A Hierarchical Linear Modeling Approach for Detecting Cheating and Aberrance

William Skorupski

University of Kansas

Karla Egan

CTB/McGraw-Hill

Paper presented at the May, 2012 Conference on the Statistical Detection of Potential Test Fraud

Lawrence, KS

Abstract

The purpose of this study was to demonstrate, through Monte Carlo simulation, the utility of a newly presented method for detecting group-level cheating and aberrance (Skorupski & Egan, 2011). The method relies on vertically scaled test scores over grades levels. Using these data, the change in individual scores, nested within groups (classrooms or schools) over time may be modeled. The approach is based on a hierarchical linear model (HLM), and evaluates unusually large group-by-time interaction effects as evidence of potential cheating or aberrance. The authors have previously demonstrated this method using real data from a large, statewide, testing program. Some external evidence of suspected cheating was available and used to cross-validate those schools flagged as potential cheaters. This approach provided some encouraging success, but because real data were used, the accuracy of the method couldn't be demonstrated. The purpose of the current study was to directly evaluate how well the method can identify simulated groups which are known a priori to demonstrate such aberrance (which could be evidence of cheating or something else unusual). This evaluation was conducted by fitting a fully Bayesian HLM and considering marginal recovery of the known parameter values from the model, as well as a determination of power and Type I error rates for identifying aberrant versus non-aberrant groups. Results indicate an acceptable Type I error rate for non-aberrant groups, with relatively high power for aberrant groups.

Introduction

Cheating on statewide assessments (SWAs) has gone on for decades. Current and former state directors of assessment can share amusing, startling, and even depressing anecdotes of the perpetrators that have been caught in their particular state. Based on a review of literature, Thiessen (2007) estimates that 25% of educators cheat on standardized tests; this cheating may involve anything from subtle forms of cheating (e.g. teaching to the test) to blatant forms of cheating (e.g. changing student answer documents). This figure provides a disquieting perspective of the validity of the student response data used to make many important decisions. Cheating on SWAs may have serious implications for the psychometric integrity of item parameters and test scores as well as the validity of how those test scores are used.

There are a myriad of methods in the literature to detect possible incidences of cheating on SWAs, such as detecting similar response string patterns (e.g., Wollack, 1997, 2003) and analyzing person-fit data (e.g., Levine & Rubin, 1979; Drasgow, Levine, & Williams, 1985). See Skorupski and Egan (2010) for a comprehensive review of these and other methods. These techniques have tried to detect cheating at the student level by means of detecting answering copying or aberrance, which is how individual cheaters cheat. However, cheating on SWAs may likely occur at the teacher or even the school level. Students have little incentive to cheat on SWAs when those tests are not tied to student grades, retention, budgets, or graduation policies. Teachers and administrators, however, may be motivated to cheat on the SWA because test results are often used for teacher, school, and/or district accountability. In some cases, teacher merit pay is tied to test results. Under adequate yearly progress (AYP) standards associated with NCLB, schools and/or districts may be shut down or taken over by the state based on test results.

The threat of not making AYP provides a large incentive for educators to cheat. In many states, AYP is based, in part, on the percentage of students reaching proficiency. This type of status model makes it relatively easy for educators to game the system. So long as educators ensure a certain percentage of students do well on the SWA, then they should make AYP (assuming that other criteria are met). Other states have started using growth models when estimating AYP. On the surface, it seems that it may be harder to game a growth model, which is based on individual student growth across years. In either case, however, the temptation to cheat must be great if it could make the difference between keeping or losing one's job.

A limited number of researchers have examined cheating at the classroom level. Jacob and Levitt (2004) examined fluctuations in student test scores and similarity of answer patterns to detect classroom-level cheating incidents at a Chicago area school. Their work uncovered cheating incidences in four to five percent of classrooms studied. Their indices, however, require that students can be grouped by teacher. Schools and districts often do not report the students' teachers. This is especially true in the upper grades where students have different teachers for each content area, and test materials allow only for a single teacher (such as a homeroom teacher) to be entered. Recently, Skorupski and Egan (2011) presented a statistical method for detecting possible group-level cheating using a Bayesian Hierarchical Linear Model (HLM). Using real data from a vertically scaled SWA, they modeled the change in individual scores, nested within groups (schools) over three years. Unusually large group-by-time interaction effects (i.e., high performance not explained by group and/or time marginal effects) were treated as evidence of potential cheating or aberrance. Some external evidence of suspected cheating was available and used to cross-validate those schools flagged as potential cheaters. This approach provided some encouraging success, but because real data were used, the accuracy of the method couldn't be demonstrated.

Statement of Problem/Purpose

The purpose of the current study was to directly evaluate how well this Bayesian HLM approach can identify simulated groups which are known a priori to demonstrate such aberrance (which could be evidence of cheating or something else unusual). The ultimate goal is to develop and validate a reliable method for identifying group-level cheating behavior, such as inappropriate coaching, widespread use of improper materials during testing, or answer changing by teachers. It is hypothesized that any such group-level cheating behavior would not manifest itself using traditional individual-level cheating detection procedures. These aberrances would likely only be detectable at a group level. This evaluation was conducted by simulating vertically scaled test scores for sixty groups over three test administrations. A fully Bayesian HLM was fit to each replicated dataset, with parameter recovery of the group-by-time interaction effect used to evaluate cheating detection. Groups simulated to display aberrant behavior were used for power analyses, non-aberrant groups were used to evaluate Type I error rates. An advantage of using the fully Bayesian framework was that stochastic inferences about cheating likelihood (given an aberrance criterion) could also be obtained. These methods are explicated below.

Method

Simulated Data

Data were simulated for this study to appear similar to vertically scaled test scores over three administration years. Based on a previous study using real data (Skorupski & Egan, 2011), groups and sample sizes within groups were created to be proportional to a geometric distribution. Figure 1 contains a histogram of school sample sizes from the Skorupski and Egan (2011) study.

