Using Nonlinear Regression to Identify Unusual Performance Level Classification Rates

Paper Presented at the Statistical Detection of Potential Test Fraud Conference
Madison, Wisconsin

J. Michael Clark III, Pearson
William P. Skorupski, University of Kansas
Stephen T. Murphy, Pearson

October 2013

Author Note

J. Michael Clark III, Psychometric Services, Pearson; William P. Skorupski, Department of Psychology and Research in Education, University of Kansas; Stephen T. Murphy, Psychometric Services, Pearson.

Special thanks to Robert Furter for his valuable input on aspects of the data simulation methodology.

Correspondence concerning this manuscript should be addressed to Mike Clark, Pearson, 2488 E. 81st St., Suite 4700, Tulsa, OK 74137. Email: mike.clark@pearson.com.

Abstract

In this paper, the authors propose a new technique for identifying unusual changes in test performance at the classroom level. Using a cumulative logit regression model, test-takers' performance level categories in the current year are predicted by their test scores in the same subject area from the prior year. Individual probabilities obtained from the cumulative logit model are aggregated across examinees within the classroom to derive expected proportions of students at each performance level. Expected and observed proportions are compared, and a standardized residual is computed for the purposes of flagging classrooms with much greater-than-expected test performance gains. Results of a simulation study found that this method has good detection power when classrooms show moderate-to-large score gains due to cheating and misconduct is not overly widespread in the test-taking population. Among the total number of classrooms flagged by this method, over 98.9% were true cheating classrooms in all included conditions of the simulation. Type I error rates for this method were conservative in all investigated conditions.

Keywords: cheating, test security, misconduct, performance gains

Using Nonlinear Regression to Identify Unusual Performance Level Classification Rates

Large-scale educational testing is becoming an increasingly high-stakes endeavor. Critical decisions affecting students' lives, such as passing courses, progressing to the next grade level, and graduating from high school may be directly impacted by performance on state-mandated standardized tests. However, high stakes associated with testing outcomes are not limited to test-takers exclusively; educators also may experience pressures brought on by accountability expectations related to their students' performance on state assessments. Given the ubiquity of large-scale assessments in the current educational landscape and the accountability pressures placed upon stakeholders, any number of individuals playing various roles in the testing process—such as test-takers, teachers, proctors, educational administrators, or others—may potentially feel pressures to engage in test misconduct. Recognizing that obtaining valid test scores is a central concern for any testing program and that test misconduct—whether it is carried out with intent or not—represents a serious threat to test score validity (Cizek, 1999; Amrein-Beardsley, Berliner, & Rideau, 2010), it is in the best interest of assessment program owners, administrators, and vendors to remain vigilant for evidence of possible misconduct.

With accountability measures directly tied to student performance on state assessments, educators wishing to gain rewards or perhaps avoid sanctions might be incentivized to engage in misconduct in an effort to boost students' scores. Numerous high-profile allegations of cheating carried out by educators have been brought to light in recent years (e.g., Vogell, 2011; Molland, 2012; National Center for Fair and Open Testing, 2013). Educators who engage in misconduct may manipulate test scores in a number of possible ways, including but not limited to obtaining test items prior to the administration and sharing answers with students; allowing assistive materials such as notes, posters, or textbooks to remain accessible to students during the test;

coaching or assisting students during the test administration; or erasing incorrect submitted responses and replacing them with correct answers (Cizek, 2001). When misconduct occurs, it may leave telltale signatures in test data, and its manifestations will likely vary depending on the form of misconduct as well as the role-players responsible. If one makes a reasonable assumption that whichever of the aforementioned forms of misconduct results in an increase in test-takers' observed performance on the assessment beyond what might otherwise be expected given their true levels of achievement, one straightforward approach to identify potential incidents of misconduct may be to model student achievement longitudinally, looking for much greater-than-predicted increases in test-taker performance in the focal year compared to prior performance, provided of course that the previous year's instructor did not also engage in misconduct.

Techniques Used to Identify Unusual Performance Gains

An observation of unusual student performance gains over time may indicate potential test misconduct (e.g., Jacob & Levitt, 2002; NCME, 2012). For example, a classroom (or more generally, unit) that includes a large number of historically low-achieving students who are now apparently demonstrating high achievement in the current time point—well beyond what might be considered a "reasonable" performance gain when consulting the results of a statistical prediction model—may elicit suspicion. Finding the most appropriate and widely-applicable statistical modeling strategy, considering the characteristics of the data and the appropriateness of outcomes for making inferences, was the primary goal of this research study.

An argument could be made that one reasonable method for identifying unusual gains in student performance over time may involve some form of a linear regression model. In a widely-publicized investigation, the Atlanta Journal-Constitution (2012) obtained mean scale scores for

schools, grades, and subject areas, and used a linear regression model, weighted by classroom size, to predict average classroom scale scores in the current year from average scale scores in the previous year. Investigators compared standardized residuals to the *t*-distribution, adjusting the flagging criterion for classroom size, and flagged classrooms with standardized residuals that had probabilities less than 0.05.

Jacob and Levitt (2003) used a very different approach to find evidence of potential misconduct in classrooms. Rather than modeling changes in mean scores over time using a regression model, they computed percentile ranks for students within classrooms across several time points and looked for evidence of potential misconduct by identifying classrooms made up of students who previously had low scores relative to the population, in terms of percentile ranks, and now showed extremely large percentile rank improvements in the current time point. Such a technique shifts focus away from the scores themselves and instead emphasizes conditional percentile ranks across years, not unlike some of the current popular growth modeling techniques (e.g., see Student Growth Percentiles; Betebenner, 2009), albeit in a less sophisticated manner.

Although these techniques are straightforward and intuitive, their limitations warrant consideration. For the purposes of identifying suspicious test scores, such techniques may not be ideal in all circumstances. Identifying large standardized residuals computed from a linear model will identify observed mean test scores with the greatest deviations from their predicted values. Similarly, any method based on percentile ranks will be most sensitive to large relative score gains across years. Although an observation of an extreme unit-level score gain across years may (rightly) elicit suspicion of the validity of the current year's test scores for a particular unit, this may not be the only possible manifestation of misconduct. Because accountability measures are tied to performance level outcomes—not scores themselves—there may not be great incentive

for those inclined to engage in misconduct to intervene in such a way that results in historically low-performing students in the unit now scoring into the upper echelons of the state's test-takers. Furthermore, larger gains like these will likely require greater levels of intervention from the educator (thus increasing the risk of being caught while performing said intervention) and invite further scrutiny and suspicion when these previously low-achieving students' scores show such improbably massive gains.

