tampering schools. At α =.01, the power was above .40 for simulated score-based tampering with 100% of schools in tampering districts.

Table 20.1: *Type I Error Rates of Individual Detection for Score-Based Tampering by Six Methods at α =.05, .01, and .001*

Methods	α	Simulation Conditions						
			Tampering					
			25% schools in 10 districts			50% schools in 10 districts		
			Score-based			Score-based		
		(0) No Tempering	(19) 25% examinees	(20) 50% examinees	(21) 100% examinees	(22) 25% examinees	(23) 50% examinees	(24) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	.05	**.0165** (.0001)	**.0133** (.0017)	**.0053** (.0003)	**.0034** (.0005)	**.0056** (.0001)	**.0037** (.0005)	**.0020** (.0003)
	.01	**.0165** (.0001)	.0050 (.0004)	**.0031** (.0004)	**.0018** (.0002)	**.0030** (.0004)	**.0019** (.0002)	**.0008** (.0002)
	.001	**.0056** (.0001)	**.0027** (.0001)	**.0020** (.0002)	**.0008** (.0002)	**.0016** (.0003)	.0010 (.0002)	**.0003** (.0001)
EDI (C=0)	.05	**.0867** (.0003)	**.0859** (.0003)	**.0862** (.0003)	**.0856** (.0003)	**.0856** (.0004)	**.0852** (.0004)	**.0840** (.0005)
	.01	**.0809** (.0003)	**.0806** (.0003)	**.0805** (.0003)	**.0800** (.0003)	**.0800** (.0004)	**.0796** (.0004)	**.0787** (.0005)
	.001	**.0791** (.0004)	**.0789** (.0003)	**.0788** (.0003)	**.0783** (.0004)	**.0780** (.0004)	**.0777** (.0004)	**.0768** (.0005)
EDI_WTR (C=0)	.05	**.0195** (.0003)	**.0194** (.0003)	**.0180** (.0013)	**.0157** (.0022)	**.0183** (.0012)	**.0194** (.0003)	**.0173** (.0016)
	.01	.0065 (.0002)	.0065 (.0002)	.0060 (.0005)	.0055 (.0007)	.0062 (.0005)	.0067 (.0001)	.0060 (.0005)
	.001	**.0017** (.0001)	**.0017** (.0001)	**.0016** (.0002)	.0015 (.0002)	**.0016** (.0001)	**.0017** (.0001)	**.0016** (.0001)

Table 20.1 Continued

Methods	α	Simulation Conditions						
		Tampering						
			25% schools in 10 districts			50% schools in 10 districts		
			Score-based			Score-based		
		(0) No Tempering	(19) 25% examinees	(20) 50% examinees	(21) 100% examinees	(22) 25% examinees	(23) 50% examinees	(24) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
EDI_WTR (C=0.5)	.05	.0631 (.0003)	.0586 (.0044)	.0634 (.0007)	.0625 (.0003)	.0625 (.0004)	.0577 (.0043)	.0616 (.0004)
	.01	**.0196** (.0002)	**.0181** (.00014)	**.0194** (.0002)	**.0191** (.0002)	**.0183** (.0014)	**.0190** (.0014)	**.0189** (.0002)
	.001	**.0061** (.0002)	**.0058** (.0004)	**.0060** (.0002)	**.0059** (.0002)	**.0061** (.0003)	**.0060** (.0004)	**.0062** (.0001)
$Diff_\theta$	.05	**.1094** (.0033)	**.1057** (.0032)	**.1055** (.0031)	**.1060** (.0032)	**.0994** (.0034)	**.0988** (.0034)	**.0996** (.0035)
	.01	.0505 (.0013)	**.0487** (.0012)	**.0486** (.0012)	**.0489** (.0012)	**.0462** (.00014)	**.0461** (.0014)	**.0465** (.0015)
	.001	**.0240** (.0007)	**.0231** (.0007)	**.0231** (.0006)	**.0232** (.0007)	**.0223** (.0006)	**.0221** (.0006)	**.0224** (.0007)
VJ	.05	.0410 (.0003)	.0395 (.0003)	.0380 (.0004)	.0358 (.0005)	.0383 (.0004)	.0360 (.0005)	.0325 (.0007)
	.01	**.0172** (.0003)	**.0167** (.0002)	**.0161** (.0003)	.0152 (.0003)	**.0160** (.0002)	.0151 (.0003)	.0137 (.0003)
	.001	**.0058** (.0001)	**.0056** (.0001)	**.0054** (.0001)	**.0051** (.0001)	**.0056** (.0001)	**.0052** (.0001)	**.0049** (.0001)

Table 20.1 Continued

Methods	α	Simulation Conditions		
		Tampering		
		100% schools in 10 districts		
		Score-based		
		(25) 25% examinees	(26) 50% examinees	(27) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	.05	**.0042** (.0006)	**.0020** (.0003)	**.0007** (.0002)
	.01	**.0020** (.0003)	**.0009** (.0002)	**.0002** (.0001)
	.001	**.0018** (.0006)	**.0002** (.0001)	**.0000** (.0000)
EDI (C=0)	.05	**.0853** (.0005)	**.0841** (.0005)	**.0819** (.0008)
	.01	**.0798** (.0005)	**.0788** (.0005)	**.0771** (.0007)
	.001	**.0779** (.0005)	**.0769** (.0011)	**.0753** (.0007)
EDI_WTR (C=0)	.05	**.0191** (.0003)	**.0172** (.0005)	**.0173** (.0005)
	.01	**.0066** (.0002)	**.0058** (.0002)	**.0060** (.0002)
	.001	.0015 (.0001)	.0014 (.0001)	.0014 (.0001)

Table 20.1 Continued

Methods	α	Simulation Conditions		
		Tampering		
		100% schools in 10 districts		
		Score-based		
		(25) 25% examinees	(26) 50% examinees	(27) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)
EDI_WTR (C=0.5)	.05	.0618 (.0003)	.0519 (.0061)	.0555 (.0043)
	.01	**.0188** (.0002)	**.0158** (.0016)	**.0159** (.0013)
	.001	**.0061** (.0002)	**.0052** (.0005)	**.0054** (.0003)
$Diff_\theta$	.05	**.0963** (.0036)	**.0950** (.0038)	**.0962** (.0039)
	.01	**.0443** (.0019)	**.0439** (.0020)	**.0447** (.0020)
	.001	**.0205** (.0011)	**.0208** (.0014)	**.0207** (.0012)
VJ	.05	.0355 (.0008)	.0321 (.0010)	.0276 (.0011)
	.01	.0148 (.0003)	.0134 (.0004)	.0116 (.0003)
	.001	**.0051** (.0001)	**.0047** (.0001)	**.0041** (.0002)

Table 20.2: *Power of Individual Detection for Score-Base Tampering by Six Methods at α =.05, .01, and .001*

Methods	α	Simulation Conditions						
			Tampering					
			25% schools in 10 districts			50% schools in 10 districts		
			Score-based			Score-based		
		(0) No Tempering	(19) 25% examinees	(20) 50% examinees	(21) 100% examinees	(22) 25% examinees	(23) 50% examinees	(24) 100% examinees
		Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)
MD	.05	-	.9395	.9119	.8744	.9149	.8838	.8222
		-	(.0092)	(.0074)	(.0095)	(.0067)	(.0128)	(.0147)
	.01	-	.8960	.8672	.8149	.8864	.8266	.7193
		-	(.0095)	(.0112)	(.0120)	(.0093)	(.0128)	(.0177)
	.001	-	-	-	.7370	-	.7467	.5799
		-	-	-	(.0185)	-	(.0200)	(.0504)
EDI (C=0)	.05	-	-	-	-	-	-	-
		-	-	-	-	-	-	-
	.01	-	-	-	-	-	-	-
		-	-	-	-	-	-	-
	.001	-	-	-	-	-	-	-
		-	-	-	-	-	-	-
EDI_WTR (C=0)	.05	-	-	-	-	-	-	-
		-	-	-	-	-	-	-
	.01	-	.7095	.7035	.7256	.7029	.6999	.7130
		-	(.0117)	(.0121)	(.0165)	(.0137)	(.0107)	(.0005)
	.001	-	-	-	.4942	-	-	-
		-	-	-	(.0420)	-	-	-

Table 20.2 Continued

Methods	α	Simulation Conditions						
			Tampering					
			25% schools in 10 districts			50% schools in 10 districts		
			Score-based			Score-based		
		(0) No Tempering	(19) 25% examinees	(20) 50% examinees	(21) 100% examinees	(22) 25% examinees	(23) 50% examinees	(24) 100% examinees
		Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)
EDI_WTR (C=0.5)	.05	- -	.9745 (.0030)	.9748 (.0036)	.9777 (.0018)	.9776 (.0016)	.9815 (.0019)	.9787 (.0014)
	.01	- -	- -	- -	- -	- -	- -	- -
	.001	- -	- -	- -	- -	- -	- -	- -
$Diff_\theta$	.05	- -	- -	- -	- -	- -	- -	- -
	.01	- -	- -	- -	- -	- -	- -	- -
	.001	- -	- -	- -	- -	- -	- -	- -
VJ	.05	- -	.9752 (.0038)	.9763 (.0027)	.9767 (.0018)	.9802 (.0016)	.9812 (.0022)	.9702 (.0019)
	.01	- -	- -	- -	- -	- -	- -	.9461 (.0026)
	.001	- -	- -	- -	- -	- -	- -	- -

Table 20.2 Continued

Methods	α	Simulation Conditions		
		Tampering		
		100% schools in 10 districts		
		Score-based		
		(25) 25% examinees	(26) 50% examinees	(27) 100% examinees
		Power (SE)	Power (SE)	Power (SE)
MD	.05	.8817 (.0156)	.8242 (.00160)	.7019 (.0279)
	.01	.8168 (.0151)	.7358 (.0235)	.5096 (.0830)
	.001	- -	.5454 (.0778)	.2559 (.0812)
EDI (C=0)	.05	- -	- -	- -
	.01	- -	- -	- -
	.001	- -	- -	- -
EDI_WTR (C=0)	.05	- -	- -	- -
	.01	.6980	.6946	.6995

	(.0116)	(.0086)	(.0077)
.001	.4409 (.0168)	.4214 (.0095)	.4321 (.0128)

Table 20.2 Continued

Methods	α	Simulation Conditions		
		Tampering		
		100% schools in 10 districts		
		Score-based		
		(25) 25% examinees	(26) 50% examinees	(27) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)
EDI_WTR (C=0.5)	.05	.9780 (.0017)	.9694 (.0078)	.9766 (.0008)
	.01	- -	- -	- -
	.001	- -	- -	- -
$Diff_\theta$	.05	- -	- -	- -
	.01	- -	- -	- -
	.001	- -	- -	- -
VJ	.05	.9759 (.0028)	.9704 (.0029)	.9518 (.0065)
	.01	.8935 (.0092)	.8862 (.0082)	.9115 (.0118)
	.001	- -	- -	- -

4.6.3 Type I Error Rates and Power of Single-Index School and District Detection Based on Flagged percentages in Score-Based Tampering

This section presents the Type I error rates and power of school and district detection based on the within-school percentages of examinees in score-based tampering. As in fixed-number tampering, only the flagged percentages by MD, EDI (C=0), EDI_WTR (C=.5), $Diff_\theta$, and VJ at α = .05 and .01 during individual detection were used.

Recall that a school would be an outlier according to Tukey's method if its flagged percentage is at least 1.5 IQR greater than the third quartile of all schools. Type I error rates for school detection are reported in Table 21.1, and a controlled error rate or at least a rate decreasing as tampering expanded are preferred. An error rate under a tampering condition would be viewed as being controlled if there is no trend for different conditions and if the value of the error rate lay between $|.5\Omega|$ and $|1.5\Omega|$, where Ω denotes the error rate from no-tampering data sets.

School-level error rates for flagged percentages by MD under multiple tampering conditions were greater than$|1.5\Omega|$, indicating loss of control. In addition, there was no trend in the change of error rates.

Most school-level Type I error rates for VJ were lower than those for no-tampering data sets. However, for the flagged percentages from individual detection at α = .05, the error rates increased when the percent of tampered examinees increased from 25% to 50% in tampering schools. For the flagged percentages from individual detection at α = .01, some uncontrolled error rates were lower than $|.5\Omega|$, and there was no conclusive trend showing that VJ's Type I error rates decreased as more examinees become victims. .

The Type I error rate for EDI_WTR (C=.5) at the school-level under tampering conditions was almost always smaller than Ω. The school-level Type I error rates tended to decrease as tampering was present in more examinees and schools. Type I error rates of EDI (C=0) shared the similar characteristics as those of EDI_WTR (C=.5).

School level Type I error rates for $Diff_\theta$ appeared to decline when more examinees became simulated victims. However, the flagged percentages from individual detection at α = .05 had some error rates that were higher than Ω , when there was 25% of tampered examines in chosen schools. For the flagged percentages from individual detection at α = .01, the decreasing trend of Type I error rates appeared to be constant over conditions.

The power of school detection based on Tukey's outliers are reported in Table 21.2. Only results for EDI_WTR (C=.5), EDI (C=0), and $Diff_\theta$are included, as they showed decreasing Type error rates when more examinees and schools were involved in tampering. All power decreased when the percentages of tampering schools increased, but there was a positive relationship between the power and the percentages of tampered examinees in the schools.

EDI_WTR (C=.5) correctly identified more schools than EDI (C=0) and $Diff_\theta$. This was particularly so for the flagged percentages based on individual flagging at α = .01, for which the power of school detection ranged from .3985 to .9635. The power improved when more percentages of examines were involved tampering. Higher percentages of tampered examinees usually increased the flagged percentages of each tampered school and made these schools more likely be identified as Tukey's outliers. In contrast, high percentages of tampering schools enlarged the IQRs and categorized more tampered school into non-outliers.

For EDI (C=0), the flagged percentage from individual detection at α = .05 yielded school-level power ranging from .0771 to .7455, higher than the ones from individual detection

at α = .01. Like EDI_WTR (C=.5), EDI (C=0) tended to detect more true tampering schools when higher percentages of examinees, albeit in fewer schools, were involved.

It was not clear whether $Diff_\theta$ worked better than EDI (C=0). For the flagged percentage from individual detection by $Diff_\theta$ at α = .01, the power of school detection ranged from .0153 to .6853.