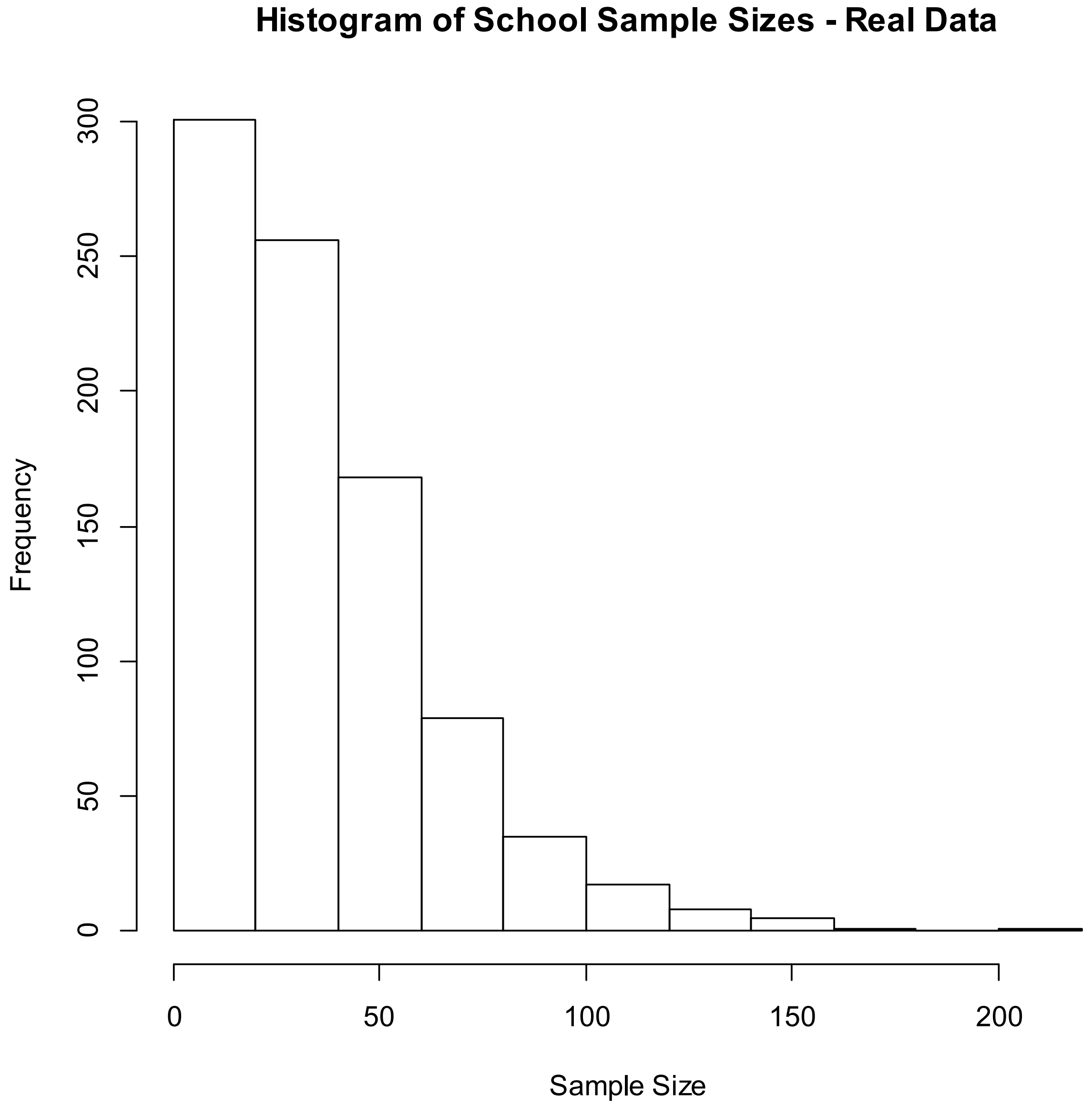


Figure 1. Histogram of school sample sizes for real data (total N of groups = 781) studied in Skorupski & Egan (2011).

For the present study, sixty groups were simulated, with group samples sizes chosen to be proportional to what was observed in the real data. Figure 2 contains a histogram of these group sample sizes. Fewer groups were used for computational purposes, but a large enough number of groups was included to simulate a realistic HLM analysis. Simulated group sizes ranged from N_g

= 10 (the minimum sample size used in Skorupski & Egan (2011)) to N_g = 260, with the total sample size equal to 4,650.

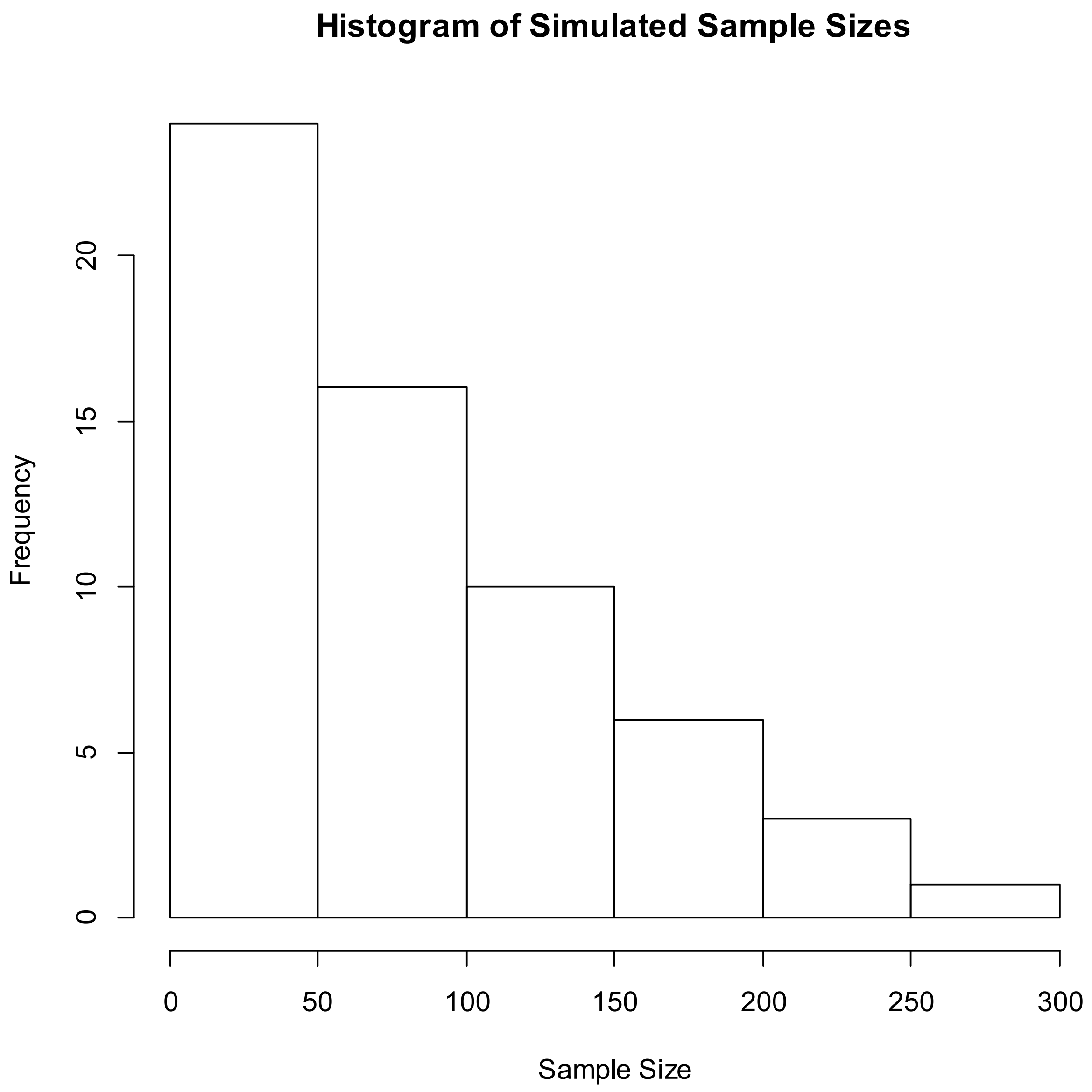


Figure 2. Histogram of simulated group sizes (total N of groups = 60), chosen to be proportional to real data studied in Skorupski & Egan (2011).