Another possible manifestation of misconduct that should warrant suspicion may be smaller, but systematic, score gains at the student level, which are sufficient to significantly improve the number of students in a unit who reach a performance level category corresponding to a level of proficient or above. If only small or moderate gains obtained through cheating are necessary to boost students' scores in a unit into a desirable performance level category, methods that are most sensitive to either large differences in unit score means or large differences in percentile ranks over time may be less capable of identifying these cases of misconduct.

For the purposes of this research study, the authors sought to investigate a modeling strategy that satisfied a number of criteria. First, the authors sought a method that is widely applicable to a variety of test delivery circumstances. For example, not every state uses a vertical scale, so methods that are scale-dependent, such as those requiring the computation of difference scores across years, were not considered. Next, the chosen technique should focus on performance level categories as opposed to scores as the outcome of interest. Accountability measures are tied to performance level outcomes; therefore, any techniques designed to flag suspicious outcomes ought to similarly focus on performance levels. Furthermore, a method predicting performance level outcomes should have equal or superior power to more traditional linear models in a wide variety of situations. For example, in circumstances in which students'

test scores in a particular unit show very large gains across years due to cheating, both linear and nonlinear methods should have reasonable detection power. The linear model will flag the unit because the large change in scores will result in a significant standardized residual, and the nonlinear model will flag the unit because more students scored into the higher performance level categories than what was predicted by the model based on their previous observed scores. However, in circumstances in which a unit contains a large percentage of students who would be expected to score somewhere reasonably close to, but below the proficient cut, but those students' scores were boosted above the cut (but again, not necessarily to the top of the scale) through cheating, nonlinear methods that compare predicted and observed performance level classification rates at the unit level should have superior power to linear models, which would be less sensitive to smaller, systematic gains such as this. Finally, because we recognize that a certain degree of classification error is inherently unavoidable when predicting categorical outcomes like performance level from continuous predictors like scale scores, any chosen modeling strategy should account for this uncertainty in predicted outcomes when computing expected values. A good modeling strategy should not produce a single expected value in the sense that a time $t - 1$ score of X results in a predicted time t performance level category of Y (alone), but rather, the modeling technique should provide *probabilities* for all possible performance level categories. Although some combinations of X and Y may be highly improbable, no combination is impossible, and there may be circumstances in which two or more categories are almost equally probable; expected values should reflect this reality.

A Cumulative Logit Regression Modeling Approach

The authors propose a strategy that uses aggregated results of an examinee-level nonlinear regression model to make unit-level inferences regarding the reasonableness of

observed proportions of test-takers classified into performance level categories (denoted $j = 1, 2, \ldots J$), given students' performance in the previous grade level of the subject area of interest. The proposed methodology models test performance longitudinally across time points and within test-takers. The outcome variable in the prediction model is the student's observed performance level category (Y) at time t, which is treated as an ordinal variable, and the predictor variable is the student's scale score in the previous grade level at time $t - 1$; therefore, this is characterized as a cumulative logit regression model (Agresti, 1996). Because students are not necessarily expected to be placed in identical classrooms—in terms of classrooms being comprised of identical peer groups—across years, we make no assumption that students are nested within units across the two time points.

As described by Agresti (1996), cumulative probabilities reflect the ordering that $P(Y \leq 1) \leq P(Y \leq 2) \leq \ldots P(Y \leq J) = 1$; the cumulative probability that $Y \leq J$ necessarily equals 1, and the logits for the first $J - 1$ cumulative probabilities are

$$\begin{aligned} \text{logit}[P(Y \leq j)] &= \log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) \\ &= \log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_j}\right), \; j = 1, \ldots, J - 1. \end{aligned} \tag{1}$$

Cumulative probabilities can be used to compute individual probabilities. For performance level category $j = 1$, the individual probability is equal to the cumulative probability for $j = 1$. For performance levels $j = 2$ through J, the individual probability is equal to the difference between the cumulative probability for category j and the cumulative probability for category $j - 1$. An example showing individual probabilities for a test with $J = 4$ performance level categories at time t is provided in Figure 1.

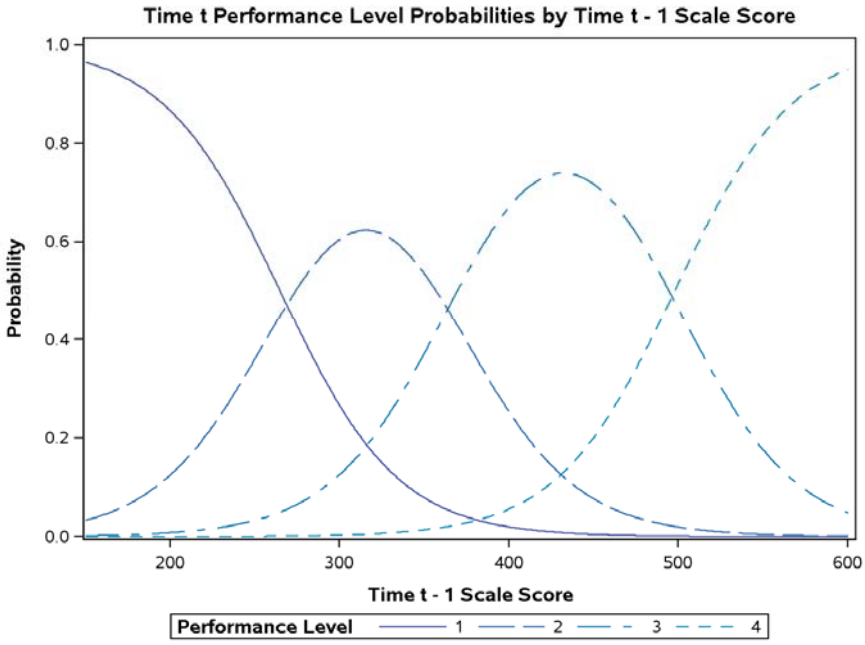

Figure 1. Example of Individual Probabilities for Four Performance Level Categories.

As is evident from this example, we can compute independent probabilities of scoring into each of the four performance levels at time *t* conditional on any observed scale score from time *t* – 1, and independent probabilities sum to 1. For example, a test-taker who received a score of 300 at time *t* – 1 has a probability of 0.266 of scoring into Performance Level 1 at time *t*, a 0.605 probability of scoring into Performance Level 2, a 0.126 probability of scoring into Performance Level 3, and a 0.003 probability of scoring into Performance Level 4. For this student, we might expect that he or she would score into Performance Level 2 at time *t*, although a score placing him/her into Performance Level 1 would not be out of the question, either. A score placing this student into Performance Level 4 at time *t*, however, would be highly improbable.