Table 21.1: *Type I Error Rates of Schools Selected as Tukey's Outliers based on the Within-School Percentages of Flagged Examinees at α =.05 and .01 for Score-Based Tampering*

Methods	α at Individual Detection	Simulation Conditions						
			Tampering					
			25% schools in 10 districts			50% schools in 10 districts		
			Score-based			Score-based		
		(0) No Tempering	(19) 25% examinees	(20) 50% examinees	(21) 100% examinees	(22) 25% examinees	(23) 50% examinees	(24) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	[.05]	.0701 (.0020)	.0978 (.0190)	**.1675** (.0123)	**.1151** (.0154)	**.1579** (.0176)	**.1170** (.0192)	.0690 (.0092)
	[.01]	.0701 (.0020)	**.1594** (.0145)	**.1063** (.0131)	.0609 (.0085)	**.1288** (.0141)	.0667 (.0078)	**.0254** (.0053)
EDI (C=0)	[.05]	.0440 (.0026)	.0336 (.0029)	.0315 (.0021)	.0313 (.0021)	.0288 (.0019)	.0260 (.0017)	.0237 (.0017)
	[.01]	.0413 (.0016)	.0384 (.0023)	.0364 (.0028)	.0295 (.0024)	.0363 (.0034)	.0281 (.0031)	.0268 (.0018)
EDI_WTR (C=0.5)	[.05]	.0329 (.0028)	.0271 (.0016)	.0255 (.0014)	.0252 (.0014)	.0240 (.0019)	.0238 (.0021)	.0223 (.0023)
	[.01]	.0728 (.0024)	.0596 (.0019)	.0569 (.0022)	.0565 (.0019)	.0486 (.0029)	.0448 (.0035)	.0450 (.0030)
$Diff_\theta$	[.05]	.0346 (.0023)	**.0399** (.0031)	.0330 (.0016)	.0289 (.0019)	**.0362** (.0024)	.0301 (.0021)	.0236 (.0022)
	[.01]	.0531 (.0008)	.0489 (.0028)	.0416 (.0029)	.0400 (.0027)	.0407 (.0025)	.0316 (.0026)	.0291 (.0027)
VJ	[.05]	.0303 (.0023)	**.0257** (.0016)	**.0265** (.0017)	**.0267** (.0016)	**.0243** (.0017)	**.0250** (.0021)	**.0258** (.0021)
	[.01]	.0622 (.0022)	**.0516** (.0021)	**.0517** (.0025)	**.0530** (.0021)	**.0437** (.0020)	**.0408** (.0028)	**.0425** (.0028)

Table 21.1 Continued

Methods	α at Individual Detection	Simulation Conditions		
		Tampering		
		100% schools in 10 districts		
		Score-based		
		(25) 25% examinees	(26) 50% examinees	(27) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	[.05]	.0858 (.0193)	.0620 (.0137)	.0258 (.0070)
	[.01]	.0710 (.0104)	.0340 (.0084)	.0064 (.0030)
EDI (C=0)	[.05]	.0256 (.0011)	.0195 (.0019)	.0132 (.0025)
	[.01]	.0321 (.0024)	.0200 (.0012)	.0141 (.0025)
EDI_WTR (C=0.5)	[.05]	.0224 (.0032)	.0152 (.0027)	.0119 (.0032)
	[.01]	.0347 (.0040)	.0232 (.0043)	.0213 (.0043)
$Diff_\theta$	[.05]	**.0415** (.0023)	.0281 (.0025)	.0187 (.0028)
	[.01]	.0407 (.0022)	.0263 (.0021)	.0196 (.0032)
VJ	[.05]	.0151 (.0022)	**.0144** (.0025)	**.0136** (.0030)
	[.01]	**.0261** (.0043)	**.0237** (.0042)	**.0219** (.0040)

Table 21.2: *Power of School-Level Detection as Tukey's Outliers based on The Within-School Percentages of Flagged Examinees at α =.05 and .01 for Score-Based Tampering*

Methods	α at Individual Detection	Simulation Conditions						
			Tampering					
			25% schools in 10 districts			50% schools in 10 districts		
			Score-based			Score-based		
		(0) No Tempering	(19) 25% examinees	(20) 50% examinees	(21) 100% examinees	(22) 25% examinees	(23) 50% examinees	(24) 100% examinees
		Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)	Power (SE)
MD	[.05]	-	-	-	-	-	-	-
		-	-	-	-	-	-	-
	[.01]	-	-	-	-	-	-	-
		-	-	-	-	-	-	-
EDI (C=0)	[.05]	-	.1520	.3686	.7455	.0949	.3076	.6669
		-	(.0168)	(.0187)	(.0202)	(.0163)	(.0218)	(.0176)
	[.01]	-	.1237	.2570	.5429	.0908	.2222	.5192
		-	(.0087)	(.0162)	(.0240)	(.0109)	(.0201)	(.0167)
EDI_WTR (C=0.5)	[.05]	-	.2420	.6339	.9104	.2051	.6323	.8933
		-	(.0165)	(.0137)	(.0158)	(.0196)	(.0162)	(.0100)
	[.01]	-	.6365	.8833	.9635	.5939	.8380	.9572
		-	(.0153)	(.0182)	(.0147)	(.0291)	(.0173)	(.0081)
$Diff_\theta$	[.05]	-	-	-	-	-	-	-
		-	-	-	-	-	-	-
	[.01]	-	.0226	.2364	.6853	.0251	.2043	.6453
		-	(.0065)	(.0268)	(.0255)	(.0039)	(.0290)	(.0255)
VJ	[.05]	-	-	-	-	-	-	-
		-	-	-	-	-	-	-
	[.01]	-	-	-	-	-	-	-
		-	-	-	-	-	-	-

Table 21.2 Continued

Methods	α at Individual Detection	Simulation Conditions		
		Tampering		
		100% schools in 10 districts		
		Score-based		
		(25) 25% examinees	(26) 50% examinees	(27) 100% examinees
		Power (SE)	Power (SE)	Power (SE)
MD	[.05]	-	-	-
		-	-	-
	[.01]	-	-	-
		-	-	-
EDI (C=0)	[.05]	.0771	.2256	.4992
		(.0104)	(.0173)	(.0450)
	[.01]	.0872	.1599	.3718
		(.0074)	(.0144)	(.0298)
EDI_WTR (C=0.5)	[.05]	.1231	.4467	.7599
		(.0161)	(.0546)	(.0727)
	[.01]	.3985	.6642	.8308
		(.0543)	(.0721)	(.0783)
$Diff_\theta$	[.05]	-	-	-
		-	-	-
	[.01]	.0153	.0889	.4413
		(.0047)	(.0133)	(.0452)
VJ	[.05]	-	-	-
		-	-	-
	[.01]	-	-	-
		-	-	-

District Type I error rate results are summarized in Table 21.3. The analysis using general linear models used the district as a predictor of the within-school percentage of examines flagged. If the intercept for a district was greater than 0 at $\alpha = .05$, it was considered a suspect.

Most of the time, Type I error rates of all tampering detection methods for district level detection fell below the liberal lower boundary, .025. When only 25% of examinees were victims, some controlled error rates were observed for EDI (C=0), although this was not consistent as all Type I error rates tended to decrease as the tampering increased.

Some error rates of EDI, EDI_WTR, $Diff_\theta$, and VJ appeared controlled, falling between .025 and .0. Relatively large standard errors were observed, however, for those with controlled Type I error rates. This was because the actual error rates of individual data sets were often 0 or less than .025 and occasionally increased to above .10, similar to the findings in fixed-number tampering. Thus, none of the power rates for this approach were interpreted for any of the simulated conditions.

Table 21.3: *Type I Error Rates of Detected Districts as Indicators of School Outliers based on the Within-School Percentages of Flagged Examinees at α =.05 for Score-Based Tampering*

Methods	α at Individual Detection	Simulation Conditions						
			Tampering					
			25% schools in 10 districts			50% schools in 10 districts		
			Score-based			Score-based		
		(0) No Tempering	(19) 25% examinees	(20) 50% examinees	(21) 100% examinees	(22) 25% examinees	(23) 50% examinees	(24) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	[.05]	.0352 (.0131)	**.0175** (.0085)	**.0000** (.0000)	**.0000** (.0000)	**.0135** (.0069)	**.0000** (.0000)	**.0000** (.0000)
EDI (C=0)	[.05]	.0593 (.0361)	.0328 (.0192)	**.0204** (.0181)	**.0068** (.0068)	.0381 (.0131)	**.0205** (.0137)	**.0068** (.0049)
EDI_WTR (C=0.5)	[.05]	**.0241** (.0083)	**.0247** (.0086)	**.0090** (.0037)	**.0000** (.0000)	.0271 (.0125)	**.0136** (.0091)	**.0023** (.0023)
$Diff_\theta$	[.05]	.0333 (.0120)	**.0223** (.0067)	.0313 (.0083)	**.0203** (.0063)	.0407 (.0182)	.0295 (.0170)	**.0159** (.0090)
VJ	[.05]	**.0222** (.0086)	**.0154** (.0057)	**.0022** (.0022)	**.0000** (.0000)	**.0203** (.0124)	**.0045** (.0030)	**.0023** (.0023)

Table 21.3 Continued

Methods	α at Individual Detection	Simulation Conditions		
		Tampering		
		100% schools in 10 districts		
		Score-based		
		(25) 25% examinees	(26) 50% examinees	(27) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)
MD	[.05]	**.0045** (.0030)	**.0000** (.0000)	**.0000** (.0000)
EDI (C=0)	[.05]	.0386 (.0107)	**.0227** (.0096)	**.0068** (.0035)
EDI_WTR (C=0.5)	[.05]	**.0230** (.0096)	**.0206** (.0072)	**.0136** (.0061)
$Diff_\theta$	[.05]	.0296 (.0113)	.0250 (.0110)	**.0137** (.0070)
VJ	[.05]	.0273 (.0088)	**.0114** (.0091)	**.0023** (.0023)

4.6.4 Type I Error Rates and Power of Single-Index School and District Detection Based on Mixed Models for Score-based Tampering

This section presents the Type I error rates and power for school and district detection based on the values of EDI_WTR (C=.5), EDI (C=0), and $Diff_\theta$.

In score-based tampering, values for EDI_WTR (C=.5), EDI (C=0), and $Diff_\theta$,, were treated as the level 1 dependent variable in the mixed models. School level detection was performed at a two-level mixed model (Equations 23 and 24). Table 22.1 summarizes Type I error for the two-level mixed models at α =.05, and .01. All Type I error rates of EDI (C=0) and $Diff_\theta$were below the liberal lower boundary of α, indicating lack of control.

EDI_WTR (C=.5) controlled the Type I error only when 25% schools in chosen districts participated score-based tampering. At α =.01 the controlled error rates presented when 25% and 50% of low achievers in those schools were chosen, As the number of tampering schools and tampered items increased, Type I error rates declined. At α =.05, EDI_WTR (C=.5) only controlled the error rate when 25% of low achievers in those schools were chosen. Beyond these, there was no more controlled error rate.

Power of EDI_WTR (C=.5) was reported in Table 22.2. At α =.01, in the conditions having controlled Type I error rate, the power ranged from .4067 to .7179. At α =.05, in the only condition having controlled Type I error rate, the power was .5068.

District level detection was performed with three-level mixed models. All Type I error rates from mixed models were 0, similar to the results of the previous district direction based on flagged percentages. Thus, no power was reported.

Table 22.1: *School-Level Type I Error Rates in Two-Level Mixed Models at α =.05 and .01 for Score-Based Tampering*

Methods	α	Simulation Conditions						
			Tampering					
			25% schools in 10 districts			50% schools in 10 districts		
			Score-based			Score-based		
		(0) No Tempering	(19) 25% examinees	(20) 50% examinees	(21) 25% examinees	(22) 50% examinees	(23) 25% examinees	(24) 50% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)	Type I error (SE)
EDI_WTR (C=0.5)	**.05**	.0185 (.0022)	.0266 (.0017)	**.0226** (.0021)	**.0154** (.0020)	**.0116** (.0013)	**.0108** (.0016)	**.0077** (.0013)
	.01	.0053 (.0008)	.0063 (.0006)	.0054 (.0008)	**.0038** (.0008)	**.0028** (.0004)	**.0021** (.0005)	**.0024** (.0005)
EDI (C=0)	**.05**	.0000 (.0000)	**.0008** (.0004)	**.0020** (.0008)	**.0033** (.0010)	**.0004** (.0003)	**.0007** (.0004)	**.0010** (.0004)
	.01	.0000 (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0001** (.0001)	**.0000** (.0000)	**.0001** (.0001)	**.0006** (.0001)
$Diff_\theta$	**.05**	.0000 (.0000)	**.0001** (.0001)	**.0003** (.0002)	**.0003** (.0003)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
	.01	.0000 (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)

Table 22.1 Continued

Methods	α	Simulation Conditions		
		Tampering		
		100% schools in 10 districts		
		Score-based		
		(25) 25% examinees	(26) 50% examinees	(27) 100% examinees
		Type I error (SE)	Type I error (SE)	Type I error (SE)
EDI (C=0)	**.05**	**.0067** (.0021)	**.0039** (.0013)	**.0017** (.0008)
	.01	**.0012** (.0005)	**.0008** (.0004)	**.0006** (.0003)
EDI_WTR (C=0.5)	**.05**	**.0003** (.0002)	**.0003** (.0002)	**.0001** (.0001)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
$Diff_\theta$	**.05**	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)
	.01	**.0000** (.0000)	**.0000** (.0000)	**.0000** (.0000)

Table 22.2: *School-Level Power in Two-Level Mixed Models at α =.05 and .01 for Score-Based Tampering*

Methods	α	**Simulation Conditions**		
		Tampering		
		25% schools in 10 districts		
		Score-based		
		(1) 25% examinees	**(2) 50% examinees**	**(3) 100% examinees**
		Type I error (SE)	**Type I error (SE)**	**Type I error (SE)**
EDI (C=0)	**.05**	.5068 (.0188)	- -	- -
	.01	.4067 (.0161)	.7179 (.0196)	- -
EDI_WTR (C=0.5)	**.05**	- -	- -	- -
	.01	- -	- -	- -
$Diff_\theta$	**.05**	- -	- -	- -
	.01	- -	- -	- -

CHAPTER 5

DISCUSSION

This dissertation presented the exploration, demonstration, and evaluation of several data forensic tools based on erasures. This topic has a very limited literature, of which few studies involve both real data sets and simulated ones. Even when they do, power and error analysis of single individual detection indices was usually the focus. The current study aimed to provide a more comprehensive picture with finer details on the characteristics of erasures, and the dynamic between testing data and various analysis techniques.

Three types of erasures were studied: i.e., WR, WW, and RW. Tampering is supposed to create more WR erasures, in particular, when it is caused by the malpractices of educators or administrators, but the other types of erasures could also be actual examinee outcomes. Each provides a part of the puzzle but not a complete picture.

MD is a common erasure analysis method in many states' test security procedures. MD focuses on outliers in the marginal distributions of total or WR erasures. It is straightforward and easily understood by the general public.