These 4,650 “examinees” within 60 groups were simulated over three time points, with a mean increase of half a standard deviation from one administration to the next. These values were chosen based on what was observed in real data from Skorupski & Egan (2011). A histogram of group means at Time 1 are contained in Figure 3.

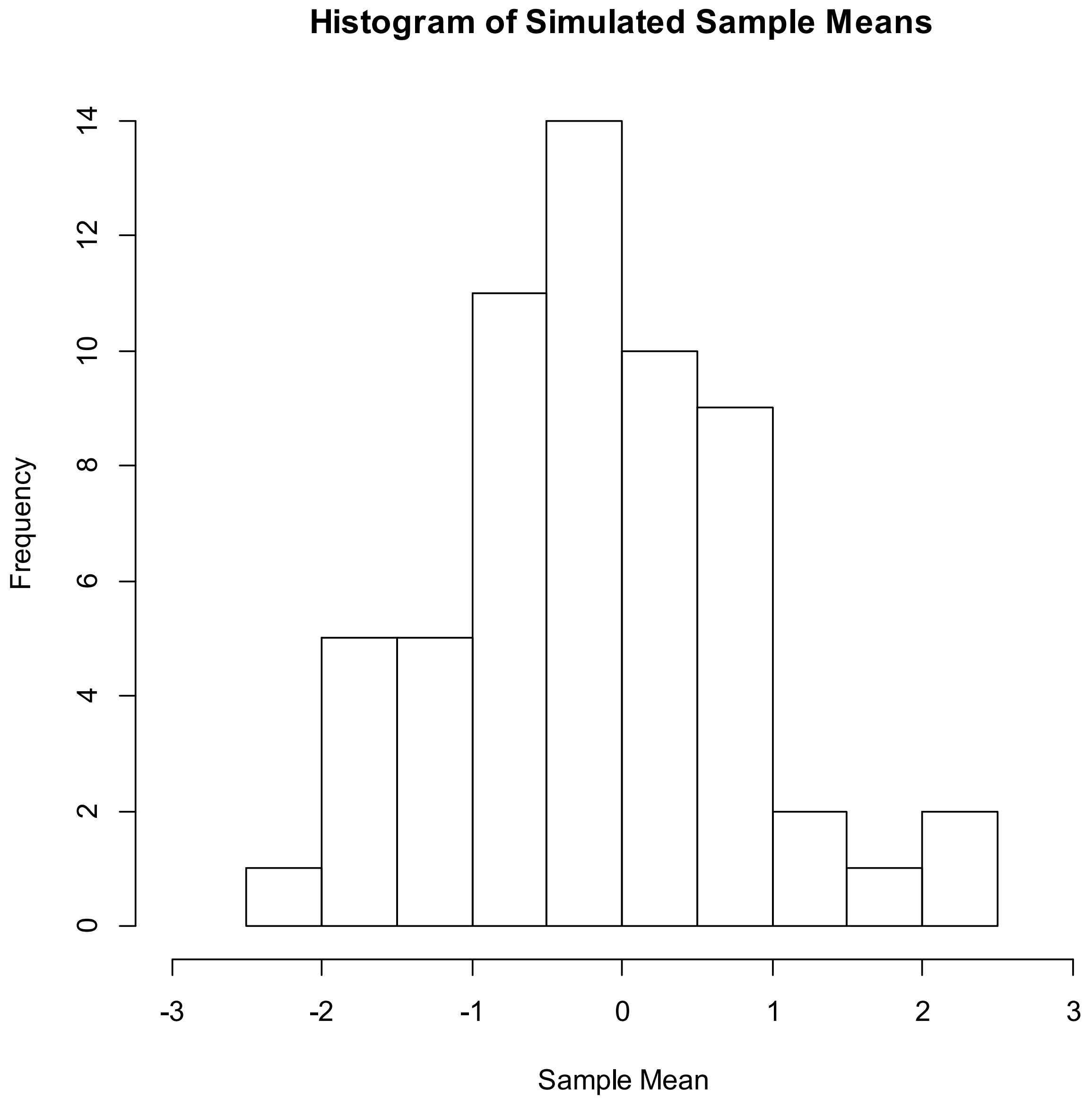


Figure 3. Histogram of simulated group means at Time 1 (total N of groups = 60).

Fifty-one means of these Time 1 means were randomly sampled from a standard normal distribution; three schools were assigned group means of -1, three others were assigned means of 0, and three other groups were assigned means of 1. These nine groups were simulated to demonstrate cheating/aberrant behavior (each of the three group means represented at each of the three time points).

Individual scores were simulated directly by sampling from a multivariate normal distribution, $\underline{\theta} \sim \text{MVN}(\underline{0},\mathbf{R})$[1], where $\underline{0}$ is a vector of zeros, and **R** is a correlation matrix with 0.77 on all off-diagonal elements. This constant correlation was chosen based on what was observed in real data from Skorupski & Egan (2011). After individual scores over three time points were simulated, each was altered by adding each individual's respective Time 1 Group mean and respective time point mean (Time point means: $\mu_1 = 0$, $\mu_1 = 0.5$, $\mu_1 = 1$). Finally, a group-by-time interaction effect was added to each individual's score. A group-by-time (60 x 3) matrix of interaction effects was created to simulate cheating/aberrance. If an element of this matrix was zero, then individuals within that group-time combination had scores which were explainable by main effects (Group and Time) alone. Any non-zero value in this matrix would produce aberrance. In the 60 x 3 matrix of interaction effects, exactly nine (5%) were selected to demonstrate aberrance, by replacing a coefficient of zero with one (i.e., scores at this time point for this group were one standard deviation larger than the main effects would predict, a "large" effect, which would be expected if caused by some purposeful cheating behavior). For each time point, three groups were selected for aberrance, one for each of the three non-randomly-sampled Time 1 group means (-1, 0, 1). Each of these group means was assigned to groups of varying size, N_g = 10, 60, or 110 to emulate "small," "medium," and "large" groups. Thus, at each time point three out of 60 groups were aberrant, so the aberrance rate was 5% at each time point, and across all time points, which has been observed in previous studies (Jacob & Levitt, 2004; Skorupski & Egan, 2001).