This example illustrates a useful benefit of this approach over other potential methods that might be considered. Rather than obtaining a single expected performance level for a conditional scale score, we obtain expected probabilities for all four performance levels for any

conditional scale score. A possible performance level-oriented alternative approach might be to predict examinees' scale scores at time *t* from their time *t* – 1 scale scores, and then apply the performance level cuts to the predicted and observed time *t* scale scores to obtain predicted and observed performance level outcomes. However, as illustrated in Figure 1, there are regions of scale scores where multiple performance level outcomes are nearly equally probable. For example, a test-taker who obtained a scale score of 364 at time *t* – 1 has a 0.461 probability of scoring into Performance Level 2 at time *t*, and a 0.465 probability of scoring into Performance Level 3. Although this test-taker has a slightly higher probability of scoring into Performance Level 3, practically speaking, the difference is trivial and we might conclude that either outcome seems fairly reasonable. When aggregating expected values across students within a unit, unit-level expected values computed through this cumulative logit model will reflect this inherent uncertainty in student-level performance level classifications, particularly when multiple performance level outcomes are likely.

Treating independent probabilities as examinee-level expected values, we can easily aggregate these and compute the expected count of examinees at performance level *j* for unit (classroom) *k* by summing independent probabilities for performance level *j* across all examinees in the unit ($i = 1, 2, \ldots, n_k$), conditioning on their respective scale scores from the prior time point. Dividing these expected counts by the total number of examinees in the unit yields the expected proportion of examinees at performance level *j*:

$$E\left(P_{jk}\right) = \frac{\sum_{i=1}^{n_k} P_{ij}}{n_k}. \tag{2}$$

The standardized residual for performance level *j* for unit *k* is the difference between the observed proportion, P_{jk}, and the expected proportion, $E(P_{jk})$, divided by the standard error:

$$SR_{jk} = \frac{P_{jk} - E(P_{jk})}{SE_{jk}}, \tag{3}$$

where the standard error is equal to the square root of the variance of the mean:

$$SE_{jk} = \sqrt{\frac{E(P_{jk})\left(1 - E(P_{jk})\right)}{n_{jk}}}. \tag{4}$$

Standardized residuals are expected to be normally distributed with a mean of 0 and a standard deviation of 1. For any performance level, positive residuals are taken as indication of a higher-than-predicted proportion of students within the unit scoring into that performance level, and a negative residual is indication of a lower-than-predicted proportion. If, for example, a given assessment program has four performance level categories, and Performance Level 3 corresponds to "proficient" or on-grade-level performance and Performance Level 4 corresponds to superior performance, then extremely large, positive residuals for either or both of these categories may be taken as a potential reason for further follow-up.

Method

Data Simulation

This proposed method was investigated using a simulation study. The simulation we conducted included several steps, including performing a review of the characteristics of data obtained from various large-scale assessments, selecting data to form the foundation for the simulation, simulating "clean" data with no cheating interventions included, and performing additional manipulations to the simulated clean data to mimic cheating. These steps are described in further detail in the following sections.

Base data

Prior to simulating data for this study, we examined a number of test data files from a variety of large-scale testing programs for the purposes of establishing reasonable parameters for the data simulation. After completing this broad examination of test data, we selected data from two adjacent years and grade levels of a reading assessment to serve as the base data for the simulation. These data were chosen to serve as the basis of the simulation after being found to be representative of the typical univariate and multivariate characteristics of longitudinal test data that were observed in the broader investigation. After matching students based on a unique identification variable across years 1 and 2 and performing typical test data clean-up steps, such as dropping invalid scores and duplicate cases found within a single administration, students' reading scores were found to have a correlation of $r = 0.759$ across years. Univariate descriptive statistics for base data are provided in Table 1, and the joint distribution is shown in the density plot provided in Figure 2.

Table 1

Descriptive Statistics for Base Data Scale Scores

Year	*N*	Mean	*SD*	Skewness	Kurtosis
1	69,709	488.032	64.710	-0.225	-0.203
2	69,709	494.139	66.423	-0.334	-0.227

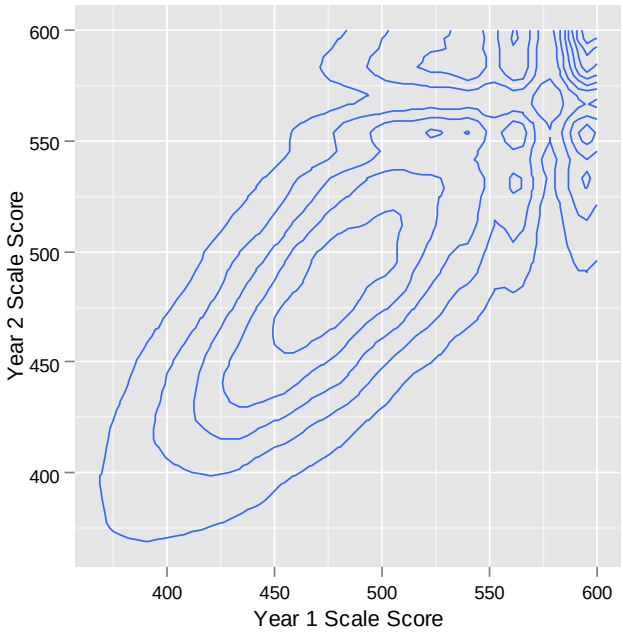

Figure 2. Joint Density of Base Data.

Simulation procedure

The data simulations were completed in two sequential steps. In the first step, data were simulated to match the observed distributional characteristics of the base data set. A total of 2,000 replicated data sets were simulated. Each replicated data set contained raw scores and scale scores computed using the actual concordance tables from two time points (denoted Year 1 and Year 2) for 70,000 test-takers. For each replication, simulated test-takers were randomly assigned to one of 2,800 units (classrooms). Although the total number of units was fixed at 2,800 for each replication, unit sizes were allowed to randomly vary across replications. The mean unit size was approximately 25 for each replication, unit sizes generally ranged between 10 and 45, and the 25th and 75th quantiles for unit size were generally near 22 and 28, respectively. Because no misconduct was simulated at this stage, these simulated data sets will be referred to as the "clean" simulation condition in the study.

In the next phase of the simulation, data from the clean condition were manipulated to simulate the effect of cheating on units. For each replicated data set, a random draw was taken from a uniform distribution ranging from 0.01 to 0.30. This value was set as the replication-specific theoretical population cheating rate. For each of the 2,800 units within that replicated data set, another random draw was pulled from a uniform distribution ranging from 0.00 to 1.00. If the unit's random draw was less than the theoretical population cheating rate, the unit was selected as a cheating unit, otherwise, it was not. Due to the probabilistic nature of the method used to assign cheating units, observed percentages of cheating units within replications occasionally varied outside of the theoretical boundaries of 1 – 30%.