EDI, EDI_WTR, $Diff_\theta$, and VJ are all IRT-based approaches and more technically sophisticated than MD. VJ estimates ability based on pre-erasure item responses and then obtains the conditional probability of observing correct post-erasure responses. This is calculated among items that were answered incorrectly before the erasure. VJ then permutes those probabilities in a generalized binomial distribution to reach a composite for all WR erasures. EDI and EDI_WTR obtained "untainted" ability estimates by excluding items having erasures. EDI calculates the

probability of observing all erasures, while EDI_WTR restricts the probability to only WR erasures. Both of these use a normal approximation to the generalized binomial distribution. $Diff_\theta$ judges extreme cases on unusually large gains in ability after the answer has been changed.

Group detection in this study was based on the individual value of tampering indices or the within-group percentages of flagged examinees. Tukey's method, general linear models, and mixed models were applied. Tukey's method for finding outliers is classical and robust, while general linear models and mixed models can accommodate the need for partialling out sources of variation as explanations for WR erasures other than tampering. Also, mixed models improve the estimation of group effects on the values of tampering indices by incorporating information from neighboring groups.

All individual and group level detections were demonstrated in an empirical data set from a large statewide testing program. This also gave insights on how prevalent erasures could be in a real testing situation, how different types of erasures occur together, how the structure of different types of erasures might affect detection results, and how results from different detection methods were similar or different from each other. To understand the power and limits of detection methods, simulated data sets with known tampered responses were created on the basis of the real data. Three factors were manipulated in the simulation: tampering strategies, numbers of schools, and numbers of examinees.

5.1 Discussion on Simulated Data Sets

The simulated datasets were designed to have the distribution of total erasures similar to the empirical one. The generating process was not simply sampling the same portions of examinees then changing certain numbers of answers for them by equal chances. Rather, the data

were generated to create random erasures under the rules of item responses models. Further, the numbers of benign erasures that an examinee created were made to depend on the number of repeated samplings of the examinee's item responses, the chosen tampered item parameters, and the examinee's ability. Given that complexity, results indicated that there were only minor differences in most erasure frequencies between the simulated data and the empirical ones. The major departure was in the frequencies of examinees' answer records with more than five erasures, where it was designed to progressively enlarge that frequency in simulated data.

Comparing the empirical data set, the no-tampering simulated data sets had almost twice as many examinees with no erasures but slightly fewer examinees in all other categories. The other 27 conditions in the simulation study had similar distributions to those in the no-tampering setting, except that more examinees gradually were selected into the category of having more than five erasures as the rate of tampering increased. However, the increase was slow, and the percentages of examinees without erasure were still larger than those in the empirical data. So, the distribution of total erasures in the simulated data was close to the real one for the most part but varied enough to provide a useful test of the robustness of each of the detection methods.

5.2 Discussion on the Interrelationships among Erasures, Pre-Erasure Estimated Ability, and Detection Indices in the Empirical Testing Data Set

In the empirical data, changing answers was a very common behavior, and nine out of ten examinees at least did it once. Most of time the changing was from wrong to right. Over half of examinees made at least one WR erasures, while less than one third made at least one WW or RW erasure. Given the same erasure count, WR erasures always had higher frequencies than WW or RW erasures. This is encouraging, as it suggests that, if examinees are assumed to be

working alone, they may be able to effectively correct wrong answers. This is also challenging, as just knowing that the majority of changes were WR does not tell us whether or not tampering occurred. On the one hand, ample erasures were available to analyze. This is particularly important for use of complex statistical tools such as the ones described in this dissertation. However, without knowing whether or not tampering actually occurred, it appeared to be difficult to distinguish aberrant responses that might be due to breach of test security from the more benign ones.

Correlations between EDI and the three types of erasures were not always linear or positive. EDI had a strongly positive linear relationship with WR erasures, a moderately negative relationship with RW erasures, and a small negative relationship with WW erasures. As a result, the overall linear relationship between EDI and total numbers of erasures was close to 0. This may help explain why EDI didn't flag an examinee who had 22 erasures, consisting of 8 WR erasures, 13 WW erasures, and 1 RW erasure. Correlations between $Diff_\theta$ and the three types of erasures showed similar patterns, which were consistent with the fact that $Diff_\theta$ and EDI were calculated on the basis of three types of erasures and both were strongly correlated.

Correlations between EDI_WTR and the three types of erasures were all positive. EDI_WTR correlated moderately with WR and WW erasures, and marginally with RW erasures. Its correlation with total number of erasures was almost a strong one.

The analysis of this real dataset also showed that examinees with higher ability were more likely to make WR erasures. So, it seems to be reasonable to control for the impact of ability on erasure, when judging which examinees are abnormal. In addition, a much larger portion of examinees making more than six WR erasures or more than seven total erasures were from lower scoring groups. These lower ability examinees tended to have higher values of EDI

(C=-.5), EDI_WTR (C=0), and $Diff_\theta$. This is consistent with the expectation that examinees of lower abilities would be more likely to either engage in cheating or to become victims of test tampering.

5.3 Discussion on Individual Tampering Detection for Both Empirical and Simulated Data Sets

This section presents a comparative discussion on both empirical and simulated results of individual tampering detection.

$Diff_\theta$ flagged the largest number of individual examinees in the empirical data at all α levels. This can be attributed, in part, to the moderate to high power observed in the simulation, which was usually above .5 and could went up to .85. Another reason could be the inflated Type I error rates observed in the simulation study in which too many innocent examinees were flagged by mistake.

In the empirical study, 30% to 45% of the cases detected by $Diff_\theta$ were also outliers detected by the MD of WR erasures. For these cases, the agreement enhanced the signal of irregularities in the response data. The existence of cases flagged by MD of WR erasures but not by $Diff_\theta$ implied that large WR erasures didn't always ensure large positive score gains. In part, this can be attributed to IRT scoring methods, like maximum a posteriori and maximum likelihood estimation, which consider the likelihood of the answering patterns of all items. An examinee's IRT score may not benefit from WR erasures, if most of WR erasures occur on items far below or very close to his or her true proficiency level and very few occur on items far above his or her level. The loss of sensitivity for detecting such cases could be a factor that suppressed the power of $Diff_\theta$ in the simulation study.

Also, $Diff_\theta$ flagged some examinees who had relatively small numbers of WR erasures and who were not flagged by VJ and MD in the empirical data set. These flagged cases by $Diff_\theta$ are likely to be false positive, since $Diff_\theta$'s Type I error rates were highly inflated and always greater than those of VJ and MD in the simulation study. Further, true detections by $Diff_\theta$ were usually smaller than those of VJ, MD of WR erasures, and EDI_WTR(C=.05). A possible reason could be that the school-level correlation between pre-erasure and post-erasure abilities might be smaller than the person-level correlation. Therefore, using the school-level correlation as the substitute for the person-level correlation in Equation 18 might underestimate the standard error of $Diff_\theta$, leading to many false positive detections.

MD of WR erasures were applied in both empirical and simulation study. Flagging for this index was mainly based on the 1.5IQR rule in the empirical study. In the simulation study, on the other hand, flagging was based on a normality assumption in order to allow comparison with other detecting indices. MD of WR erasures based on the 1.5IQRs rule had the second largest individual flagging in the empirical data. Even if switching to the normality assumption, the numbers of flagged examinees by MD of WR erasures were still the second largest. This can be attributed, in part, to the high power observed in the simulation.

However, there were three issues that arose with this index. First, when using a normality assumption to find extreme cases, the Type I error rates in the simulation study were constantly uncontrolled, indicating the inappropriateness of the normality assumption. Therefore, the risk of false positives is likely to depart from the chosen α. Second, when more examinees and schools were involved in tampering, the Type I error of MD decreased rapidly, while those of other methods stayed roughly the same. Third, the power in the simulation was high for 5-item tampering but decreased for 10 item tampering. For score-based tampering, the power of MD

decreased to a greater extent, while the power of other indices became either higher or decreased only slightly compared with what they did for fixed-item tampering. These results suggest that MD based on the normality is more sensitive to the sampling distribution of erasures. Detections can be more easily distorted by examinees with extremely large counts of erasures. Applying the 1.5IQR rule might reduce such sensitivity, but it still did not stabilize the Type I error rates.

MD detected fewer true tampered examinees than VJ across simulation conditions. True detections with MD was also less than EDI_WTR(C=.05) at α=.05 for score-based tampering. Type I error rates for MD were the smallest in the simulation study. So, it was unexpected that the total flagged examinees by MD were greater than those of VJ in the empirical data set. One explanation could be that in reality there were more examinees with large benign erasures due to randomness, misalignment, speededness or other factors which the current simulation missed. If this were the case, it would possibly indicate that the true Type I error rates for MD were actually greater than what were observed in the simulation study.

VJ would appear to be the best method for individual detection. In the simulation study, VJ had controlled Type I error rates at α =.05 and .01 and the highest power (most were greater than 0.97) no matter what tampering strategy was used and what percentages of involved examinees and schools. In the empirical data set, between 73.72% and 100% of the suspected tampering cases detected by VJ were also outliers in both WR erasures and total erasures. Further, these suspected cases were more likely to come from low achieving groups.

Type I error rates for VJ showed a tendency to decrease when more items, examines, and schools were involved in tampering. This may not be a big concern, since the reduction in Type I error rates was slow. The uncontrolled error rates mainly presented when 100% of examinees in

100% of schools in 10 chosen districts were involved in 10-item tampering. This simulation condition is quite extreme and thus likely to be encountered only rarely in practice.

In the simulation study, EDI without any correction consistently had inflated Type I error rates. Adding a negative correction constant (for example, C = -.5) may have restricted Type I error rates but appeared to degrade power, which, without the correction on EDI, this was already the smallest. These results were consistent with the observation in the real data set that EDI had the poorest detection for individuals, and similar to $Diff_\theta$, ignored some examinees who had large numbers of WR erasures and also who were flagged by VJ and MD. One explanation for the low power of EDI could be that EDI positively correlated with WR erasures but negatively correlated with RW and WW erasures, as noted in the previous section.

EDI_WTR without any correction controlled Type I error rates at α =.01, but its error rates fell below the lower bound at α =.05. No clear trend of Type I error rates presented when simulation conditions changed. When the Type I error rates were controlled, the power of EDI_WTR (C=0) was lower than that of VJ. This was consistent with the finding that EDI_WTR flagged fewer suspected cases than did VJ in the real data.

With a correction of C=.5, EDI_WTR still presented controlled Type I error at α=.05, where its power was greater power than EDI_WTR (C=0) and always above .78 at α=.05. The power decreased, when more items, examinees, and schools were contaminated, but increased above .97, when switching from fixed-number tampering to score-based tampering. At α=.01 and .001, the Type I error rate was inflated and the resulting power was not evaluated. In general, the results of this study suggest that EDI_WTR with a continuity correction is another good choice to for individual detection. With regards to continuity correction, since every testing program is unique, it would be more reasonable to use responses and items from the target

testing program to perform tampering simulation and then decide which correction value might be better.

5.4 Discussion on Group Tampering Detection for Both Empirical and Simulated Data Sets

In both empirical and simulation studies, two approaches of school detection were tried out. The first approach was applying Tukey's 1.5IQR rule to the within-school percentages of flagged examines by MD, VJ, EDI, EDI_WTR, and $Diff_\theta$. The 1.5IQR is a conventional albeit arbitrary limit. In a real testing program, it may be more useful to examine real or simulated no-tampering data to determine the test-specific choice for the numbers of IQR. The second approach was applying mixed models to the individual values of EDI, EDI_WTR, and $Diff_\theta$.

MD didn't perform well with the first approach. In the simulation study, the school-level Type I error rates under tampering conditions were not maintained within the acceptable neighborhood of the Type I error rates under the null (i.e., no-tampering) condition. Further, the error rates did not decrease when the numbers of tampered items, examinees or schools increased. So, there was no way to estimate the average or maximum risk of false positive errors when flagging schools with this method in real application. The failure might be attributed to the fact the IRQ of the flagged percentages by MD was almost always zero, generating the lowest bar to judging school outliers and a relatively large number of false detections.

When Tukey's 1.5IQR rule was applied to the flagged percentages by VJ, the results were not satisfactory either. In fixed-number tampering, the school-level Type I error rates tended to decrease, when more schools were involved in tampering. These rates tended to increase, however, as the numbers of tampered items or examinees increased. In score-based

tampering, the school-level Type I error rates also decreased with more tampering schools, and showed no clear trend when tampered examinees increased.

For $Diff_\theta$, EDI (C=0), and EDI_WTR(C=.5), Type I error rates from tampering conditions were less than the ones from no-tampering conditions and generally decreased with more tampered items, examinees, and schools. Therefore, in a real application, the school-level Type I error rates from no-tampering flagged percentages by these detection indices might be viewed as the maximum school-level Type I error rates from potentially tampered data sets. The greater within-school percentages of tampered examinees were also associated with an increased probability of detecting tampering schools, although when more schools were involved in tampering, power decreased. Among the three detection indices, the flagged percentages by EDI_WTR(C=.5) resulted in the highest school-level power, ranging from .43 to 1. This was consistent with the finding from the empirical study in which the flagged percentages by EDI_WTR detected more suspicious schools than the other two.

In the second approach for school detection, two-level mixed models of EDI_WTR (C=.5), EDI (C=0), and $Diff_\theta$ resulted in overly small and uncontrolled Type I error rates in many simulation conditions. Only EDI_WTR(C=.5) had controlled error rates, and most of the time in fixed-number tampering with small numbers of tampered items, examinee victims, and tampering schools. In score-based tampering, the only controlled error rates were from EDI_WTR(C=.5) when 25% of schools in chosen districts were involved. The Type I error rates tended to decline, when increasing numbers of tampered items, examinee victims, or tampering schools, or switching from fixed-number tampering to score-based tampering.

When school-level Type I error rates of mixed models were controlled, EDI_WTR(C=.5) yielded power between .40 and .92 at α =.01, and between .50 and .98 at α =.05. The power

tended to rise when increasing numbers of tampered items, examinee victims, or tampering schools, or switching from score-based tampering to fixed-number tampering. Also, EDI_WTR(C=.5) always made more true detection than EDI (C=0), and $Diff_\theta$. This was consistent with the finding from the empirical study that using school as the predictor of EDI_WTR (C=0) in mixed models or ANOVA flagged more schools. These results support the application of EDI_WTR with mixed modeling for group-level tampering detection.

The district tampering detection also took two approaches. The first one was using district to predict the within-school flagged percentages by MD, VJ, EDI, EDI_WTR, and $Diff_\theta$. The second one was using district as one predictor of EDI, EDI_WTR, and $Diff_\theta$ in three-level mixed models. Both approaches showed the same problem of overly small, even zero Type I error rates. One explanation for this failure could be that the normal assumption in testing whether the district random intercept was 0 was not valid. Alternatively, this might be attributed to the possibility that the simulation setting had too few examinees who had large and benign WR erasures than the real data.