[1] These scores were treated as observed scores in the Bayesian HLM (that is, they were treated as being perfectly reliable). This was done to present a "best-case scenario" for the approach. However, a measurement model could easily be nested within this methodology if it were desirable to account for unreliability in the estimation.

Figure 4 presents a graphic representation of five of the 60 simulated groups over the three time points (a subsample chosen to avoid clutter in the figure). The values represented in this figure are true group means, and don't represent individual variability within group.

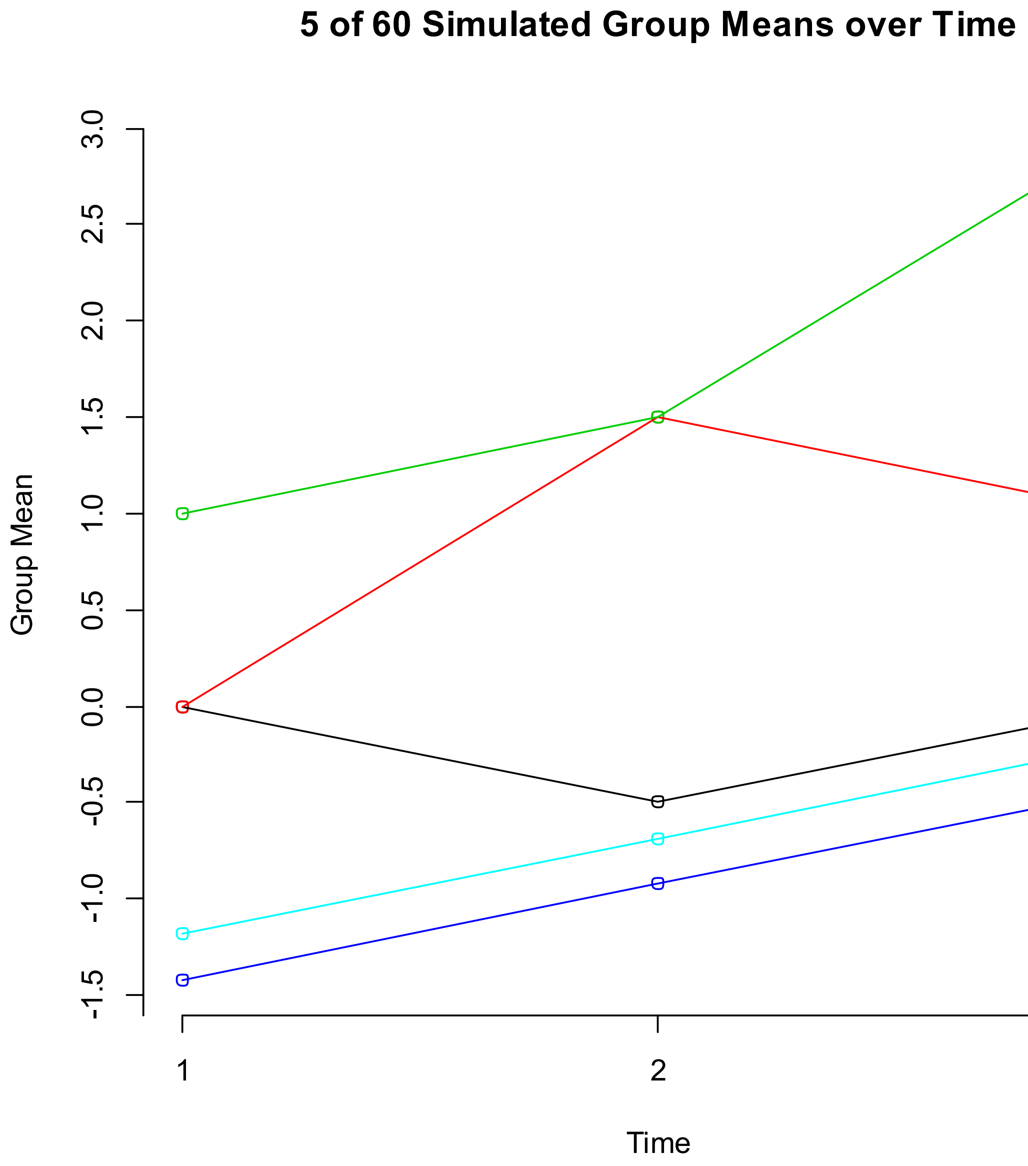


Figure 4. *Group means over three time points for five of the 60 simulated groups.*

Groups indicated by Blue and Teal lines are non-cheating/aberrant (i.e., main effects of Group and Time account for all growth). Groups indicated in Black, Red, and Green are simulated to be cheating/aberrant. The Black group has a Time 1 mean of -1, but demonstrates cheating/aberrance at Time 1, so this mean was actually 0. No cheating/aberrance was simulated at Times 2 and 3, so these individuals subsequently revert back to their main effects (thus, they decline at Time 2 and return to "baseline" at Time 3). The Red group has a Time 1 mean of 0, and demonstrates cheating/aberrance at Time 2. So, while other groups increase 0.5 standard deviations from Time 1 to Time 2, this group increases 1.5 standard deviations. No cheating/aberrance was simulated at Time 3, so these individuals subsequently revert back to their main effects and thus appear to decline. The Green group has a Time 1 mean of 1, and demonstrates cheating/aberrance at Time 3. They increase 0.5 standard deviations from Time 1 to Time 2, but then increase 1.5 standard deviations from Time 2 to Time 3.

Each of these datasets, containing 4,650 individuals nested within 60 groups over three simulated time points, was replicated 50 times. Each of these replicated datasets was analyzed with the fully Bayesian HLM approach (described in the following section). Results were averaged over replications to insure the reliability of results.

Analysis

A hierarchical growth model was fit to each of the 50 replicated datasets, with scores over time nested within students, who in turn are nested within groups[2]. The parameters of the HLM were estimated within a fully Bayesian network, implemented with Markov Chain Monte Carlo (MCMC) techniques using the freeware WinBUGS 1.4 (Lunn, Thomas, Best, & Spiegelhalter, 2000). The basic premise of a hierarchical model is to establish a series of nested equations, wherein independent variables from one level of the model become dependent variables at the next level. The complete model for an individual's score at time point *t* (indicated by "T") nested within student *i* within Group *g* (indicated by "G") was as follows:

[2] As a practical consideration with real data, the analysis is more robust if scores for examinees are available for all Time points, and if they stay in the same Group for all administrations. However, dynamic group-changing and missing data extensions to this approach are possible.