For units randomly selected to be impacted by cheating, scores for all test-takers within that unit were increased by some fixed amount. There were three different cheating conditions: in the first condition, Year 2 scores were increased between 0.5 and 1.0 standard deviations above the originally-simulated "clean" Year 2 scores; in the second condition, scores in cheating units were increased between 1.0 and 1.5 standard deviations; and in the third condition, scores in cheating units were increased between 1.5 and 2.0 standard deviations. In terms of raw scores, the standard deviation for the selected Year 2 base data was approximately 6, so for the 0.5 to 1.0 standard deviation boost condition, test scores in cheating units were increased between 3 and 6 raw scores; for the 1.0 to 1.5 *SD* boost condition, tests within cheating units saw increases between 6 and 9; and for the 1.5 to 2.0 boost *SD* condition, tests within cheating units saw increases between 9 and 12. Boosted raw scores were converted to scale scores using the base test's raw score to scale score conversion table. Score increases were not allowed to exceed the highest obtainable scale score, 600. Data simulations for both the clean and cheating conditions were performed using R, version 3.0.1.

Analyses

Cumulative logit regression model

For each replicated data set, performance levels in Year 2 were predicted by Year 1 scale scores in a cumulative logit regression model, estimated using PROC LOGISTIC in SAS, version 9.2. Examinee-level individual probabilities for the four performance levels were exported from the model results and used to estimate unit-level expected proportions at each performance level. Standardized residuals for each unit and performance level were computed based on observed and expected proportions.

Because scoring into either Performance Level 3 or 4 constitutes passing on the base test, flagging outcomes were based on the values of unit-level standardized residuals for those performance levels. The theoretical flagging criterion chosen for this study was 3.00. Because units could be flagged based on the value of the standardized residual for either Performance Levels 3 or 4, the flagging criterion was adjusted to maintain control over the Type I error rate. To account for two comparisons being performed for each unit, the adjusted flagging criterion was set to be approximately equal to 3.205, which corresponds to the critical value associated with $\alpha / 2$. Cumulative logit models were fit to the clean (no cheating) data condition, as well as the three conditions in which examinees in simulated cheating units had their Year 2 scores boosted by between 0.5 to 1.0 standard deviations, 1.0 to 1.5 standard deviations, or 1.5 to 2.0 standard deviations.

Weighted least squares regression model

To contrast this proposed cumulative logit regression model with a linear prediction model, we computed unit mean scores for Year 1 and Year 2. Simulated examinees were grouped in both years based on their Year 2 unit IDs. Using SAS PROC REG, we predicted Year

2 unit means from Year 1 unit means, weighting by unit size. Standardized residuals were exported from the model results and compared to various flagging criteria, which will be discussed further in the Results section.

Evaluating the Efficacy of the Investigated Methods

Simulation research, whether it is focused on test misconduct or identification rates for any other phenomenon of interest, commonly reports four outcomes of interest. Reported in the context of a study such as this, those outcomes would be sensitivity, or the ratio of correctly-identified cheating units to the total number of true cheating units; specificity, or the ratio of correct non-identifications to the total number of non-cheating units; type I error rates, or the ratio of false identifications to the total number of non-cheating units; and finally, type II error rates, or the ratio of falsely non-identified cheating units to the total number of cheating units. However, given the circumstances unique to test misconduct, it is critical that we examine the performance of these indicators more fully. If we make the very reasonable assumptions that 1) classification errors are unavoidable, 2) false accusations of misconduct are potentially damaging to the reputations of both the accused and the accuser, and 3) state education agencies (SEAs) and local education agencies (LEAs) have finite budgets and therefore likely cannot investigate every case of potential misconduct, then it becomes very important to adopt a more Bayesian-like approach in evaluating the performance of these statistics (Skorupski & Wainer, 2013).

As previously mentioned, sensitivity represents the ratio of correctly-identified cheating units to the total number of cheating units, which is very useful for evaluating how often cheating units might be expected to be identified by some means. However, one could make a reasonable argument that perhaps an even more important consideration, particularly from a policy standpoint, may be how often those units that are identified by the statistic are truly

cheating units—information that is not directly provided by sensitivity. We suggest that the positive predictive value—or, the ratio of correctly-identified cheating units to the total number identified as cheating units by the method—holds great value in examining the usefulness of any statistic used in misconduct research. This value will be reported and discussed. Further, the negative predictive value—or ratio of correctly non-identified units to the total number of non-identified units—and efficiency—or ratio of correct outcomes (i.e., true positives and true negatives) to total outcomes—will be reported and discussed as well.

The proposed cumulative logit regression method was expected to have good detection power, in terms of correctly identifying cheating units, and reasonable type I error rates relative to the flagging criterion. Power was expected to be adequate even when rather small cheating effect sizes were simulated, power should increase with larger simulated cheating effect sizes, and the proposed nonlinear approach was expected to demonstrate equal or superior power, positive predictive value, and efficiency to the linear model in all investigated conditions.

Results

Cumulative Logit Regression Method

Marginal results for flagging outcomes from the cumulative logit model are shown in Table 2. As expected, detection power was greatest when simulated examinees within cheating units experienced score gains greater than or equal to 1.5 standard deviations, with approximately 63.7% of cheating units being correctly identified in this condition. Detection power greatly decreased in conditions where misconduct had less impact on scores within cheating units, with only 8.5% of these units being detected in the condition where misconduct resulted in score boosts less than or equal to 1 standard deviation. Positive predictive power remained strong, however, with true cheating units making up more than 98.9% of the total

number of flagged units in all three conditions. Efficiency ranged from 85.8 – 94.3%, with the greatest efficiency observed in the condition with the largest cheating effect size. In terms of type I errors, a nominal false positive rate of approximately 0.135% would be expected with a theoretical flagging criterion set to 3.00. As shown in this table, empirical type I error rates were conservative in all three cheating conditions. Although not presented in this table, results from the clean data simulation condition, in which no cheating was simulated, showed a marginal type I error rate of 0.003%.

Table 2

Marginal Results for Cumulative Logit Method with Flagging set to 3.00

ES[a]	Sensitivity	Specificity	Type I	Type II	PPV[b]	NPV[c]	Efficiency
0.5 – 1.0	8.450	99.984	0.016	91.550	98.976	85.670	85.845
1.0 – 1.5	34.869	99.965	0.035	65.131	99.450	89.454	89.991
1.5 – 2.0	63.733	99.939	0.061	36.267	99.487	93.683	94.266

Note. All values are presented as percentages.
[a]Effect size impact of cheating on affected scores, in standard deviations.
[b]Positive predictive value.
[c]Negative predictive value.