Another issue from the first district-level detection approach is that it was problematic to estimate district means of flagged percentages by averaging out the values from schools of different sizes. A school having small numbers of examinees with erasures in the empirical study, like those with only two examinees, could have a 100-percentage flagging rate. This, in turn, could increase the district's mean flagging percentages to a worrisome level, although the rest of the schools in the same district might be free of flags and take up much greater proportions of examinees.

5.2 Limitation and Future

All simulation results are subject to the settings used in the generated data. The initial item responses and random erasures were generated through random sampling with the NRM, based on item and person parameters estimated from the empirical data set. Parameter estimation errors are always inevitable. The tampering, misalignment, and speededness also introduced further difficulties in model recovery. Simulation study results are sometimes useful but typically not a complete reflection of the complexities of real data. As a result, the fit of VJ, EDI, and EDI_WTR to real data may not be as good in the current study, Type I error rates could increase, and power may decline. This particularly would pose a challenge on VJ, which estimates item parameters from the post-erasure item responses given their initial responses are incorrect. These types of conditional responses could be relatively sparse in a longer test but with few examinees, leading to poor estimation for VJ.

When more schools and examinees were involved in tampering, the numbers of examinees having very large (i.e., >5) numbers of erasures also increased from 2.6% to 15% in the current simulation study and were greater than the 2.59 % observed in the empirical data. The performance of various detection indexes might change in unknown ways in a new simulation condition in which tampering schools and tampered examines are increasing while the percentage of examinees with very large (>5) numbers of erasures remains very low (e.g., lower than 2.6 %, even 1%),

As discussed in the previous sections, it is likely that the current simulated data sets had fewer proportions of examinees with larger numbers of benign erasures than the real data set. Thus, the changes in results might also occur in data sets where greater proportions of examinees have large numbers of benign erasures. The Type I error rates might increase. This is not

necessarily bad, since it actually might serve to bring overly small Type I error rates under control, such as the rates for VJ in individual detection at α =.01 and .001, and the rates for EDI and $Diff_\theta$ in group detection through mixed modeling.

An interesting finding in the simulation study is that Type I error rates of EDI, EDI_WTR and $Diff_\theta$ varied only slightly in individual detection across simulation conditions, although the error rates often departed from chosen α levels. One method to bring them under control could be to adopt non-normal assumptions about the distribution of the indices and figure out the true α levels, since the small variations of those Type I error rates could imply that EDI, EDI_WTR and $Diff_\theta$ actually have some possibly stable non-normal distributions. Under the normal assumption for EDI, EDI_WTR and $Diff_\theta$ in individual detection, null hypothesis tests compare the upper tail probabilities of 1.64, 2.33, and 3.09 times the standard errors with $\alpha = .05, .01,$ and .001. If the distributions are not normal, these probabilities should corresponded to other α levels. And one way to estimate the correct α levels might be to look at the false positive rates of EDI, EDI_WTR and $Diff_\theta$ in individual detection for tampering-free real data sets of targeting test programs.

Finally, VJ and EDI_WTR had good detection performance in this study, although they only considered tampering that results in WR erasures. There are other potential tampering measurements that could be developed to address the problem from different perspectives. For example, RW and WW erasures might also be the result of tampering, if they occur on the same particular items for multiple examinees of a group, and large score differences between pre and post-erasure responses, or between different years, could be part of tampering signals too. And, it doesn't have to be limited to irregular evidence. If an examinee has a consistent excellent score record over multiple years, that could be evidence to ease suspicion due to the false positive

flagging by other indices. In sum, synthesizing positive or negative evidence from different methods or sources may be a way to decrease Type I errors and increase power of detection. Factor analysis or principal components analysis also can simplify this process, for example, by extracting one tampering factor score or component score as a composite signal, based on the underlying correlation structure among multiple evidences. These are topics can be explored in further study.

REFERENCES

Al-Hamly, M., & Coombe, C. (2005). To change or not to change: Investigating the value of MCQ answer changing for Gulf Arab students. *Language Testing*, *22*(1), 509–531.

Amrein-Beardsley, A., Berliner, D. C., & Rideau, S. (2010). Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. *Education Policy Analysis Archives, 18*, 14-49.

Angoff, W. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, *69*(345), 44–49.

Baker, Al (2013, April 11). Allegations of test help by teachers. *The New York Times.* Retrieved from http://www.nytimes.com/.

Bath, J. (1967). Answer-changing behavior on objective examinations. *The Journal of Educational Research*, *6*, 105–107.

Belov, D. I., & Armstrong, R. D. (2010). Automatic detection of answer copying via Kullback–Leibler divergence and K-index. *Applied Psychological Measurement*, *34*, 379–392.

Belov, D. I. (2014) Analysis of answer changes via kullback-leibler divergence. Paper presented at the annual meeting of the Conference on Test Security, Iowa City, Iowa.

Birenbaum, M. (1985). Comparing the effectiveness of several IRT based appropriateness measures in detecting unusual response patterns. *Educational and Psychological Measurement, 45*(3), 523-534.

Blume, H. (2011, June 22). South L.A. charter schools may get a reprieve after cheating scandal. *Los Angeles Times.* Retrieved from http://articles.latimes.com/ Bock, J. (2012, August 17). Crackdown is likely reason for plummeting scores at St. Louis school. *The St. Louis Post-Dispatch.* Retrieved from http://www.stltoday.com/

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Bock, R. D. (1996). Title of chapter. In W. J. van der Linden & R. K. Hambleton, (Eds.), Handbook of modern item response theory (pp.33-49). New York, NY: Springer.

Bradley, J. V. (1978). Robustness?. *British Journal Of Mathematical And Statistical Psychology, 31*(2), 144-152. doi:1.1111/j.2044-8317.1978.tb00581.x

Casella, George, & Berger, Roger L. (2002). *Statistical Inference*. Belmont: Duxbury Press.

Cavalcanti, E. R., Pires, C. E., Cavalcanti, E. P., & Pires, V. F. (2012). Detection and Evaluation of Cheating on College Exams using Supervised Classification. Informatics In Education, *11*(2), 169-190.

cheat. (2012). Merriam-webster.com. Retrieved December 12, 2012, from http://www.merriam-webster.com/dictionary/cheating.

Chen, S. Y. & Wollack, J. A. (in preparation). *Detection of answer copying on multiple-choice tests*. Madison, WI: University of Wisconsin.

Chute, E., & Niederberger, M. (2012, September 9). Inquiry continues into cheating on PSSA tests. *The Pittsburgh Post-Gazette.* Retrieved from http://www.post-gazette.com/

Clark III, J. M., Skorupski, W. P., & Murphy, S. T. (2013) Using nonlinear regression to identify unusual performance level classification rates. Paper presented at the annual meeting of the Conference on Statistical Detection of Potential Test Fraud, Madison, Wisconsin.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In. R. L. Brennan (Ed.) *Educational measurement* (4th ed., pp. 355-386). Westport, CT: American Council on Education/Praeger. Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Copeland, D. A. (1972). Should chemistry students change answers on multiple-choice tests?. *Journal of Chemical Education, 49,* 258–26.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement, 6,* 475-494.

Dawson, Robert (2011). How Significant Is a Boxplot Outlier?. *Journal of Statistics Education, 19*(2).

Emons, W. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*(3), 224-247.

Fernandez, M. (2012, October 13). El paso schools confront scandal of students who 'disappeared' at test time. *The New York Times.* Retrieved from http://www.nytimes.com/

Frary, R. B. (1993). Statistical detection of multiple-choice answer copying: Review and commentary. *Applied Measurement in Education, 6,* 153–165.

Fremer, John J., & Ferrara, Steve (2013). Security in large-scale paper and pencil testing. In James A. Wollack & John J. Fremer (Eds.), *Handbook of Test Security* (pp18). New York: Routledge.

Frysh, P. (2011, August 8). Cheating report confirms teacher's suspicions. *CNN.* Retrieved from http://www.cnn.com/

Geiger, M. (1991a).Changing multiple choice answers: A validation and extension. *College Student Journal, 25,* 181–86.

Gelman, A., Jakulin, A., Pittau, M., & Su, Y. (2009). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics, 2*(4) 1360–1383. doi:10.1214/08-AOAS191

Government Accountability Office (2007, September). No child left behind: Education should clarify guidance and address potential compliance issue for schools in corrective action and restructuring status. Retrieved from http://www.gao.gov/products/GAO-07-1035

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(2), 139-150. doi:10.2307/2086306

Benton, J., & Hacker, H. K., (2004, December 18). Exclusive: Poor schools' TAKS surges raise cheating questions. *The Dallas Morning News*. Retrieved from http://www.parentadvocates.org/nicecontent/dsp_printable.cfm?articleID=5321

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*(3), 133-146. doi:10.1111/j.1745-3984.1981.tb00848.x

Herold, B. (2012, November, 21). Test cheating: Former Philadelphia principal saliya cruz details destructive impact. *The Huffington Post.* Retrieved from http://www.huffingtonpost.com/

Huang, T. W. (2007, July). *Establishing cutoffs for two new aberrance indices: the within-ability-concern index and the beyond-ability-surprise index.* Paper presented at the annual meeting of the International Conference on the Teaching of Psychology, Vancouver, Canada.

Huang, T. (2012). Aberrance detection powers of the BW and person-fit indices. *Educational Technology & Society, 15*(1), 28-37.

Hulin, C. L., Drasgow, F., Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow Jones-Irvin.

Jacob, B. A., & Levitt, S. D. (2003). Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *The Quarterly Journal of Economics, 3*. 843.

Johnson, P. (2011, July 5). America's biggest teacher and principal cheating scandal unfolds in atlanta. *The Christian Science Monitor*. Retrieved from http://www.csmonitor.com/

Josephson Institute of Ethics (2012). 2012 Report Card on the Ethics of American Youth. Retrieved from http://charactercounts.org/pdf/reportcard/2012 /ReportCard-2012-DataTables.pdf

Kao, Shun-chuan, Woo, A., Gorham, J. (2013). *Analysis of ability changes for repeating examinees using growth models*. Paper presented at the annual meeting Statistical Detection of Potential Test Fraud Conference, Madison, WI.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*(4), 277-298.

Li, M. F., & Olejnik, S. (1997). The power of rasch person–fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21*(3), 215-231. doi:10.1177/01466216970213002

Marianti, S., Fox, J., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational & Behavioral Statistics*, *39*(6), 426-451. doi:10.3102/1076998614559412

Mason, R. L., Gunst, R. F., & Hess, J. L. (1989). *Statistical design and analysis of experiments: With applications to engineering and science*. New York: Wiley.

Mathews, J. (2012, February 14). Admissions 101: How badly do school test-tampering scandals hurt college applicants?. *The Washington Post*. Retrieved from http://www.washingtonpost.com/

Mathews, J. (2012, July 03). Baltimore fights cheating. D.C. punts. *The Washington Post.* Retrieved from http://www.washingtonpost.com/.

Maynes, D. D. (2009b, April). Combining statistical evidence for increased power in detecting cheating. Paper Presented at the annual meeting of the National Council of Measurement in Education in San Diego, CA.

McLeod, L. D., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, *27*(2), 121-137. doi:10.1177/0146621602250534.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit, *Applied Psychological Measurement*, *25*(2), 107-135. doi: 10.1177/01466210122031957

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*(1), 3-8. doi:10.1207/s15324818ame0901_2

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

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

Otterman, S. (2011, September 23). State says it analyzed test erasures for cheating; 62 schools proved suspect. *The New York Times.* Retrieved from http://www.nytimes.com/

Pell, M. B. (2012, September 29). More cheating scandals inevitable, as states can't ensure test integrity. *Atlanta Journal-Constitution.* Retrieved from http://www.ajc.com/

Pine, S. M. (1977). Applications of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing: Proceedings of a symposium presented at the 18th annual convention of the Military Testing Association* (Research Rep. No. 77-1, pp. 37-43). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Primoli, V., Liassou, D., Bishop, N. S., & Nhouyvanisvong, A. (2011, April). *Erasure descriptive statistics and covariates,* Paper presented at the annual meeting of the National Council on Measurement in education, New Orleans, LA.

Prinsell, C., Ramsey, P.H. & Ramsey, P.P. (1994). Score gains, attitudes and behavior changes due to answer-changing instruction. *Journal of Educational Measurement, 31*, 327–37.

Qualls, A. L. (2001). Can Knowledge of Erasure Behavior Be Used as an Indicator of Possible Cheating?. *Educational Measurement: Issues and Practice*, *20*(1), 9-16.

Resmovits, J. (2011, July 5). Atlanta public schools shaken by cheating report. *The Huffington Post*.

Richmond, P., & Roehner, B. M. (2015). The detection of cheating in multiple choice examinations. *Physica A: Statistical Mechanics and Its Applications*, *436*, 418-429. doi:10.1016/j.physa.2015.05.040

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science, 6* (1): 15–32.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review 15*, 351–57.

Seo, Songwon (2006). *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets* (Unpublished master's thesis). University of Pittsburgh, Pittsburgh.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66*, 191-208.

Skinner, N. F. (1983). Switching answers on multiple-choice questions: Shrewdness or shibboleth? *Teaching of Psychology, 10*(4), 220-222.

Skorupski, W., & Egan, K., (2012, May). *A hierarchical linear modeling approach for detecting cheating and aberrance*. Paper presented at the annual meeting of the Statistical Detection of Potential Test Fraud Conference, Lawrence, KS.

Snijders, Tom A. B. & Bosker, R. J. (2012). Multilevel analysis (2nd Edition). London: Sage Publishers.

Sotaridona, L. S., van der Linden, W. J., & Meijer, R. R. (2006). Detecting answer copying using the kappa statistic. *Applied Psychological Measurement, 30,* 412–431.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295.

Thiessen, B. A. (2007). *Case study – policies to address educator cheating.* Unpublished manuscript.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 274-286.

Thissen, D. (2003). MULTILOG 7: Multiple, categorical item analysis and test scoring using item response theory [Computer program]. Chicago, IL: Scientific Software.

Toppo, G., Amos, D., Gillum, J., & Upton, J. (2011, March 17). When test scores seem too good to believe. *USA Today.* Retrieved from http://www.usatoday.com/

Tukey, John W. (1977). *Exploratory Data Analysis*. New York: Addison-Wesley.

U.S. Department of Education. (2012). Teacher incentive fund. Retrieved from http://www2.ed.gov/programs/teacherincentive/index.html

Van der Linden, Wim J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics, 37*(1), 180-199. doi: 1.2307/41429217

Van der Linden, W. J., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics, 31*, 283-304.

Verbeke, G., Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer.

Wesolowsky G.O. (2000) "Detecting Excessive Similarity in Answers on Multiple Choice Exams", *Journal of Applied Statistics*, 27, 909-921.