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

where Y_{igt} is the (vertically linked) score of student *i* in Group *g* at time *t*, *i*=1,…,*N*(*g*), the number of individuals in Group *g*, *g*=1,..,60, the total number of groups, and *t*= 1,2,3. The coefficients in the model are indexed to indicate their level of the hierarchy: β_0 is the common intercept (the grand mean), and all other effects are centered around it. β_{1g} is the main effect for Group, accounting for the average performance of groups around the grand mean, β_{2t} is the main effect for Time, accounting for the average rate of growth, β_{3gt} is the interaction effect for each Group by Time, accounting for any unique effects occurring for a particular group at a particular time that cannot be explained by the main effects, and ε_{igt} is a random error term, reflecting individual variability within Group-by-Time clusters. Thus, this model can capture all of the variability of test scores over time, nested within students within groups. When Y_{igt} is replaced with its expected value in the model, the random error term drops out (its expected value is zero). Thus, β_{3gt} in the model will represent the interaction effect plus any additional random error, the traditional approach in such ANCOVA-type designs. Thus, the "signal" of this effect should only be detected if it is stronger than the "noise" created by within-group variability. Using this approach one can monitor individual growth of students within schools over time, effects which might help us evaluate potential sources of cheating evidence. It should be noted that this model, as specified, only considers linear trends, which was appropriate given the simulation. However, other applications of this method could certainly consider quadratic growth as a possibility (or higher-order polynomials if more time points were observed).

After fitting the model, posterior distributions were evaluated for the convergence of their solutions. The successful convergence of the models to the data from the MCMC processes was assessed using techniques suggested by Gelman, Carlin, Stern, and Rubin (1995). Parameter estimates from WinBUGS were used to evaluate these Group, Time, and Group-by-Time interaction effects. Large, positive values for the β_{3gt} coefficients are indicative of cheating/aberrance behavior. That is, if a group were to perform unusually well at a given time, relative to that time's mean and the group's own average performance, this coefficient could be "flagged" for further review, as some unusual group-level aberrance has occurred (whether this performance is commendable or condemnable would have to be independently investigated). Furthermore, the β_{1g} coefficients (group level main effects) could also be monitored for evidence of unusually high performance across all three years. That is, if a school's performance was consistently very high, its β_{1g} coefficient would be large, though its β_{3gt} coefficients would be close to zero. Because aberrance was only simulated at individual time points, these were not considered in this study, but they could be used in real data analysis. Such a pattern would indicate systematically high performance, which could represent excellence, but may also be consistent with a pattern of persistent group-level cheating behavior.

The delta (δ) statistic (Cohen, 1988) is a simple, standardized measure of an effect size. It does not incorporate information with regard to sample size. The result is that δ is equal to an estimate of how many standard deviations different a group's performance is from an expected baseline level. Its general form is to divide an observed difference by an estimate of the population standard deviation. In order to better interpret the magnitude and practical significance of the model parameter estimates, δ statistics were computed for each group-based coefficient, the β_{1g} values for group main effects and the β_{3gt} values for each Group-by-Time interaction effect.

$$\delta_g = \frac{\beta_{1g} - 0}{\sqrt{a \sum a'}} \text{ and } \delta_{gt} = \frac{\beta_{3gt} - 0}{\sqrt{\sigma_t^2}} \ ,$$

where a is a 1x3 row vector with each element equal to 1/3, $a = [1/3, 1/3, 1/3]$, $\mathbf{\Sigma}$ is the 3x3 variance-covariance matrix of scores over the three Time points (thus the denominator for δ_g is the standard deviation of scales scores averaged over time points), and σ_t^2 is the t^{th} diagonal element of $\mathbf{\Sigma}$. Following standard conventions proposed by Cohen (1988), $\delta \geq 0.8$ was considered a large effect and used as criterion for flagging groups as potential cheaters (obviously, other choices could be used here).

The parameters of these models could be evaluated using any software capable of handling multilevel data (e.g., HLM, or SEM software like LISREL, for example). However, this model was fit using WinBUGS 1.4 (Lunn et al, 2000) in order to take advantage of the stochastic inference MCMC output can provide. Specifically, it was desirable to obtain estimates of the Group-by-Time interaction effects (β_{3gt}) to evaluate potential outliers which might be construed as “unusual” (and therefore possibly cheating). An MCMC algorithm produces random draws from the posterior distribution of each parameter being estimated. As such, these values can be used to make any sort of probabilistic inference desired, without having to assume a known density function for the parameter (which would be necessary in order to make similar kinds of inferences using parameter estimates and standard errors from a maximum likelihood estimation procedure). For example, one can iteratively test a series of potential “cutscores” for flagging potentially cheating groups by specifying a baseline expected value and determining the posterior probability that a group’s performance is unusual compared to this. This probability is easily calculated once the MCMC output is available: all one needs to do is set the cutscore, count the number of random posterior draws that appear above that threshold, and divide that number by the total posterior draws available.

This approach is methodologically quite similar to a method developed by Wainer, Wang, Skorupski, and Bradlow (2005) for evaluating the reliability of pass/fail decisions. Wainer et al (2005) define the Posterior Probability of Passing for an examinee as the proportion of posterior draws at or above a given cutscore. Adapted from this, the current approach is to define the Poster Probability of Cheating (PPoC) for a group based on the proportion of posterior draws at or above a given threshold. For this study, a baseline was established by treating the threshold as zero, that is, PPoC is equal to the proportion of posterior samples greater than zero.

Cross Validation of Cheating Detection

Based on analyses in Skorupski and Egan (2011), both the delta statistic and PPoC were used to flag potentially cheating groups. To avoid unnecessary Type I error, relatively strict criteria were employed for detection. Group-by-Time interaction effects were flagged as cheating/aberrant if δ_{gt} was greater than or equal to 0.5 and PPoC_{gt} was 0.75 or greater (indicating a group/time mean at least 0.5 standard deviations above expected, with a 75% or greater chance of being a greater-than-zero effect). These values were selected based on real-data analyses from Skorupski and Egan (2011), which maintained an overall detection rate of less than 10%.

Summary of Analysis Steps

Operationally, implementation of this procedure should take place as a series of eight steps, detailed below:

1. CALIBRATE. Conduct an Item Response Theory (IRT) calibration of item response data.
2. LINK. Use anchor item parameters from the previously established vertical scale to link examinee scale scores on a common metric.