Looking beyond marginal outcomes from the three levels of cheating effect size, we examined the impact of the extent of cheating within the data set on flagging outcomes. As previously discussed, we simulated theoretical population cheating percentages ranging from 1 – 30%, and the extent of cheating in the population was allowed to vary randomly between those theoretical boundaries across the 2,000 simulated data sets within each of the three cheating effect size conditions. The following three figures illustrate the relationship between detection power and the extent of cheating simulated within the data set. In these scatter plots each point represents the outcome from a single replicated data set within that condition. As shown in these figures, detection power decreased in all three conditions as simulated misconduct became more widely prevalent within the data set. Type I error rates and positive predictive value, which are

not presented in these figures, were not systematically impacted by the extent of cheating in the data set.

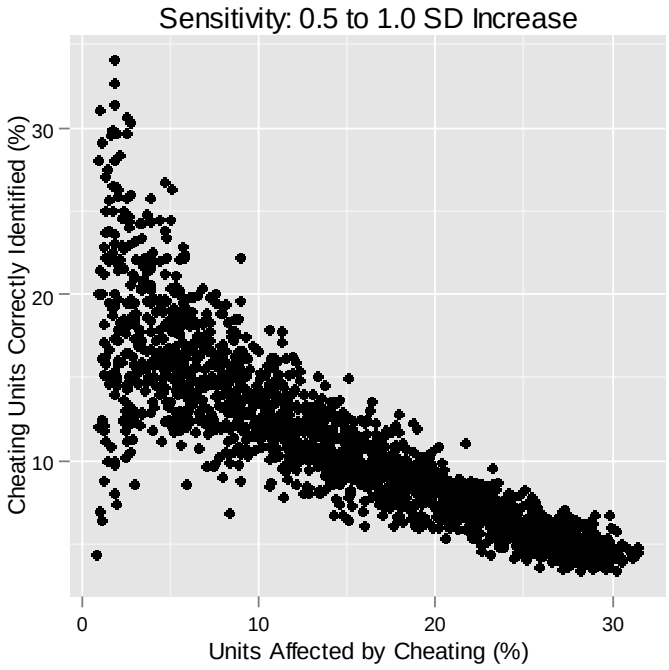

Figure 3. Sensitivity by Extent of Cheating: 0.5 to 1.0 *SD* Increase Condition.

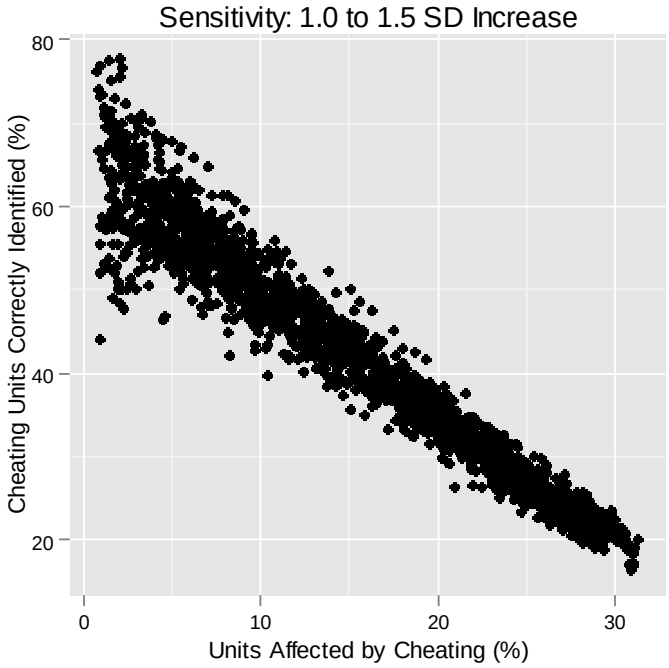

Figure 4. Sensitivity by Extent of Cheating: 1.0 to 1.5 *SD* Increase Condition.

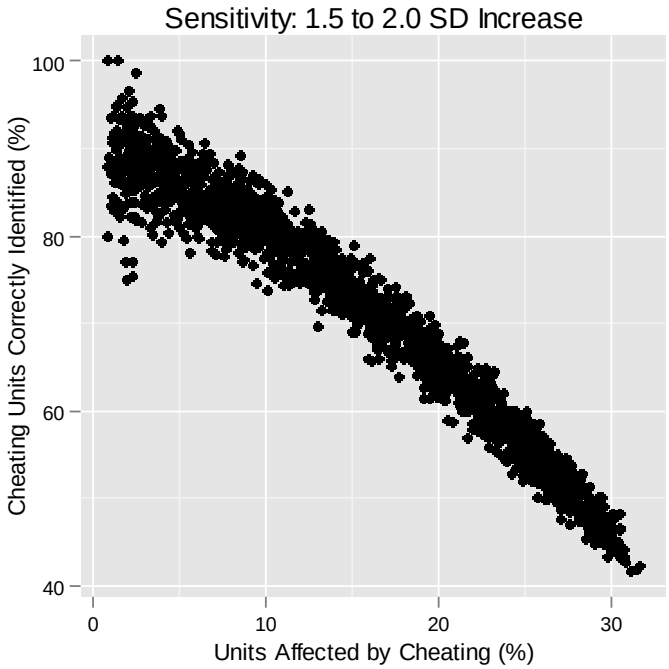

Figure 5. Sensitivity by Extent of Cheating: 1.5 to 2.0 *SD* Increase Condition.

We also investigated the impact of unit size on detection power, type I error rates, and positive predictive value. Unit sizes varied within each data set. For reporting and discussion purposes, we grouped units into four categories based on size, splitting them at approximately the 25th, 50th, and 75th percentiles. The unit sizes grouped into these four categories are 1) less than 22 students, 2) 22 – 24 students, 3) 25 – 28 students, and 4) 29 or more students. As shown in the following figures, detection power was greatest for larger units. Not shown in these figures, type I error rates and positive predictive value remained consistent across the unit size groups.

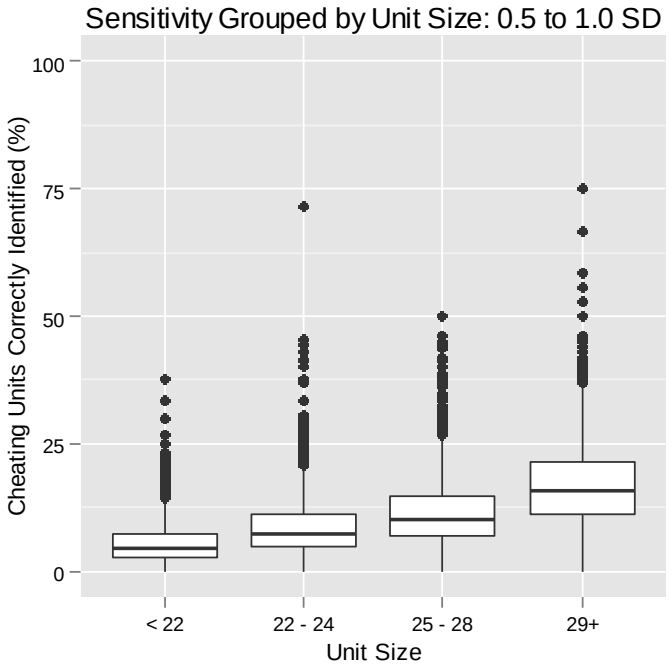

Figure 6. Sensitivity by Unit Size: 0.5 to 1.0 *SD* Increase Condition.