Williams, D. A. (1987), Generalized linear model diagnostics using the deviance and single case deletions, *Applied Statistics*, *36*, 181–191.Zumbo, B. D., & Jennings, M. J. (2002). The

robustness of validity and efficiency of the related samples t-test in the presence of outliers. *Psicológica*, *23*(2), 415-450.

Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement*, *22*, 144–152.

Wollack, J. A., & Maynes, D. (2011). *Detection of test collusion using item response data.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Zhang, Y., Searcy, C. A., & Horn, L. (2011). *Mapping clusters of aberrant patterns in item responses.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

APPENDIX A

AN EXAMPLE OF R SYNTAX FOR GENERATING SIMULATED DATA

```
setwd("F:/REMS/Erasure/eworking/set1a")

#Import data

multi<-
    read.table("F:/REMS/Erasure/Erasure
    Simulation/multi2.txt", header =
    TRUE,sep=",")

multi2<-multi[order(multi$nonthe),]



theta=multi2$nonthe

schID=multi2$schoolID2

disID=multi2$District3

size=multi2$schsi

ID=multi2$ID

ID2=seq(1,35280,1)



thecut<-quantile(theta, .45)

lambda<-
    read.table("F:/REMS/Erasure/Erasure
    programming/sim_lam.txt", header=F,
    sep=",")

zeta<- read.table("F:/REMS/Erasure/Erasure
    programming/sim_zeta.txt", header=F,
    sep=",")

N=35280

T=45

NC=4

tt<-matrix(data=NA,nrow=45,ncol=4)

p<-matrix(data=NA,nrow=45,ncol=4)

P4<-matrix(data=NA,nrow=35280,ncol=1)

x<-matrix(data=NA,nrow=35280,ncol=180)



#for (k in 1:45) {

#  for (l in 1:4) {

#     tt[k,l]<-exp(zeta[k,l]+lambda[k,l]*0.043)

#  }

#  for (l in 1:4) {

#    p[k,l]<-tt[k,l]/(sum(tt[k,]))

#  }

#  p}
```

```
#sum(p[,4])

#[1] 29.68927


for (j in 1:N) {

  for (k in 1:45) {

    for (l in 1:4) {

      tt[k,l]<-
      exp(zeta[k,l]+lambda[k,l]*theta[j])

    }

    for (l in 1:4) {

      p[k,l]<-tt[k,l]/(sum(tt[k,]))

    }

    p      }

  x[j,1:180] <-as.vector(t(p))

  x}


x6<-round(x, digits = 6)




##################Starting
    set####################
```

```
N=35280

item<-matrix(data=0,nrow=N,ncol=45)

eras<-matrix(data=0,nrow=N,ncol=45)

raw<-matrix(data=NA,nrow=N,ncol=1)

tamper<-matrix(data=0,nrow=N,ncol=1)

missa<-matrix(data=0,nrow=N,ncol=1)

speed<-matrix(data=0,nrow=N,ncol=1)

tam<-matrix(data=0,nrow=N,ncol=1)

sam<-matrix(data=NA,nrow=180,ncol=N)

sam2<-matrix(data=NA,nrow=180,ncol=N)

ERA<-matrix(data=0,nrow=N,ncol=1)

ERAT<-matrix(data=0,nrow=N,ncol=1)

rpo<-matrix(data=NA,nrow=N,ncol=180)

rpo2<-matrix(data=NA,nrow=N,ncol=180)


nres<-matrix(data=0,nrow=N,ncol=45)

nres2<-matrix(data=0,nrow=N,ncol=45)


rrap<-runif(N, 0, 1)

erap<-runif(N, 0, 1)


for (j in 1:N) {

  set.seed(j+1)
```

```
for (m in 0:44){
 st=1+m*4
 num=1+m
 st1=st+1
 st2=st+2
 en=st+3
 sam[st:en,j]<-rmultinom(1, 1, prob =
  x6[j,st:en])
 rpo[j,st]=sam[st,j]
 rpo[j,st1]=sam[st1,j]
 rpo[j,st2]=sam[st2,j]
 rpo[j,en]=sam[en,j]


 if(rpo[j,st] == 1) {
  nres[j,num] = 1
 }
 if(rpo[j,st1] == 1) {
  nres[j,num] = 2
 }
 if(rpo[j,st2] == 1) {
  nres[j,num] = 3
 }
 if(rpo[j,en] == 1) {
  nres[j,num] = 4
   }
  }
 }

####################Misalignment and
    Speededness#####################
    ###########


as.numeric(Sys.time())-> t

set.seed(t)


Q1<-sample(seq(1,7056,1), 56, replace =
    FALSE)

q1 <- sample(Q1, 28, replace=FALSE)

qq1<-setdiff(Q1, q1)


Q2<-sample(seq(7056+1,7056*2,1), 56,
    replace = FALSE)

q2 <- sample(Q2, 28, replace=FALSE)

qq2<-setdiff(Q2, q2)


Q3<-sample(seq(7056*2+1,7056*3,1), 56,
    replace = FALSE)

q3 <- sample(Q3, 28, replace=FALSE)

qq3<-setdiff(Q3, q3)
```

```
Q4<-sample(seq(7056*3+1,7056*4,1), 56,
    replace = FALSE)

q4 <- sample(Q4, 28, replace=FALSE)

qq4<-setdiff(Q4, q4)


Q5<-sample(seq(7056*4+1,7056*5,1), 56,
    replace = FALSE)

q5 <- sample(Q5, 28, replace=FALSE)

qq5<-setdiff(Q5, q5)


msam<-c(q1,q2,q3,q4,q5)

misam<-sort(msam)


misrop<-rpo

dim(misrop)


for (j in misam){
  missa[j,1]=1
  Lm<-rbinom(1, 45, 0.2)
  ta<-sample(seq(1,44,1), 1, replace =
    FALSE)
  max=ta+Lm-1
  if(max<45){
  for (m in ta:max){
   m1=1+(m-1)*4
   m4=m1+3
   om1=m4+1
   om4=m4+4
   misrop[j,m1:m4]=rpo[j,om1:om4]
   }
  }
  if(max >= 45){
   for (m in ta:44){
   m1=1+(m-1)*4
   m4=m1+3
   om1=m4+1
   om4=m4+4
   misrop[j,m1:m4]=rpo[j,om1:om4]
   }
  }
}


ssam<-c(qq1,qq2,qq3,qq4,qq5)
stsam<-sort(ssam)


for (j in stsam){
 speed[j,1]=1
```

```
  for (m in 37:45){
    m1=1+(m-1)*4
    m4=m1+3
    sam2[m1:m4,j]<-rmultinom(1, 1, prob
      =c(.25,.25,.25,.25))
    misrop[j,m1]=sam2[m1,j]
    misrop[j,m1+1]=sam2[m1+1,j]
    misrop[j,m1+2]=sam2[m1+2,j]
    misrop[j,m4]=sam2[m4,j]
  }
}



####################Random
    Erasures##########################
    #########


ERA2<-matrix(data=0,nrow=N,ncol=1)
ERAT2<-matrix(data=0,nrow=N,ncol=1)


MIS<-c(misam,stsam)
RAM<-ID2[-MIS]



t<-as.numeric(Sys.time())
set.seed(t)


rpo2<-rpo


for (j in 1:N) {
  set.seed(j+t)
  choice<-sample(seq(1,45,1), 29, replace =
    FALSE)



    st1=1+(choice[1]-1)*4
    st11=st1+1
    st12=st1+2
    en1=st1+3
    sam2[st1:en1,j]<-rmultinom(1, 1, prob =
      x6[j,st1:en1])
    rpo2[j,st1]=sam2[st1,j]
    rpo2[j,st11]=sam2[st11,j]
    rpo2[j,st12]=sam2[st12,j]
    rpo2[j,en1]=sam2[en1,j]


    if(rrap[j]>.3266){
    st2=1+(choice[2]-1)*4
    en2=st2+3
```

```
st21=st2+1

st22=st2+2

sam2[st2:en2,j]<-rmultinom(1, 1, prob =
  x6[j,st2:en2])

rpo2[j,st2]=sam2[st2,j]

rpo2[j,st21]=sam2[st21,j]

rpo2[j,st22]=sam2[st22,j]

rpo2[j,en2]=sam2[en2,j]

}


if(rrap[j]>.7266){

st3=1+(choice[3]-1)*4

st31=st3+1

st32=st3+2

en3=st3+3

sam2[st3:en3,j]<-rmultinom(1, 1, prob =
  x6[j,st3:en3])

rpo2[j,st3]=sam2[st3,j]

rpo2[j,st31]=sam2[st31,j]

rpo2[j,st32]=sam2[st32,j]

rpo2[j,en3]=sam2[en3,j]

}


if(rrap[j]>.848){

st4=1+(choice[4]-1)*4

st41=st4+1

st42=st4+2

en4=st4+3

sam2[st4:en4,j]<-rmultinom(1, 1, prob =
  x6[j,st4:en4])

rpo2[j,st4]=sam2[st4,j]

rpo2[j,st41]=sam2[st41,j]

rpo2[j,st42]=sam2[st42,j]

rpo2[j,en4]=sam2[en4,j]

}


if(rrap[j]>0.8741){

st5=1+(choice[5]-1)*4

st51=st5+1

st52=st5+2

en5=st5+3

sam2[st5:en5,j]<-rmultinom(1, 1, prob =
  x6[j,st5:en5])

rpo2[j,st5]=sam2[st5,j]

rpo2[j,st51]=sam2[st51,j]

rpo2[j,st52]=sam2[st52,j]

rpo2[j,en5]=sam2[en5,j]

}
```

```
if(rrap[j]>0.8859){
st6=1+(choice[6]-1)*4
st61=st6+1
st62=st6+2
en6=st6+3
sam2[st6:en6,j]<-rmultinom(1, 1, prob =
  x6[j,st6:en6])
rpo2[j,st6]=sam2[st6,j]
rpo2[j,st61]=sam2[st61,j]
rpo2[j,st62]=sam2[st62,j]
rpo2[j,en6]=sam2[en6,j]
}

if(rrap[j]>0.9923){
st7=1+(choice[7]-1)*4
st71=st7+1
st72=st7+2
en7=st7+3
sam2[st7:en7,j]<-rmultinom(1, 1, prob =
  x6[j,st7:en7])
rpo2[j,st7]=sam2[st7,j]
rpo2[j,st71]=sam2[st71,j]
rpo2[j,st72]=sam2[st72,j]
rpo2[j,en7]=sam2[en7,j]
}

if(rrap[j]>0.9953){
st8=1+(choice[8]-1)*4
st81=st8+1
st82=st8+2
en8=st8+3
sam2[st8:en8,j]<-rmultinom(1, 1, prob =
  x6[j,st8:en8])
rpo2[j,st8]=sam2[st8,j]
rpo2[j,st81]=sam2[st81,j]
rpo2[j,st82]=sam2[st82,j]
rpo2[j,en8]=sam2[en8,j]
}

if(rrap[j]>0.9974){
st9=1+(choice[9]-1)*4
st91=st9+1
st92=st9+2
en9=st9+3
sam2[st9:en9,j]<-rmultinom(1, 1, prob =
  x6[j,st9:en9])
rpo2[j,st9]=sam2[st9,j]
```

```
rpo2[j,st91]=sam2[st91,j]

rpo2[j,st92]=sam2[st92,j]

rpo2[j,en9]=sam2[en9,j]

}

if(rrap[j]>0.9985){

st10=1+(choice[10]-1)*4

st101=st10+1

st102=st10+2

en10=st10+3

sam2[st10:en10,j]<-rmultinom(1, 1, prob
 = x6[j,st10:en10])

rpo2[j,st10]=sam2[st10,j]

rpo2[j,st101]=sam2[st101,j]

rpo2[j,st102]=sam2[st102,j]

rpo2[j,en10]=sam2[en10,j]

}

if(rrap[j]>0.9989){

st11=1+(choice[11]-1)*4

st111=st11+1

st112=st11+2

en11=st11+3

sam2[st11:en11,j]<-rmultinom(1, 1, prob
 = x6[j,st11:en11])

rpo2[j,st11]=sam2[st11,j]

rpo2[j,st111]=sam2[st111,j]

rpo2[j,st112]=sam2[st112,j]

rpo2[j,en11]=sam2[en11,j]

}


if(rrap[j]>0.9990){

st12=1+(choice[12]-1)*4

st121=st12+1

st122=st12+2

en12=st12+3

sam2[st12:en12,j]<-rmultinom(1, 1, prob
 = x6[j,st12:en12])

rpo2[j,st12]=sam2[st12,j]

rpo2[j,st121]=sam2[st121,j]

rpo2[j,st122]=sam2[st122,j]

rpo2[j,en12]=sam2[en12,j]

}


if(rrap[j]>0.9992){

st13=1+(choice[13]-1)*4

st131=st13+1

st132=st13+2

en13=st13+3
```

```
sam2[st13:en13,j]<-rmultinom(1, 1, prob
  = x6[j,st13:en13])

rpo2[j,st13]=sam2[st13,j]

rpo2[j,st131]=sam2[st131,j]

rpo2[j,st132]=sam2[st132,j]

rpo2[j,en13]=sam2[en13,j]

}


if(rrap[j]>0.9992){

st14=1+(choice[14]-1)*4

st141=st14+1

st142=st14+2

en14=st14+3

sam2[st14:en14,j]<-rmultinom(1, 1, prob
  = x6[j,st14:en14])

rpo2[j,st14]=sam2[st14,j]

rpo2[j,st141]=sam2[st141,j]

rpo2[j,st142]=sam2[st142,j]

rpo2[j,en14]=sam2[en14,j]

}


if(rrap[j]>0.9992){

st15=1+(choice[15]-1)*4

st151=st15+1

st152=st15+2

en15=st15+3

sam2[st15:en15,j]<-rmultinom(1, 1, prob
  = x6[j,st15:en15])

rpo2[j,st15]=sam2[st15,j]

rpo2[j,st151]=sam2[st151,j]

rpo2[j,st152]=sam2[st152,j]

rpo2[j,en15]=sam2[en15,j]

}


if(rrap[j]>0.9992){

st16=1+(choice[16]-1)*4

st161=st16+1

st162=st16+2

en16=st16+3

sam2[st16:en16,j]<-rmultinom(1, 1, prob
  = x6[j,st16:en16])

rpo2[j,st16]=sam2[st16,j]

rpo2[j,st161]=sam2[st161,j]

rpo2[j,st162]=sam2[st162,j]

rpo2[j,en16]=sam2[en16,j]


st17=1+(choice[17]-1)*4

st171=st17+1
```

```
st172=st17+2

en17=st17+3

sam2[st17:en17,j]<-rmultinom(1, 1, prob
  = x6[j,st17:en17])

rpo2[j,st17]=sam2[st17,j]

rpo2[j,st171]=sam2[st171,j]

rpo2[j,st172]=sam2[st172,j]

rpo2[j,en17]=sam2[en17,j]

}


if(rrap[j]>0.9999){

st18=1+(choice[18]-1)*4

st181=st18+1

st182=st18+2

en18=st18+3

sam2[st18:en18,j]<-rmultinom(1, 1, prob
  = x6[j,st18:en18])

rpo2[j,st18]=sam2[st18,j]

rpo2[j,st181]=sam2[st181,j]

rpo2[j,st182]=sam2[st182,j]

rpo2[j,en18]=sam2[en18,j]


st19=1+(choice[19]-1)*4

st191=st19+1

st192=st19+2

en19=st19+3

sam2[st19:en19,j]<-rmultinom(1, 1, prob
  = x6[j,st19:en19])

rpo2[j,st19]=sam2[st19,j]

rpo2[j,st191]=sam2[st191,j]

rpo2[j,st192]=sam2[st192,j]

rpo2[j,en19]=sam2[en19,j]


st20=1+(choice[20]-1)*4

st201=st20+1

st202=st20+2

en20=st20+3

sam2[st20:en20,j]<-rmultinom(1, 1, prob
  = x6[j,st20:en20])

rpo2[j,st20]=sam2[st20,j]

rpo2[j,st201]=sam2[st201,j]

rpo2[j,st202]=sam2[st202,j]

rpo2[j,en20]=sam2[en20,j]


st21=1+(choice[21]-1)*4

st211=st21+1

st212=st21+2

en21=st21+3
```

```
sam2[st21:en21,j]<-rmultinom(1, 1, prob
 = x6[j,st21:en21])

rpo2[j,st21]=sam2[st21,j]

rpo2[j,st211]=sam2[st211,j]