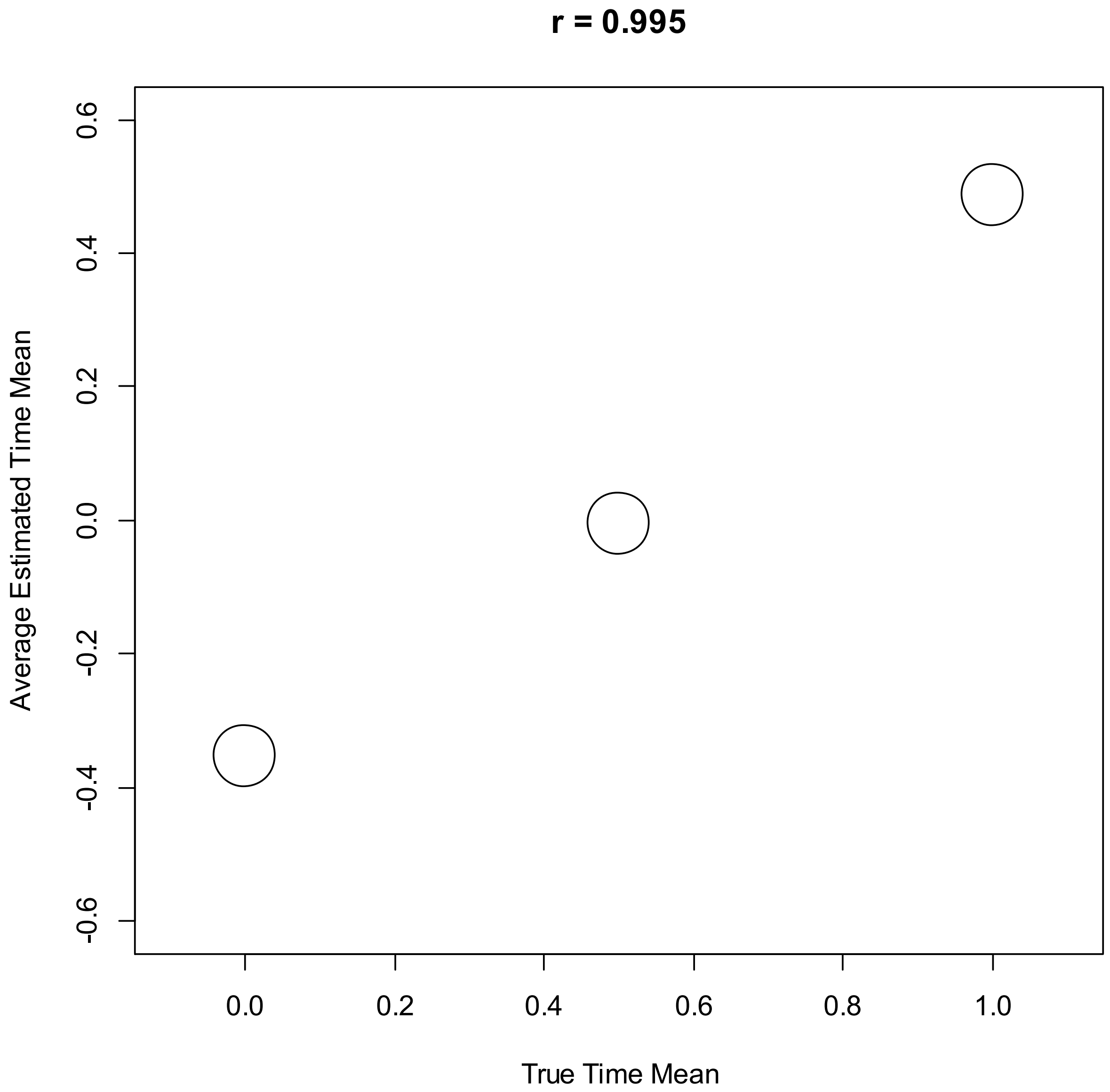
3. MODEL: A Bayesian hierarchical growth model is fit to the scale scores, and parameter estimates and posterior draws from the MCMC algorithm are retained. Note: Steps 1 - 3 could be combined into a single analytic step.
4. CLEAN. Any groups with N<10 (or another criterion) are removed from consideration.
5. FLAG. Groups with parameter estimates that have δ and PPoC statistics greater than criterion levels are flagged as potential cheaters.
6. VALIDATE. Any group flagged for aberrance is compared to any previously established examples of documented security breaches which have occurred.
7. EXPLORE. Examining the sampling distributions of posterior draws from parameter estimates may help to refine a process for when to flag a group's performance.
8. INVESTIGATE. Any group ultimately flagged as potentially cheating would need to be investigated. Such statistical criteria are never "proof" of cheating. It is even possible that these criteria could alternatively be used to identify exemplary groups. Conversely, group estimates in the extreme opposite direction (i.e., decreases relative to baseline) might be used to identify "at risk" groups.

Results and Discussion

Based on MCMC convergence diagnostics, results indicate that the parameters of the hierarchical growth model converged to stable solutions[3]. Parameter recovery was very good for these analyses. Figure 5 contains a scatterplot of true Time means versus average Time mean over replications. It is readily apparent that these true and estimated values are highly correlated (r = 0.995). The scale of the estimates is different than that of the true due to the indeterminacy

[3] Additional MCMC details: Parameters for each of the 50 replicated datasets were estimated by creating two independent Markov Chains for every parameter, each of which was 30,000 iterations long, with a burn-in of 25,000. Retained posterior draws from these chains all represented converged solutions.

of the overall metric (i.e., the “zero” was at Time 1 in the simulation, but was placed at Time 2 for the estimates), but this has no effect on the relative position of groups.



Title 5. Scatterplot of true Time means by average estimated Time means.

Figure 6 contains a scatterplot of true Group means versus average estimated delta values over replications. It is also apparent that these true and estimated values are highly correlated (r = 0.95). The scale of these estimates is more in line with the generating metric. It is also clear that group mean recovery was not a function of sample size, and that the nine groups simulated for aberrance (circles indicated in Red) all had their marginal means well recovered.