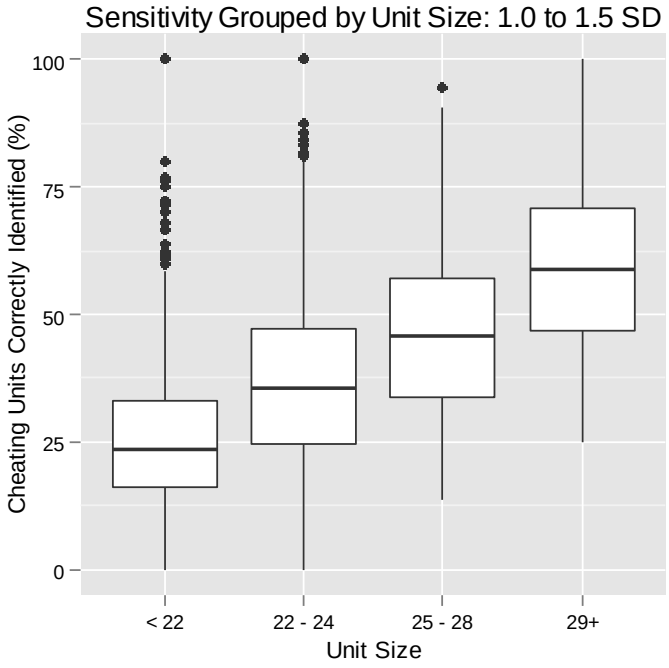

Figure 7. Sensitivity by Unit Size: 1.0 to 1.5 *SD* Increase Condition.

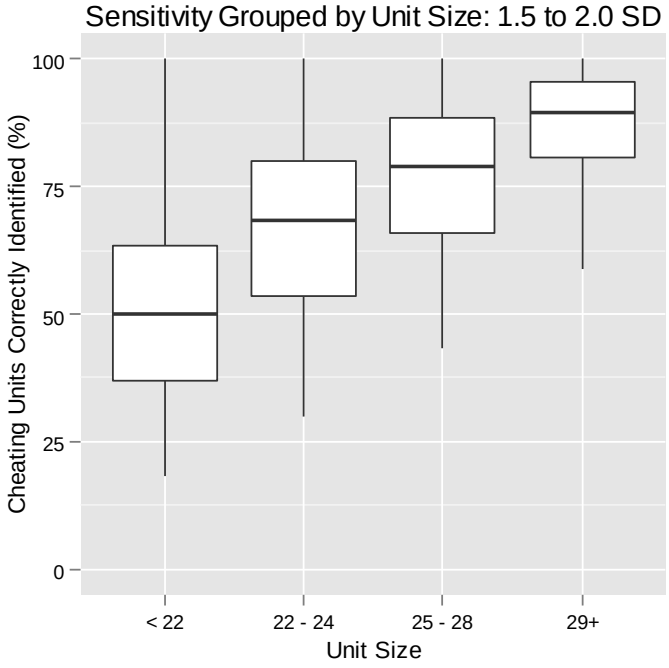

Figure 8. Sensitivity by Unit Size: 1.5 to 2.0 *SD* Increase Condition.

Weighted Least Squares Regression Method

Marginal results for the three cheating conditions from the weighted least squares regression model are shown in Table 3. As was true for the cumulative logit regression method, flagging was initially set at 3.00 for the standardized residuals. Sensitivity was poor in all three conditions, with hit rates ranging from only 9.0 – 9.9%. Type I error rates were extremely conservative in all three conditions, and the positive predictive value exceeded 99.9% in all three conditions. Efficiency held steady at approximately 86% in all three conditions. The type I error rate from the clean condition, which is not presented in this table, was very close to the nominal type I error rate at 0.142%.

Table 3

Marginal Results for WLS Method with Flagging set to 3.00

ES	Sensitivity	Specificity	Type I	Type II	PPV	NPV	Efficiency
0.5 – 1.0	9.043	99.999	0.001	90.957	99.943	85.751	85.950
1.0 – 1.5	9.876	100.000	0.000	90.124	99.993	85.978	86.191
1.5 – 2.0	9.468	100.000	0.000	90.532	100.000	85.601	85.815

Because sensitivity was so poor for the WLS method in all three cheating conditions, we followed-up by exploring the performance of this method using additional flagging criteria. Performance measures were recomputed based on two additional flagging criteria for the standardized residuals. Results from these additional investigations are shown in Table 4 and Table 5.

Table 4

Marginal Results for WLS Method with Flagging set to 2.00

ES	Sensitivity	Specificity	Type I	Type II	PPV	NPV	Efficiency
0.5 – 1.0	39.411	99.948	0.052	60.589	99.281	90.030	90.597
1.0 – 1.5	42.488	99.987	0.013	57.512	99.837	90.573	91.177
1.5 – 2.0	41.512	99.995	0.005	58.488	99.934	90.198	90.832

Table 5

Marginal Results for WLS Method with Flagging set to 1.00

ES	Sensitivity	Specificity	Type I	Type II	PPV	NPV	Efficiency
0.5 – 1.0	86.441	98.501	1.499	13.559	91.328	97.547	96.638
1.0 – 1.5	96.784	99.358	0.642	3.216	96.462	99.418	98.963
1.5 – 2.0	98.925	99.599	0.401	1.075	97.866	99.800	99.494

As shown by these tables, sensitivity improves when lower flagging thresholds are used—an expected outcome. Type I error rates remain extremely conservative compared to their respective nominal levels, and as a result, positive predictive value and efficiency were strong at these lower flagging thresholds. However, the outcomes from the clean simulation condition once again reveal type I error rates at these new flagging criteria that are extremely close to the nominal levels for these new flagging criteria. When the flagging criterion was set to 2.00 and no cheating was present, approximately 2.3% of units were incorrectly flagged, and when the flagging criterion was set to 1.00 and no cheating was present, approximately 15.9% of cheating units were incorrectly flagged. These outcomes were both consistent with the nominal Type I error rates that would be expected when no cheating is present.

The extent of cheating in the data set was found to have a rather profound impact on the type I error rates of the weighted least squares method. When cheating is present, type I error rates are extremely conservative, and this effect is multiplied as the impact of cheating on test scores increases, but when cheating is scarce or altogether absent, type I error rates are at their nominal levels. To further illustrate this phenomenon, we plotted empirical type I error rates observed across the 2,000 replications of the cheating condition in which scores in cheating units were boosted between 0.5 and 1.0 standard deviation. As shown in Figure 9, empirical type I error rates for the WLS method were impacted by the percentage of cheating units in the data set.