rpo2[j,st212]=sam2[st212,j]

rpo2[j,en21]=sam2[en21,j]


st22=1+(choice[22]-1)*4

st221=st22+1

st222=st22+2

en22=st22+3

sam2[st22:en22,j]<-rmultinom(1, 1, prob
 = x6[j,st22:en22])

rpo2[j,st22]=sam2[st22,j]

rpo2[j,st221]=sam2[st221,j]

rpo2[j,st222]=sam2[st222,j]

rpo2[j,en22]=sam2[en22,j]

}


if(rrap[j]>0.99995){

st23=1+(choice[23]-1)*4

st231=st23+1

st232=st23+2

en23=st23+3

sam2[st23:en23,j]<-rmultinom(1, 1, prob
 = x6[j,st23:en23])

rpo2[j,st23]=sam2[st23,j]

rpo2[j,st231]=sam2[st231,j]

rpo2[j,st232]=sam2[st232,j]

rpo2[j,en23]=sam2[en23,j]


st24=1+(choice[24]-1)*4

st241=st24+1

st242=st24+2

en24=st24+3

sam2[st24:en24,j]<-rmultinom(1, 1, prob
 = x6[j,st24:en24])

rpo2[j,st24]=sam2[st24,j]

rpo2[j,st241]=sam2[st241,j]

rpo2[j,st242]=sam2[st242,j]

rpo2[j,en24]=sam2[en24,j]


st25=1+(choice[25]-1)*4

st251=st25+1

st252=st25+2

en25=st25+3

sam2[st25:en25,j]<-rmultinom(1, 1, prob
 = x6[j,st25:en25])

rpo2[j,st25]=sam2[st25,j]
```

```
rpo2[j,st251]=sam2[st251,j]

rpo2[j,st252]=sam2[st252,j]

rpo2[j,en25]=sam2[en25,j]


st26=1+(choice[26]-1)*4

st261=st26+1

st262=st26+2

en26=st26+3

sam2[st26:en26,j]<-rmultinom(1, 1, prob
  = x6[j,st26:en26])

rpo2[j,st26]=sam2[st26,j]

rpo2[j,st261]=sam2[st261,j]

rpo2[j,st262]=sam2[st262,j]

rpo2[j,en26]=sam2[en26,j]

}


if(rrap[j]>0.99995){

st27=1+(choice[27]-1)*4

st271=st27+1

st272=st27+2

en27=st27+3

sam2[st27:en27,j]<-rmultinom(1, 1, prob
  = x6[j,st27:en27])

rpo2[j,st27]=sam2[st27,j]

rpo2[j,st271]=sam2[st271,j]

rpo2[j,st272]=sam2[st272,j]

rpo2[j,en27]=sam2[en27,j]


st28=1+(choice[28]-1)*4

st281=st28+1

st282=st28+2

en28=st28+3

sam2[st28:en28,j]<-rmultinom(1, 1, prob
  = x6[j,st28:en28])

rpo2[j,st28]=sam2[st28,j]

rpo2[j,st281]=sam2[st281,j]

rpo2[j,st282]=sam2[st282,j]

rpo2[j,en28]=sam2[en28,j]


st29=1+(choice[29]-1)*4

st291=st29+1

st292=st29+2

en29=st29+3

sam2[st29:en29,j]<-rmultinom(1, 1, prob
  = x6[j,st29:en29])

rpo2[j,st29]=sam2[st29,j]

rpo2[j,st291]=sam2[st291,j]

rpo2[j,st292]=sam2[st292,j]
```

```
   rpo2[j,en29]=sam2[en29,j]
   }
}


ERA<-matrix(data=0,nrow=N,ncol=1)
ERAT<-matrix(data=0,nrow=N,ncol=1)
nres2<-matrix(data=0,nrow=N,ncol=45)

for (j in 1:N) {
  for (n in 0:44){
   st=1+n*4
   num=1+n
   st1=st+1
   st2=st+2
   en=st+3
  if(rpo2[j,st] == 1) {
   nres2[j,num] = 1
  }
  if(rpo2[j,st1] == 1) {
   nres2[j,num] = 2
  }
  if(rpo2[j,st2] == 1) {
   nres2[j,num] = 3
  }
  if(rpo2[j,en] == 1) {
   nres2[j,num] = 4
  }

  if( nres[j,num] != nres2[j,num] ){
    ERA[j,1]=1
    ERAT[j,1]=ERAT[j,1]+1
  }
 }

 if (ERAT[j,1] == 0) {
  t<-as.numeric(Sys.time())
  set.seed(t)
  cho<-sample(seq(1,45,1), 1, replace =
   FALSE)
  st1=1+(cho-1)*4
  st11=st1+1
  st12=st1+2
  en1=st1+3
  sam2[st1:en1,j]<-rmultinom(1, 1, prob =
   x6[j,st1:en1])
  rpo2[j,st1]=sam2[st1,j]
```

```
   rpo2[j,st11]=sam2[st11,j]
   rpo2[j,st12]=sam2[st12,j]
   rpo2[j,en1]=sam2[en1,j]
  }

}

ERA<-matrix(data=0,nrow=N,ncol=1)
ERAT<-matrix(data=0,nrow=N,ncol=1)
nres2<-matrix(data=0,nrow=N,ncol=45)

for (j in 1:N) {
  for (n in 0:44){
   st=1+n*4
   num=1+n
   st1=st+1
   st2=st+2
   en=st+3
  if(rpo2[j,st] == 1) {
   nres2[j,num] = 1
  }
  if(rpo2[j,st1] == 1) {
   nres2[j,num] = 2
  }
 if(rpo2[j,st2] == 1) {
  nres2[j,num] = 3
 }
 if(rpo2[j,en] == 1) {
  nres2[j,num] = 4
 }
 if( nres[j,num] != nres2[j,num] ){
   ERA[j,1]=1
   ERAT[j,1]=ERAT[j,1]+1
 }
}

 if (ERAT[j,1] == 0) {
 t<-as.numeric(Sys.time())
 set.seed(t)
 cho<-sample(seq(1,45,1), 1, replace =
  FALSE)
 st1=1+(cho-1)*4
 st11=st1+1
 st12=st1+2
 en1=st1+3
 sam2[st1:en1,j]<-rmultinom(1, 1, prob =
  x6[j,st1:en1])
```

```
   rpo2[j,st1]=sam2[st1,j]
   rpo2[j,st11]=sam2[st11,j]
   rpo2[j,st12]=sam2[st12,j]
   rpo2[j,en1]=sam2[en1,j]
  }
}

ERA<-matrix(data=0,nrow=N,ncol=1)
ERAT<-matrix(data=0,nrow=N,ncol=1)
nres1<-matrix(data=0,nrow=N,ncol=45)
nres2<-matrix(data=0,nrow=N,ncol=45)

for (j in 1:N) {
  for (n in 0:44){
   st=1+n*4
   num=1+n
   st1=st+1
   st2=st+2
   en=st+3
  if(rpo2[j,st] == 1) {
   nres2[j,num] = 1
  }
  if(rpo2[j,st1] == 1) {
   nres2[j,num] = 2
  }
  if(rpo2[j,st2] == 1) {
   nres2[j,num] = 3
  }
  if(rpo2[j,en] == 1) {
   nres2[j,num] = 4
  }

  if(misrop[j,st] == 1) {
   nres1[j,num] = 1
  }
  if(misrop[j,st1] == 1) {
   nres1[j,num] = 2
  }
  if(misrop[j,st2] == 1) {
   nres1[j,num] = 3
  }
  if(misrop[j,en] == 1) {
   nres1[j,num] = 4
  }

  if( nres[j,num] != nres2[j,num] ){
```

```
    ERA[j,1]=1
    ERAT[j,1]=ERAT[j,1]+1
  }
 }
}

round(table(ERAT)/N, digits = 4)



##############Tampering districts and 25% schools####################
dsize<-cbind(schID,disID)
dim(dsize)
unique<-dsize[!duplicated(dsize),]


unique1<-as.data.frame(unique)
unique2<-unique1[order(unique1$disID),]
dim(unique2)



t1<-table(unique2$disID)
disize<-cbind(as.numeric(names(t1)),as.numeric(t1))
tendis<-sample(as.numeric(names(t1)),10)

colnames(disize) <- c("disID","feq")
mque2<-merge(unique2, disize, by="disID")
dissch<-matrix(data=0,nrow=905,ncol=1)
tamsch<-matrix(data=0,nrow=905,ncol=1)


for (z in 1:905){
  if( unique2$disID[z]==1 &&
      z == 1) {
      dissch[z]=1
    }
  if( unique2$disID[z]==1 &&
      z !=1) {
      dissch[z]=1+dissch[z-1]
    }
  if( unique2$disID[z]!=1 &&
      unique2$disID[z] != unique2$disID[z-1]) {
      dissch[z]=1
    }
  if( unique2$disID[z]!=1 &&
      unique2$disID[z] == unique2$disID[z-1]) {
      dissch[z]=1+dissch[z-1]
    }
```

```
}


mque2<-cbind(mque2,dissch)

t<-as.numeric(Sys.time())


for (z in 1:905){

  set.seed(mque2$disID[z]+t)

  schoice<-
    sample(1:mque2$feq[z],as.integer(mque
    2$feq[z]/4))

  if( mque2$disID[z] %in% tendis &&

      mque2$dissch[z] %in% schoice) {

      tamsch[z]=1

  }

}



mque3<-cbind(mque2,tamsch)

sesch<-mque3$schID[mque3$tamsch==1]



write.table(mque3,
    paste(606,c("sesch25.dat"), sep=""),

        quote=FALSE, sep="\t", row.names
    =F,col.names =F)
```

```
##############25%/50%/100%
    Tampering
    Students####################


Ssize<-cbind(ID2,schID,speed,missa)

dim(Ssize)

colnames(Ssize) <-
    c("ID2","schID","speed","missa")

#unique3<-Ssize[!duplicated(Ssize),]

unique3<-as.data.frame(Ssize)


unique4<-unique3[order(unique3$schID),]

dim(unique4)


t1<-table(unique4$schID)

SSsize<-
    cbind(as.numeric(names(t1)),as.numeric(
    t1))

colnames(SSsize) <- c("schID","feq")

mque4<-merge(unique4, SSsize,
    by="schID")

dim(mque4)

mque4[1:10,]


schper<-matrix(data=0,nrow=N,ncol=1)

tamper25<-matrix(data=0,nrow=N,ncol=1)
```

```
tamper50<-matrix(data=0,nrow=N,ncol=1)

tamper100<-matrix(data=0,nrow=N,ncol=1)


for (z in 1:N){

  if( z == 1) {

     schper[z]=1

    }

  if( unique4$schID[z] == unique4$schID[1]
    &&

     z !=1) {

     schper[z]=1+schper[z-1]

    }

  if( unique4$schID[z] != unique4$schID[1]
    &&

     unique4$schID[z] != unique4$schID[z-
    1]) {

     schper[z]=1

    }

  if( unique4$schID[z] != unique4$schID[1]
    &&

      unique4$schID[z] ==
    unique4$schID[z-1]) {

     schper[z]=1+schper[z-1]

    }

}
```

```
mque4<-cbind(mque4,schper)

t<-as.numeric(Sys.time())


for (z in 1:N){

  set.seed(mque4$schID[z]+t)

  pchoice1<-
    sample(1:mque4$feq[z],as.integer(mque
    4$feq[z]/4))

  pchoice2<-
    sample(1:mque4$feq[z],as.integer(mque
    4$feq[z]/2))

  pchoice3<-
    sample(1:mque4$feq[z],as.integer(mque
    4$feq[z]))


  if( mque4$speed[z] !=1 &&
    mque4$missa[z] !=1 &&

     unique4$schID[z] %in% sesch &&

     mque4$schper[z] %in% pchoice1) {

     tamper25[z]=1

  }

  if( mque4$speed[z] !=1 &&
    mque4$missa[z] !=1 &&

     unique4$schID[z] %in% sesch &&

     mque4$schper[z] %in% pchoice2) {

     tamper50[z]=1
```

```
  }

  if( mque4$speed[z] !=1 &&
    mque4$missa[z] !=1 &&

     unique4$schID[z] %in% sesch &&

     mque4$schper[z] %in% pchoice3) {

     tamper100[z]=1

  }

}

dim(mque4)

mque5<-
    cbind(mque4,tamper25,tamper50,tamper
    100)

t25<-mque5$ID2[mque5$tamper25==1]

t50<-mque5$ID2[mque5$tamper50==1]

t100<-mque5$ID2[mque5$tamper100==1]

ttt3<-union(t25,t50)

ttt3<-union(ttt3,t100)

ttt3<-sort(ttt3)

length(ttt3)

ttt3 <- ttt3[!duplicated(ttt3)]

table(tamper25)

table(mque5$tamper25)


############## 5-item fix tampering
    ####################
```

```
fix25<-rpo

fix50<-rpo

fix100<-rpo


tamper25<-matrix(data=0,nrow=N,ncol=1)

tamper50<-matrix(data=0,nrow=N,ncol=1)

tamper100<-matrix(data=0,nrow=N,ncol=1)


for (j in ttt3){

   witem=0

   FL=0


   for (n in 0:44){

    st=1+n*4

    num=1+n

    en=st+3

    if ( rpo[j,en] == 0 ) {

    witem<-rbind(witem,num)

    }

   }


  wrong<-data.frame(witem)

  twro<-wrong[witem>0]
```

```
FL<-length(twro)


if (any(t25 ==ID2[j]) && FL > 0 && FL
  <= 5  ) {
 tamper25[j]=1
for (n in 0:44){
 st=1+n*4
 num=1+n
 en=st+3
 fix25[j,st]=0
 fix25[j,st+1]=0
 fix25[j,st+2]=0
 fix25[j,en]=1
}
}


if (any(t25 ==ID2[j]) && FL > 5) {
 tamper25[j]=1
FS<-sample(twro, 5, replace = FALSE)
SFS<-sort(FS)
for (n in SFS) {
 st=1+(n-1)*4
 en=st+3
 fix25[j,st]=0
 fix25[j,st+1]=0
 fix25[j,st+2]=0
 fix25[j,en]=1
 }
 }


if (any(t50 ==ID2[j]) && FL > 0 && FL
  <= 5  ) {
tamper50[j]=1
for (n in 0:44){

 st=1+n*4
 num=1+n
 en=st+3
 fix50[j,st]=0
 fix50[j,st+1]=0
 fix50[j,st+2]=0
 fix50[j,en]=1
 }
 }

if (any(t50 ==ID2[j]) && FL > 5) {
tamper50[j]=1
FS<-sample(twro, 5, replace = FALSE)
```

```
SFS<-sort(FS)
for (n in SFS) {
 st=1+(n-1)*4
 en=st+3
 fix50[j,st]=0
 fix50[j,st+1]=0
 fix50[j,st+2]=0
 fix50[j,en]=1
 }
 }


if (any(t100 ==ID2[j]) && FL > 0 && FL
  <= 5  ) {
tamper100[j]=1
for (n in 0:44){
 st=1+n*4
 num=1+n
 en=st+3
 fix100[j,st]=0
 fix100[j,st+1]=0
 fix100[j,st+2]=0
 fix100[j,en]=1
}
}
  if (any(t100 ==ID2[j]) && FL > 5) {
  tamper100[j]=1
  FS<-sample(twro, 5, replace = FALSE)
  SFS<-sort(FS)
  for (n in SFS) {
   st=1+(n-1)*4
   en=st+3
   fix100[j,st]=0
   fix100[j,st+1]=0
   fix100[j,st+2]=0
   fix100[j,en]=1
   }
   }


}


############## 10-item fix tampering
    ####################


fix225<-rpo
fix250<-rpo
fix2100<-rpo
```

```
tamper225<-matrix(data=0,nrow=N,ncol=1)
tamper250<-matrix(data=0,nrow=N,ncol=1)
tamper2100<-matrix(data=0,nrow=N,ncol=1)

for (j in ttt3){
   witem=0
   FL=0

   for (n in 0:44){
    st=1+n*4
    num=1+n
    en=st+3
    if (  rpo[j,en] == 0 ) {