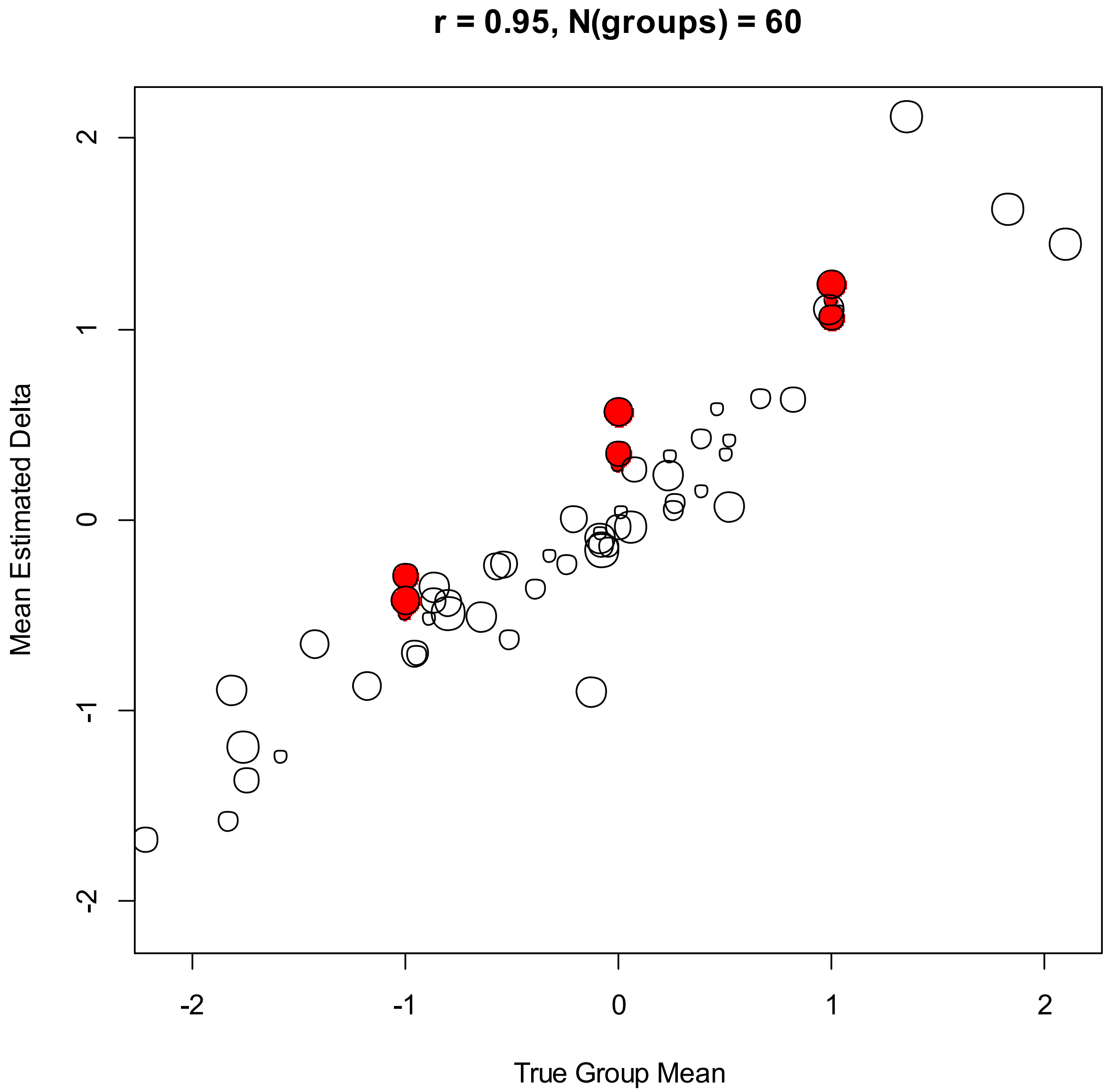


Figure 6. Scatterplot of true group means by estimated delta values. Circle sizes are proportional to sample size (N) within group. Circles shaded in red indicate groups simulated with aberrant/cheating behavior at one of three times.

Figure 7 contains a scatterplot of mean estimated delta by mean PPoC values for all 60 x 3 interaction terms (the cheating/aberrance indicators). Not surprisingly, there is a strong, somewhat curvilinear relationship between these indicators (delta has no lower/upper bounds, while PPoC is bounded by 0 and 1).