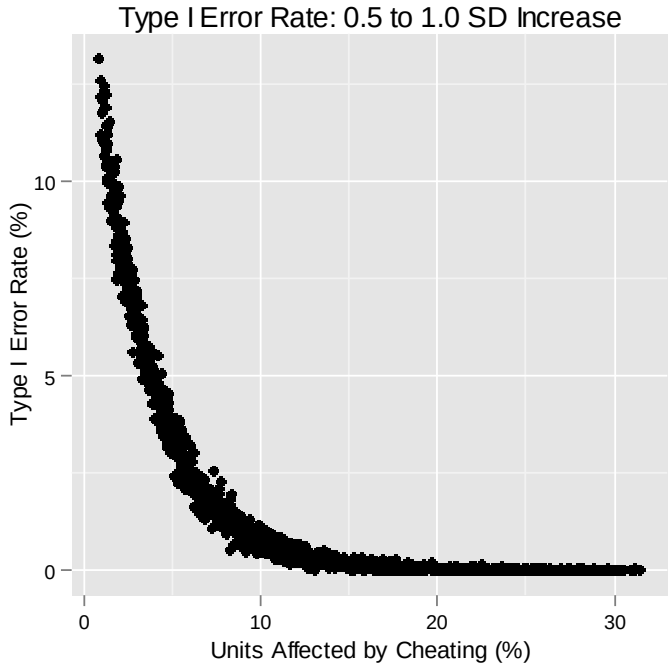

Figure 9. Empirical Type I Error Rates by Percent of Cheating Units for the Weighted Least Squares Regression Method: 0.5 to 1.0 Standard Deviation Increase Condition; Flagging at 1 Standard Deviation.

In contrast to this observation from the WLS method, Figure 10 illustrates empirical type I error rates for all 2,000 replications from the same cheating condition when using the cumulative logit regression model. With the exception of a handful of outliers, type I error rates were largely consistent and conservative across replications, regardless of the percentage of cheating units included within the replicated data set.

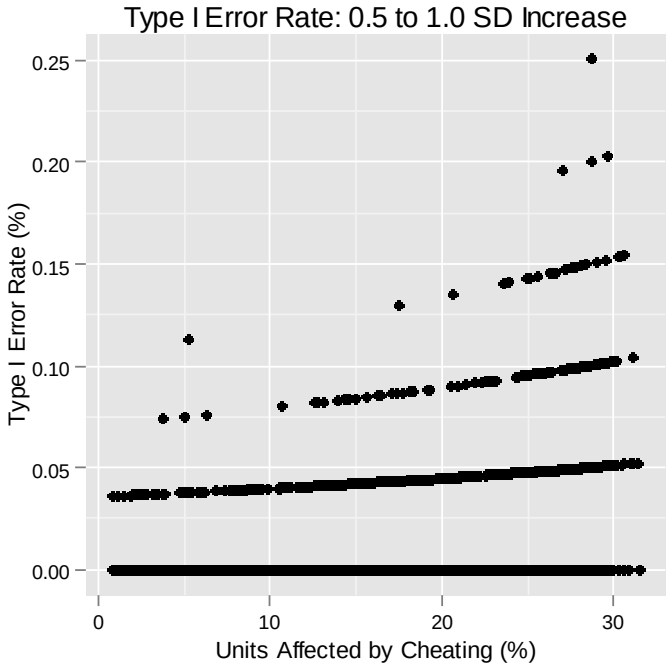

Figure 10. Empirical Type I Error Rates by Percent of Cheating Units for the Cumulative Logit Regression Method: 0.5 to 1.0 Standard Deviation Increase Condition.

Discussion

Overall, results from this study demonstrate that predicting performance level outcomes at the current time point from prior test scores holds value in detecting unusual changes in performance over time. Sensitivity was strong when cheating was not extensive in the population and moderate-to-large effect sizes for misconduct were simulated, and type I error rates remained stable and conservative across conditions. As expected, detection power improved as the effect size of cheating increased. In spite of the poor marginal detection power noted in the 0.5 – 1.0 *SD* score boost condition, positive predictive value results were encouraging throughout, with true cheating units making up more than 98.9% of all flagged units in all investigated cheating conditions.

The impact of the extent of cheating in the population on detection rates is also noteworthy. In all three cheating conditions, detection power degraded as cheating became more

widespread in the test-taking population. This study included theoretical population cheating rates ranging between 1 – 30%, and detection rates substantially declined as population cheating rates approached 30%. This outcome is not altogether surprising, considering that the regression coefficients were estimated from the available data, and the estimation of coefficients will necessarily be influenced by the data used in the prediction model. As misconduct becomes more widely prevalent in the population, any one unit exhibiting performance gains due to misconduct will appear less unusual in comparison to the rest of the population. This outcome is consistent with findings from previous research in the area of person-fit, which similarly found evidence of degrading detection power as cheating became more widespread (e.g., Clark, 2012; Karabatsos, 2003). Although detection power declines as cheating becomes more prevalent in the test-taking population, positive predictive value remains uniformly strong, and type I error rates remain consistently at or below the nominal level.

In addition to being affected by population cheating rates, detection power was observed to vary across unit sizes. Generally speaking, this method had greater success detecting cheating units with larger student populations, with the observed effect most pronounced for conditions with moderate or large impacts of misconduct on test scores. Consistent with all previously-noted findings for this method, positive predictive value remained extremely strong, and type I error rates remained stable, regardless of unit size.

Results from the weighted least squares model provided an informative contrast to outcomes from the proposed cumulative logit regression-based method. When flagging units at a threshold equivalent to the level used for the cumulative logit model, the WLS model showed poor power, with less than 10% of simulated cheating units being correctly identified in all three cheating effect size conditions, although positive predictive value was strong due to extremely

conservative type I error rates in the cheating conditions. Unlike the cumulative logit regression method, which showed a positive relationship between detection power and cheating effect size, detection power was mostly unaffected by the effect size of cheating when using the WLS regression method. Negative predictive value and efficiency were largely steady, both of which remained near 86% in all cheating conditions, in contrast to outcomes of the cumulative logit model, which showed NPV and efficiency climb from approximately 86% up to approximately 94% as cheating effect sizes increased.