    witem<-rbind(witem,num)
    }
   }

  wrong<-data.frame(witem)
  twro<-wrong[witem>0]
  FL<-length(twro)

if (any(t25 ==ID2[j]) && FL > 0 && FL
  <= 10  ) {
tamper225[j]=1
for (n in 0:44){
 st=1+n*4
 num=1+n
 en=st+3
 fix225[j,st]=0
 fix225[j,st+1]=0
 fix225[j,st+2]=0
 fix225[j,en]=1
}
}

if (any(t25 ==ID2[j]) && FL > 10) {
tamper225[j]=1
FS<-sample(twro, 10, replace = FALSE)
SFS<-sort(FS)
for (n in SFS) {
 st=1+(n-1)*4
 en=st+3
 fix225[j,st]=0
 fix225[j,st+1]=0
 fix225[j,st+2]=0
```

```
 fix225[j,en]=1
 }
 }

if (any(t50 ==ID2[j]) && FL > 0 && FL
  <= 10  ) {
tamper250[j]=1
for (n in 0:44){
 st=1+n*4
 num=1+n
 en=st+3
 fix250[j,st]=0
 fix250[j,st+1]=0
 fix250[j,st+2]=0
 fix250[j,en]=1
}
}

if (any(t50 ==ID2[j]) && FL > 10) {
tamper250[j]=1
FS<-sample(twro, 10, replace = FALSE)
SFS<-sort(FS)
for (n in SFS) {
 st=1+(n-1)*4
 en=st+3
 fix250[j,st]=0
 fix250[j,st+1]=0
 fix250[j,st+2]=0
 fix250[j,en]=1
 }
 }

if (any(t100 ==ID2[j]) && FL > 0 && FL
  <= 10  ) {
tamper2100[j]=1
for (n in 0:44){
 st=1+n*4
 num=1+n
 en=st+3
 fix2100[j,st]=0
 fix2100[j,st+1]=0
 fix2100[j,st+2]=0
 fix2100[j,en]=1
}
}

if (any(t100 ==ID2[j]) && FL > 10) {
tamper2100[j]=1
```

```
  FS<-sample(twro, 10, replace = FALSE)
  SFS<-sort(FS)
  for (n in SFS) {
   st=1+(n-1)*4
   en=st+3
   fix2100[j,st]=0
   fix2100[j,st+1]=0
   fix2100[j,st+2]=0
   fix2100[j,en]=1
   }
   }
}


##############25%/50%/100%
    Tampering
    Students( Low )####################


Ssize<-cbind(ID2,schID,theta)
dim(Ssize)
#unique3<-Ssize[!duplicated(Ssize),]


unique3<-as.data.frame(Ssize)
unique31<-unique3[unique3$theta < 0.043,]
unique4<-unique31[order(unique31$schID),]
low<-nrow(unique4)


unique4[1:10,]
dim(unique3)


t1<-table(unique4$schID)
SSsize<-
    cbind(as.numeric(names(t1)),as.numeric(
    t1))
colnames(SSsize) <- c("schID","feq")
mque4<-merge(unique4, SSsize,
    by="schID")


schlow<-matrix(data=0,nrow=low,ncol=1)
talow25<-matrix(data=0,nrow=low,ncol=1)
talow50<-matrix(data=0,nrow=low,ncol=1)
talow100<-matrix(data=0,nrow=low,ncol=1)


for (z in 1:low){
  if( z == 1) {
     schlow[z]=1
    }
```

```
  if( unique4$schID[z] == unique4$schID[1]
    &&

     z!=1) {

     schlow[z]=1+schlow[z-1]

    }

  if( unique4$schID[z] != unique4$schID[1]
    &&

     unique4$schID[z] != unique4$schID[z-
    1]) {

     schlow[z]=1

    }

  if( unique4$schID[z] != unique4$schID[1]
    &&

      unique4$schID[z] ==
    unique4$schID[z-1]) {

     schlow[z]=1+schlow[z-1]

    }

}



mque4<-cbind(mque4,schlow)

dim(mque4)

mque4[1:10,]



t<-as.numeric(Sys.time())



for (z in 1:low){

  set.seed(mque4$schID[z]+t)

  pchoice1<-
    sample(1:mque4$feq[z],as.integer(mque
    4$feq[z]/4))

  pchoice2<-
    sample(1:mque4$feq[z],as.integer(mque
    4$feq[z]/2))

  pchoice3<-
    sample(1:mque4$feq[z],as.integer(mque
    4$feq[z]))



  if( mque4$schID[z] %in% sesch &&

     mque4$schlow[z] %in% pchoice1) {

     talow25[z]=1

  }

  if( mque4$schID[z] %in% sesch &&

     mque4$schlow[z] %in% pchoice2) {

     talow50[z]=1

  }

  if( mque4$schID[z] %in% sesch &&

     mque4$schlow[z] %in% pchoice3) {

     talow100[z]=1

  }

}



mque5<-
    cbind(mque4,talow25,talow50,talow100)
```

```
mque5[1:10,]

tlow25<-
    sort(mque5$ID2[mque5$talow25==1])

tlow50<-
    sort(mque5$ID2[mque5$talow50==1])

tlow100<-
    sort(mque5$ID2[mque5$talow100==1])

ttt4<-union(tlow25,tlow50)

ttt4<-union(ttt4,tlow100)

ttt4<-sort(ttt4)

length(ttt4)

ttt4 <- ttt4[!duplicated(ttt4)]


############### score-based tampering
    ####################


stam25<-rpo

stam50<-rpo

stam100<-rpo

praw<-matrix(data=0,nrow=N,ncol=1)


talow25<-matrix(data=0,nrow=N,ncol=1)

talow50<-matrix(data=0,nrow=N,ncol=1)

talow100<-matrix(data=0,nrow=N,ncol=1)


for (j in ttt4){
   witem=0
   FL=0


   for (n in 1:45){
   en=n*4
   praw[j]=praw[j]+rpo[j,en]
   if (rpo[j,en] == 0 ) {
     witem<-rbind(witem,n)
    }
   }


  cush<-sample(1:5,1)
  it<-30-praw[j]+cush
  wrong<-data.frame(witem)
  twro<-wrong[witem>0]
  FL<-length(twro)


  if ( any(tlow25 ==ID2[j]) &&
```

```
 FL >0 && FL <= it && FL <= 10) {
 talow25[j]=1
for (n in 0:44) {
 st=1+n*4
 num=1+n
 en=st+3
 stam25[j,st]=0
 stam25[j,st+1]=0
 stam25[j,st+2]=0
 stam25[j,en]=1
 }
}


if ( any(tlow25 ==ID2[j]) &&
  FL <=  it && FL > 10) {
talow25[j]=1
FS<-sample(twro, 10, replace = FALSE)
SFS<-sort(FS)
for (n in SFS) {
 st=1+(n-1)*4
 en=st+3
 stam25[j,st]=0
 stam25[j,st+1]=0
 stam25[j,st+2]=0
 stam25[j,en]=1
 }
}


if ( any(tlow25 ==ID2[j]) &&
    FL > it && it <=10 && it > 0) {
talow25[j]=1
FS<-sample(twro, it, replace = FALSE)
SFS<-sort(FS)
for (n in SFS) {
 st=1+(n-1)*4
 en=st+3
 stam25[j,st]=0
 stam25[j,st+1]=0
 stam25[j,st+2]=0
 stam25[j,en]=1
 }
}


if ( any(tlow25 ==ID2[j]) &&
```

```
  FL > it && it > 10 ) {
talow25[j]=1
FS<-sample(twro, 10, replace = FALSE)
SFS<-sort(FS)
for (n in SFS) {
 st=1+(n-1)*4
 en=st+3
 stam25[j,st]=0
 stam25[j,st+1]=0
 stam25[j,st+2]=0
 stam25[j,en]=1
 }
}


if ( any(tlow50 ==ID2[j]) &&
  FL >0 && FL <= it && FL <= 10) {
 talow50[j]=1
for (n in 0:44) {
 st=1+n*4
 num=1+n
 en=st+3
 stam50[j,st]=0
 stam50[j,st+1]=0
 stam50[j,st+2]=0
 stam50[j,en]=1
 }
}

if ( any(tlow50 ==ID2[j]) &&
  FL <=  it && FL > 10) {
talow50[j]=1
FS<-sample(twro, 10, replace = FALSE)
SFS<-sort(FS)
for (n in SFS) {
 st=1+(n-1)*4
 en=st+3
 stam50[j,st]=0
 stam50[j,st+1]=0
 stam50[j,st+2]=0
 stam50[j,en]=1
 }
}

if ( any(tlow50 ==ID2[j]) &&
```

```
   FL > it && it <=10 && it > 0) {
talow50[j]=1
FS<-sample(twro, it, replace = FALSE)
SFS<-sort(FS)
for (n in SFS) {
 st=1+(n-1)*4
 en=st+3
 stam50[j,st]=0
 stam50[j,st+1]=0
 stam50[j,st+2]=0
 stam50[j,en]=1
 }
}



if ( any(tlow50 ==ID2[j]) &&
  FL > it && it > 10 ) {
talow50[j]=1
FS<-sample(twro, 10, replace = FALSE)
SFS<-sort(FS)
for (n in SFS) {
 st=1+(n-1)*4
 en=st+3
 stam50[j,st]=0
 stam50[j,st+1]=0
 stam50[j,st+2]=0
 stam50[j,en]=1
 }
}

if ( any(tlow100 ==ID2[j]) &&
  FL >0 && FL <= it && FL <= 10) {
 talow100[j]=1
for (n in 0:44) {
 st=1+n*4
 num=1+n
 en=st+3
 stam100[j,st]=0
 stam100[j,st+1]=0
 stam100[j,st+2]=0
 stam100[j,en]=1
 }
}

if ( any(tlow100 ==ID2[j]) &&
  FL <=  it && FL > 10) {
```

```
talow100[j]=1
FS<-sample(twro, 10, replace = FALSE)
SFS<-sort(FS)
for (n in SFS) {
 st=1+(n-1)*4
 en=st+3
 stam100[j,st]=0
 stam100[j,st+1]=0
 stam100[j,st+2]=0
 stam100[j,en]=1
 }
}


if ( any(tlow100 ==ID2[j]) &&
   FL > it && it <=10 && it > 0 ) {
talow100[j]=1
FS<-sample(twro, it, replace = FALSE)
SFS<-sort(FS)
for (n in SFS) {
 st=1+(n-1)*4
 en=st+3
 stam100[j,st]=0
 stam100[j,st+1]=0
 stam100[j,st+2]=0
 stam100[j,en]=1
 }
}

if ( any(tlow100 ==ID2[j]) &&
   FL > it && it > 10 ) {
talow100[j]=1
FS<-sample(twro, 10, replace = FALSE)
SFS<-sort(FS)
for (n in SFS) {
 st=1+(n-1)*4
 en=st+3
 stam100[j,st]=0
 stam100[j,st+1]=0
 stam100[j,st+2]=0
 stam100[j,en]=1
 }
 }
}

#####################################
 ########################
```

```
nf25<-matrix(data=0,nrow=N,ncol=45)
nf50<-matrix(data=0,nrow=N,ncol=45)
nf100<-matrix(data=0,nrow=N,ncol=45)

nf225<-matrix(data=0,nrow=N,ncol=45)
nf250<-matrix(data=0,nrow=N,ncol=45)
nf2100<-matrix(data=0,nrow=N,ncol=45)

nst25<-matrix(data=0,nrow=N,ncol=45)
nst50<-matrix(data=0,nrow=N,ncol=45)
nst100<-matrix(data=0,nrow=N,ncol=45)


for (j in 1:N) {
  for (n in 0:44){
   st=1+n*4
   num=1+n
   st1=st+1
   st2=st+2
   en=st+3
  if(fix25[j,st] == 1) {
   nf25[j,num] = 1
                                    }
                                    if(fix25[j,st1] == 1) {
                                     nf25[j,num] = 2
                                    }
                                    if(fix25[j,st2] == 1) {
                                     nf25[j,num] = 3
                                    }
                                    if(fix25[j,en] == 1) {
                                     nf25[j,num] = 4
                                    }


                                     if(fix50[j,st] == 1) {
                                     nf50[j,num] = 1
                                    }
                                    if(fix50[j,st1] == 1) {
                                     nf50[j,num] = 2
                                    }
                                    if(fix50[j,st2] == 1) {
                                     nf50[j,num] = 3
                                    }
                                    if(fix50[j,en] == 1) {
                                     nf50[j,num] = 4
```

```
}

if(fix100[j,st] == 1) {
 nf100[j,num] = 1
}
if(fix100[j,st1] == 1) {
 nf100[j,num] = 2
}
if(fix100[j,st2] == 1) {
 nf100[j,num] = 3
}
if(fix100[j,en] == 1) {
 nf100[j,num] = 4
}

if(fix225[j,st] == 1) {
 nf225[j,num] = 1
}
if(fix225[j,st1] == 1) {
 nf225[j,num] = 2
}
if(fix225[j,st2] == 1) {
 nf225[j,num] = 3
}
if(fix225[j,en] == 1) {
 nf225[j,num] = 4
}

if(fix250[j,st] == 1) {
 nf250[j,num] = 1
}
if(fix250[j,st1] == 1) {
 nf250[j,num] = 2
}
if(fix250[j,st2] == 1) {
 nf250[j,num] = 3
}
if(fix250[j,en] == 1) {
 nf250[j,num] = 4
}

if(fix2100[j,st] == 1) {
 nf2100[j,num] = 1
}
if(fix2100[j,st1] == 1) {
 nf2100[j,num] = 2
```

```
}
if(fix2100[j,st2] == 1) {
 nf2100[j,num] = 3
}
if(fix2100[j,en] == 1) {
 nf2100[j,num] = 4
}

  if(stam25[j,st] == 1) {
 nst25[j,num] = 1
}
if(stam25[j,st1] == 1) {
 nst25[j,num] = 2
}
if(stam25[j,st2] == 1) {
 nst25[j,num] = 3
}
if(stam25[j,en] == 1) {
 nst25[j,num] = 4
}

  if(stam50[j,st] == 1) {
 nst50[j,num] = 1
}
if(stam50[j,st1] == 1) {
 nst50[j,num] = 2
}
if(stam50[j,st2] == 1) {
 nst50[j,num] = 3
}
if(stam50[j,en] == 1) {
 nst50[j,num] = 4
}

if(stam100[j,st] == 1) {
 nst100[j,num] = 1
}
if(stam100[j,st1] == 1) {
 nst100[j,num] = 2
}
if(stam100[j,st2] == 1) {
 nst100[j,num] = 3
}
if(stam100[j,en] == 1) {
 nst100[j,num] = 4
}
```

```
  }

}




allsim<-cbind(ID2,ID, disID, schID, theta,
    size, nres,

          missa,speed,nres1,nres2,

          tamper25,tamper50,tamper100,
    nf25, nf50, nf100,

          nf225, nf250, nf2100,

          talow25, talow50,talow100,
    nst25,nst50,nst100,

          tamper225,tamper250,tamper2100)

write.table(allsim,
    paste(828,c("set1sch25.dat"), sep=""),

         quote=FALSE, sep="\t", row.names
    =F,col.names =F)



# size is NULL, only 553 variables
```