Type I error rates for the weighted least squares approach were alarmingly volatile in relation to extent of simulated misconduct in the test data. In simulated cheating conditions, empirical type I error rates were extremely conservative, yet they were very close to nominal levels when cheating was scarce or absent from the simulation. The most extreme example of this observed phenomenon comes from the condition in which units with standardized residuals greater than or equal to 1.0 were flagged. At this flagging level, we would expect to observe an empirical type I error rate of 16%. In the condition in which cheating was simulated to increase examinees' scores between 1.5 to 2.0 standard deviations, the marginal empirical type I error rate was 0.4% at this flagging threshold (although a clear trend relating empirical type I error rates and percentage of cheating units emerged). In the clean simulation condition, in which no cheating was simulated in the test data, an empirical type I error rate of 15.9% was observed.

One great benefit of a conducting a simulation study is having direct knowledge of the true nature of data used in the study, in particular, which units were impacted by cheating and which ones were not. In an application of a weighted least squares regression modeling technique on real-world test data, the investigator would be placed in the unenviable position of selecting amongst several less-than ideal choices. This study demonstrated that a weighted least

squares model has very poor detection power when a flagging criterion of 3.00 is selected. Furthermore, type I error rates will fluctuate greatly depending on the extent of cheating in the population. If the investigator chooses a stringent flagging criterion, detection power will be extremely poor, even when cheating has a large impact on test scores. If a lower flagging threshold is chosen in an effort to improve detection power, the extent of cheating in the population (an unknown value) will have a great deal of impact on the type I error rate. If cheating is moderate to widespread, type I error rates will be very conservative, but if cheating is relatively rare, type I error rates will be near their nominal levels, which may be too unacceptably high to be practical. In addition to evaluating the impact of the extent of cheating on outcomes, issues surrounding model selection bear discussion as well.

Choosing the right predictor variables to include is a very important consideration when building any regression model. In the model proposed in this paper, we included the prior year's test score as the only predictor variable in the model. A valid argument could be made that adding predictor variables, such student gender, ethnicity, or socio-economic status, for example, may account for a significant amount of variability in performance level outcomes at time *t*. Indeed, this consideration in the present context is not unlike discussions on what predictors to include in growth and value-added models. We recognize that the subject of misconduct is a sensitive one for stakeholders, and although there may be valid reasons—from a statistical standpoint—to control for student characteristics in a prediction model such as this, we opted to exclude any such covariates from the model due to concerns for potential face validity issues when sharing results with the public at large. The potential impact of including or excluding such covariates on flagging outcomes is a topic that should be considered in future research in this area.

Limitations

Although it is perfectly reasonable to assume that cheating—when effective—results in some boost to test scores above and beyond what would otherwise be expected in the absence of misconduct, the actual specifics of *how* various cheating behaviors affect scores are not known. If, for example, the classroom teacher shares correct answers from compromised items with students prior to the test administration, it would be reasonable to assume that the score boost achieved due to the misconduct would be influenced by some interaction between the teacher's actions and the characteristics of the students in the classroom. The methodology used to simulate cheating in this study—adding a fixed amount to all scores within a simulated cheating unit—is a rather simple approach toward simulating the impact of cheating on test scores. More research needs to be conducted to better understand how cheating impacts test scores so future simulations can have greater fidelity, particularly when simulating cheating behaviors in which characteristics or behaviors of students and third parties may potentially interact.

Although the positive predictive value for the cumulative logit method was excellent in all cheating conditions, marginal sensitivity was poor in the condition in which test scores within cheating units were boosted between 0.5 and 1.0 standard deviation. This proposed cumulative logit method was anticipated to outperform the weighted least squares approach when small cheating effects were simulated, but both methods showed poor sensitivity in this condition. It is worthwhile to consider that the cheating effect simulated in this condition resulted in increases of only 3 to 6 raw scores, which would likely prove difficult for any method to detect consistently. Future research to further explore this proposed method will include a "targeted" cheating condition, in which test scores within cheating units are boosted just enough to reach the performance level associated with proficiency. Including such a condition in a future study will

better address the question of how well a method like this identifies small to moderate cheating—in terms of score boosts—that has a large impact on performance level classification rates at the unit level.

Conclusion

The major purpose of this study was to evaluate the performance of a novel technique for identifying unusual unit-level performance changes under a variety of circumstances. We investigated a wide range of cheating effect sizes and extent of cheating within data sets. As expected, detection power varied as a function of the effect of cheating on test scores, with the method having greatest success identifying cheating that is most impactful on students' test scores. Detection power also was observed to vary as a function of how widespread misconduct appeared in the data, with detection power being greatest when cheating is a relatively isolated occurrence within the test-taking population. Observed marginal positive predictive value results were encouraging throughout, with true cheating units making up more than 98.9% of all flagged units in all conditions at the investigated flagging criterion. The longitudinal modeling technique proposed in this paper shows potential to provide a useful framework for comparing observed with expected outcomes and flagging those cases where observed performance greatly exceeds what is expected, as defined by the prediction model.

References

Agresti, A. (1996). *An introduction to categorical data analysis.* Hoboken, NJ: John Wiley & Sons.

Amrein-Beardsley, A., Berliner, D.C., & Rideau, S. (2010) Cheating in the first, second, and third degree: Educators' responses to high-stakes testing. Educational Policy Analysis Archives, 18. Retrieved from http://epaa.asu.edu/ojs/article/view/714

Betebenner, D. (2009), Norm- and Criterion-Referenced Student Growth. *Educational Measurement: Issues and Practice, 28,* 42–51.

Cheating our children: The AJC's methodology behind suspicious school test scores (2012, March 27). *The Atlanta Journal-Constitution.* Retrieved from http://www.ajc.com/news/news/local/cheating-our-children-the-ajcs-methodology-behind-/nQSTN/

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cizek, G. J. (2001). An overview of issues concerning cheating on large-scale tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Clark, J. M. (2012). Nested factor analytic model comparison as a means to detect aberrant response patterns. Paper presented at the Conference on Statistical Detection of Potential Test Fraud, Lawrence, KS.

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

Molland, J. (2012, Oct. 19). Texas guide to raising test scores: "Disappear" your students. Retrieved from http://www.care2.com/causes/texas-cheating-scandal-disappearing-students-to-improve-test-score.html

National Center for Fair and Open Testing (2013, March 27). *Standardized exam cheating in 37 states and DC; New report shows widespread test score corruption.* Retrieved from http://www.fairtest.org/2013-Cheating-Report-PressRelease

National Council on Measurement in Education (2012). *Testing and data integrity in the administration of statewide student assessment programs.* Retrieved from http://ncme.org/default/assets/File/Committee%20Docs/Test%20Score%20Integrity/Test%20Integrity-NCME%20Endorsed%20(2012%20FINAL).pdf

Skorupski, W. P. & Wainer, H. (2013). The "P" you really want to know: Why you should detect cheating the Bayesian way. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Vogell, H. (2011, July 6). Investigation into APS cheating finds unethical behavior across every level. *The Atlanta Journal-Constitution.* Retrieved from http://www.ajc.com/news/investigation-into-aps-cheating-1001375.html