APPENDIX B

AN EXAMPLE OF R SYNTAX FOR VJ AT INDIVIDUAL DETECTION

```
setwd("F:/REMS/Erasure/eworking/set1a")

F25_VS<-
    read.table("F:/REMS/Erasure/eworking/s
    et1a/sa25_F_VS.dat", quote="\"")

F25_VS[1,]

P2L3<-F25_VS[,2:46]

F25_VSco<-
    read.table("F:/REMS/Erasure/eworking/s
    et1a/sa25_F_VJ_score.SCO", quote="\"")

F25_VSco[11:17,]

ID2<-F25_VSco$V3

non_the<-F25_VSco$V1

non_se<-F25_VSco$V2

wt_VS<-matrix(data=0,nrow=35280,ncol=1)

for (j in 1:35280) {
 for (i in 1:45) {
 if(P2L3[j,i]==1) {
 wt_VS[j]=wt_VS[j]+1}
}
}

table(wt_VS)

F25_PE<-
    read.table("F:/REMS/Erasure/eworking/s
    et1a/sa25_F_PE.dat", quote="\"")

F25_PE<-F25_PE[,2:46]

F25_PE[1,]

==============================
    ==============================
    ==========

library(arm)
```

```
M1 <- bayesglm (F25_VS$V2 ~ 1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

display (M1)
lgM1 <- predict(M1)
PM1<-exp(lgM1)/(exp(lgM1)+1)

M2 <- bayesglm (F25_VS$V3 ~ 1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

display (M2)
lgM2 <- predict(M2)
PM2<-exp(lgM2)/(exp(lgM2)+1)

M3 <- bayesglm (F25_VS$V4 ~ 1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

display (M3)
lgM3 <- predict(M3)
PM3<-exp(lgM3)/(exp(lgM3)+1)

M4 <- bayesglm (F25_VS$V5 ~ 1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
```

```
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

display (M4)
lgM4 <- predict(M4)
PM4<-exp(lgM4)/(exp(lgM4)+1)

M5 <- bayesglm (F25_VS$V6 ~ 1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

lgM5 <- predict(M5)
PM5<-exp(lgM5)/(exp(lgM5)+1)

M6 <- bayesglm (F25_VS$V7 ~ 1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

lgM6 <- predict(M6)
PM6<-exp(lgM6)/(exp(lgM6)+1)

M7 <- bayesglm (F25_VS$V8 ~ 1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
```

```
        prior.df.for.intercept=1
)

lgM7 <- predict(M7)
PM7<-invlogit(lgM7)

M8 <- bayesglm (F25_VS$V9 ~ 1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

lgM8 <- predict(M8)
PM8<-invlogit(lgM8)

M9 <- bayesglm (F25_VS$V10 ~ non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

lgM9 <- predict(M9)
PM9<-exp(lgM9)/(exp(lgM9)+1)

M10 <- bayesglm (F25_VS$V11 ~ 1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

lgM10 <- predict(M10)
PM10<-exp(lgM10)/(exp(lgM10)+1)
```

```
M11 <- bayesglm (F25_VS$V12 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

lgM11 <- predict(M11)
PM11<-invlogit(lgM11)

M11 <- bayesglm (F25_VS$V12 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

lgM11 <- predict(M11)
PM11<-invlogit(lgM11)

M12 <- bayesglm (F25_VS$V13 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

lgM12 <- predict(M12)
PM12<-invlogit(lgM12)

M13 <- bayesglm (F25_VS$V14 ~
    1+non_the,
```

```
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

lgM13 <- predict(M13)
PM13<-invlogit(lgM13)

plot(non_the,PM13)

M14 <- bayesglm (F25_VS$V15 ~
    1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

lgM14 <- predict(M14)
PM14<-invlogit(lgM14)

plot(non_the,PM14)

M15 <- bayesglm (F25_VS$V16 ~
    1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

lgM15 <- predict(M15)
PM15<-invlogit(lgM15)
```

```
plot(non_the,PM15)



M16 <- bayesglm (F25_VS$V17 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)



lgM16 <- predict(M16)

PM16<-invlogit(lgM16)



plot(non_the,PM16)



M17 <- bayesglm (F25_VS$V18 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)



lgM17 <- predict(M17)

PM17<-invlogit(lgM17)


plot(non_the,PM17)


M18 <- bayesglm (F25_VS$V19 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)
```

```
lgM18 <- predict(M18)

PM18<-invlogit(lgM18)


plot(non_the,PM18)


M19 <- bayesglm (F25_VS$V20 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)


lgM19 <- predict(M19)

PM19<-invlogit(lgM19)


plot(non_the,PM19)


M20 <- bayesglm (F25_VS$V21 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)


lgM20 <- predict(M20)

PM20<-invlogit(lgM20)


plot(non_the,PM20)


M21 <- bayesglm (F25_VS$V22 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
```

```
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

lgM21 <- predict(M21)
PM21<-invlogit(lgM21)

plot(non_the,PM21)

M22 <- bayesglm (F25_VS$V23 ~
    1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

lgM22 <- predict(M22)
PM22<-invlogit(lgM22)

plot(non_the,PM22)

M23 <- bayesglm (F25_VS$V24 ~
    1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

lgM23 <- predict(M23)
PM23<-invlogit(lgM23)

plot(non_the,PM23)

M24 <- bayesglm (F25_VS$V25 ~
    1+non_the,
        family=binomial(link="logit"),
```

```
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

lgM24 <- predict(M24)
PM24<-invlogit(lgM24)

plot(non_the,PM24)

M25 <- bayesglm (F25_VS$V26 ~
    1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

lgM25 <- predict(M25)
PM25<-invlogit(lgM25)

M26 <- bayesglm (F25_VS$V27 ~
    1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

lgM26 <- predict(M26)
PM26<-invlogit(lgM26)
```

```
M27 <- bayesglm (F25_VS$V28 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)


display (M27)
lgM27 <- predict(M27)
PM27<-invlogit(lgM27)



M28 <- bayesglm (F25_VS$V29 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)


display (M28)
lgM28 <- predict(M28)
PM28<-invlogit(lgM28)



M29 <- bayesglm (F25_VS$V30 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)


display (M29)
lgM29 <- predict(M29)
PM29<-invlogit(lgM29)
```

```
M30 <- bayesglm (F25_VS$V31 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

display (M30)
lgM30 <- predict(M30)
PM30<-invlogit(lgM30)

M31 <- bayesglm (F25_VS$V32 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

display (M31)
lgM31 <- predict(M31)
PM31<-invlogit(lgM31)

M32 <- bayesglm (F25_VS$V33 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

lgM32 <- predict(M32)
PM32<-invlogit(lgM32)
```

```
M33 <- bayesglm (F25_VS$V34 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

display (M33)
lgM33 <- predict(M33)
PM33<-invlogit(lgM33)

M34 <- bayesglm (F25_VS$V35 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

display (M34)
lgM34 <- predict(M34)
PM34<-invlogit(lgM34)

M35 <- bayesglm (F25_VS$V36 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

display (M35)
lgM35 <- predict(M35)
PM35<-invlogit(lgM35)

M36 <- bayesglm (F25_VS$V37 ~
    1+non_the,
```

```
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

display (M36)
lgM36 <- predict(M36)
PM36<-invlogit(lgM36)

M37 <- bayesglm (F25_VS$V38 ~
    1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

display (M37)
lgM37 <- predict(M37)
PM37<-invlogit(lgM37)

M38 <- bayesglm (F25_VS$V39 ~
    1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

display (M38)
lgM38 <- predict(M38)
PM38<-invlogit(lgM38)

M39 <- bayesglm (F25_VS$V20 ~
    1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
```

```
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

display (M39)
lgM39 <- predict(M39)
PM39<-invlogit(lgM39)

M40 <- bayesglm (F25_VS$V41 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

display (M40)
lgM40 <- predict(M40)
PM40<-invlogit(lgM40)

M41 <- bayesglm (F25_VS$V42 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
            prior.scale.for.intercept=10,
            prior.df.for.intercept=1
)

display (M41)
lgM41 <- predict(M41)
PM41<-invlogit(lgM41)

M42 <- bayesglm (F25_VS$V43 ~
    1+non_the,
            family=binomial(link="logit"),
            prior.mean=0,
            prior.scale=2.5,
            prior.df=1,
            prior.mean.for.intercept=0,
```

```
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

display (M42)
lgM42 <- predict(M42)
PM42<-invlogit(lgM42)

M43 <- bayesglm (F25_VS$V44 ~
  1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

display (M43)
lgM43 <- predict(M43)
PM43<-invlogit(lgM43)

M44 <- bayesglm (F25_VS$V45 ~
  1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
        prior.df=1,
        prior.mean.for.intercept=0,
        prior.scale.for.intercept=10,
        prior.df.for.intercept=1
)

display (M44)
lgM44 <- predict(M44)
PM44<-invlogit(lgM44)

plot(non_the,PM44)

M45 <- bayesglm (F25_VS$V46 ~
  1+non_the,
        family=binomial(link="logit"),
        prior.mean=0,
        prior.scale=2.5,
```

```
        prior.df=1,

        prior.mean.for.intercept=0,

        prior.scale.for.intercept=10,

        prior.df.for.intercept=1

)



display (M45)

lgM45 <- predict(M45)

PM45<-invlogit(lgM45)



plot(non_the,PM45)



vmF25<-
    cbind(PM1,PM2,PM3,PM4,PM5,PM6,P
    M7,PM8,PM9,PM10,PM11,PM12,PM1
    3,PM14,PM15,


    PM16,PM17,PM18,PM19,PM20,PM21,
    PM22,PM23,


    PM24,PM25,PM26,PM27,PM28,PM29,
    PM30,PM31,PM32,PM33,PM34,PM35,
    PM36,


    PM37,PM38,PM39,PM40,PM41,PM42,
    PM43,PM44,PM45)



dim(vmF25)

write.table(vmF25, paste(c("vmF25.dat"),
    sep=""),

        quote=FALSE, sep="\t", row.names
    =F,col.names =F)



==============================
    ==============================
    =====

#vm<-
    read.table("C:/Users/Sherry/Documents/
    828vmF25.dat", quote="\"")

#vm[201,]



PEw_VS<-
    matrix(data=0,nrow=35280,ncol=1)

empty_vm<-
    matrix(data=NA,nrow=35280,ncol=45)



for (j in 1:35280) {

 for (i in 1:45) {

 if(F25_PE[j,i] != 4) {

 PEw_VS[j]=PEw_VS[j]+1

 empty_vm[j,i]=vmF25[j,i]}

}

}
```

```
table(PEw_VS)

Prv<-matrix(data=0,nrow=35280,ncol=1)

N=35280

for (j in 1: N){

 if(PEw_VS[j] == wt_VS[j]) {

 nm<-na.omit(empty_vm[j,])

 prod(nm)

 }

 if( PEw_VS[j] != wt_VS[j] &  wt_VS[j] >
    0 ) {

 M=10000

 n=PEw_VS[j]

 x=numeric(M)

 brk=0:n

 nm<-na.omit(empty_vm[j,])

 for (i in 1:M) {x[i]=sum(rbinom(n,1,nm))}

 dis<-hist(x, freq=F, breaks=brk,
    ylim=c(0,1),xlab="",ylab="")

 w<-wt_VS[j]

 Prv[j]<-sum(dis$density[w : n])

 }

 }

write.table(Prv, paste(c("PrvF25.dat"),
    sep=""),

      quote=FALSE, sep="\t", row.names
    =F,col.names =F)
```