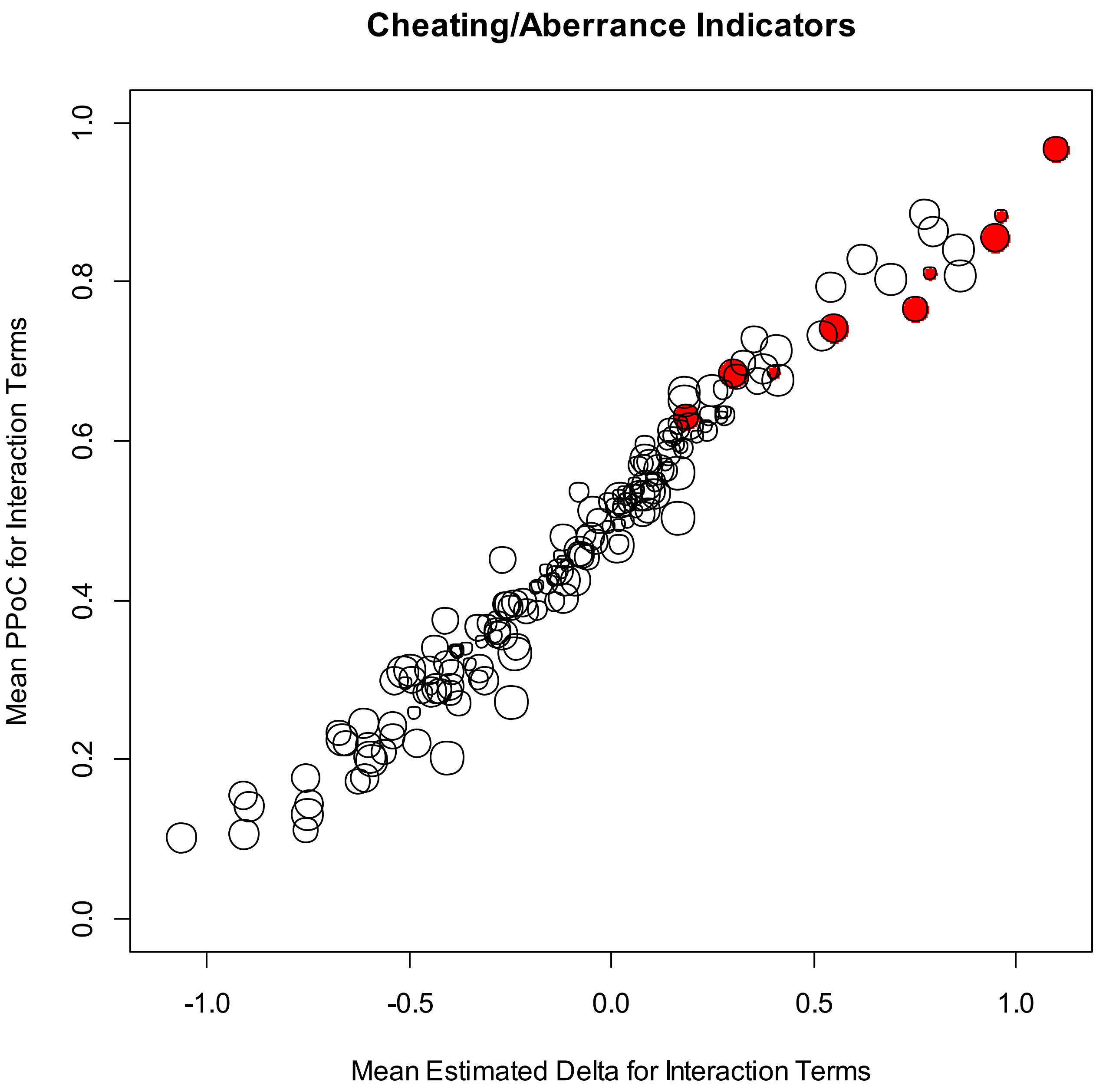


Figure 7. *Scatterplot of mean estimated delta by mean PPoC value for interaction terms across three time points. Circle are proportional to group size. Those shaded in red indicate groups simulated with cheating/aberrant behavior at one of three times.*

It was very encouraging to see that all of the nine simulated aberrant groups (Red circles) were located in the upper right quadrant of this graph, with relatively few non-aberrant groups present there. These results are further broken down, and better understood, by evaluating cheating/aberrance detection at each time point separately. Figures 8 through 10 contain these results.

Figure 8 contains a scatterplot of the mean estimated delta by mean PPoC values (over replications) at Time 1. Using criteria that groups would be flagged as potential cheaters if δ_{gt} >= 0.5 and PPoC $_{gt}$ >= 0.75, the detection power at Time 1 was only 0.07. The Type I error rate was 0.04, close to a reasonable accepted value of 0.05. Figure 9 contains a scatterplot of mean estimated delta by mean PPoC values at Time 2. Using the same criteria, the detection power at Time 2 was 0.71, a considerable improvement over the Time 1 power rate. As with the Time 1 restults, the Type I error rate was 0.04, close to a reasonable accepted value of 0.05. Lastly, figure 10 contains a scatterplot of mean estimated delta by mean PPoC values at Time 3. Using the same criteria, the detection power at Time 3 was 1.0 (perfect identification), while maintaining a Type I error rate of 0.05.

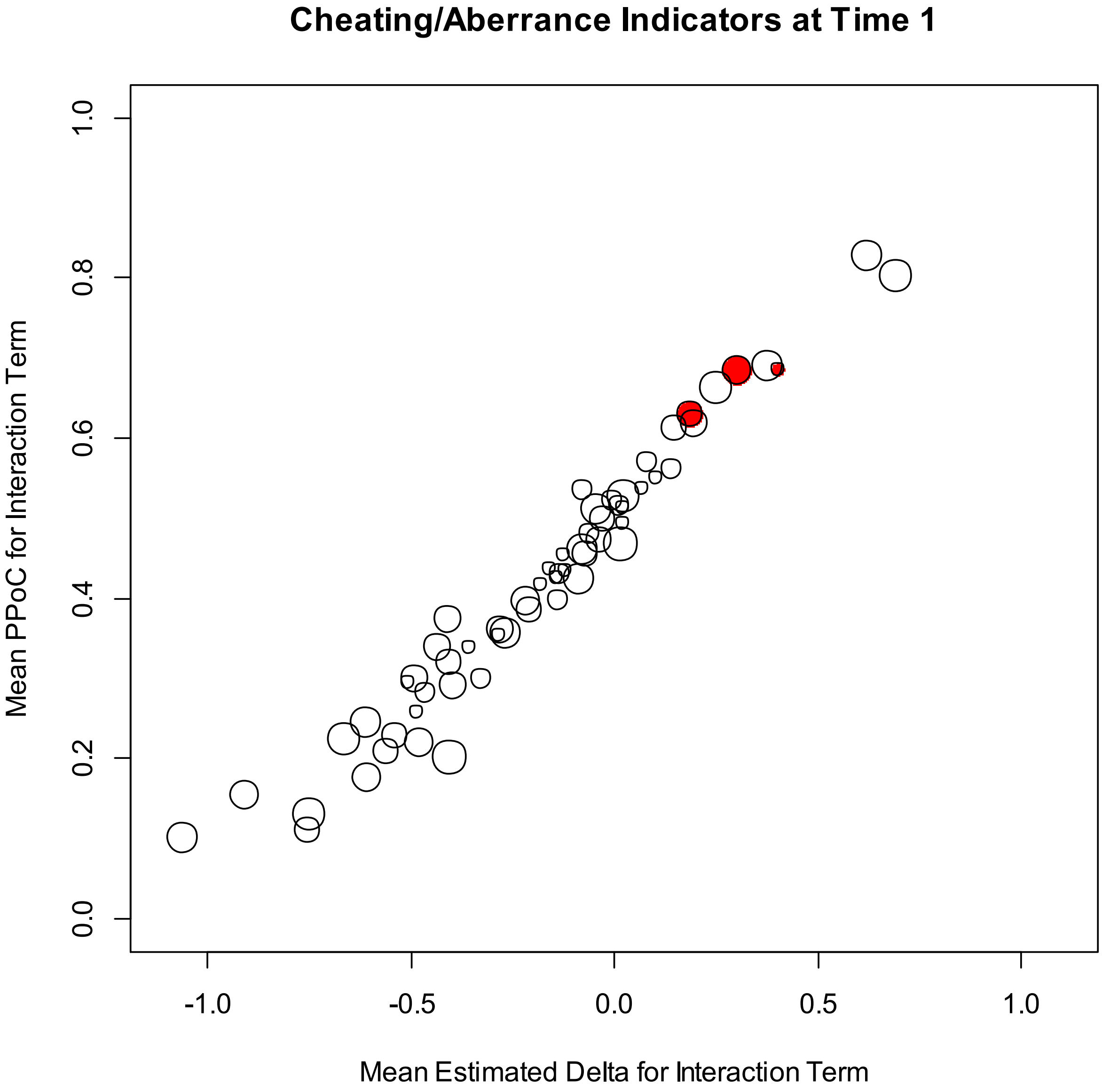


Figure 8. *Scatterplot of mean estimated delta by mean PPoC value for interaction terms at Time 1. Circle sizes are proportional to group size. Circles shaded in red indicate groups simulated with cheating/aberrant behavior.*

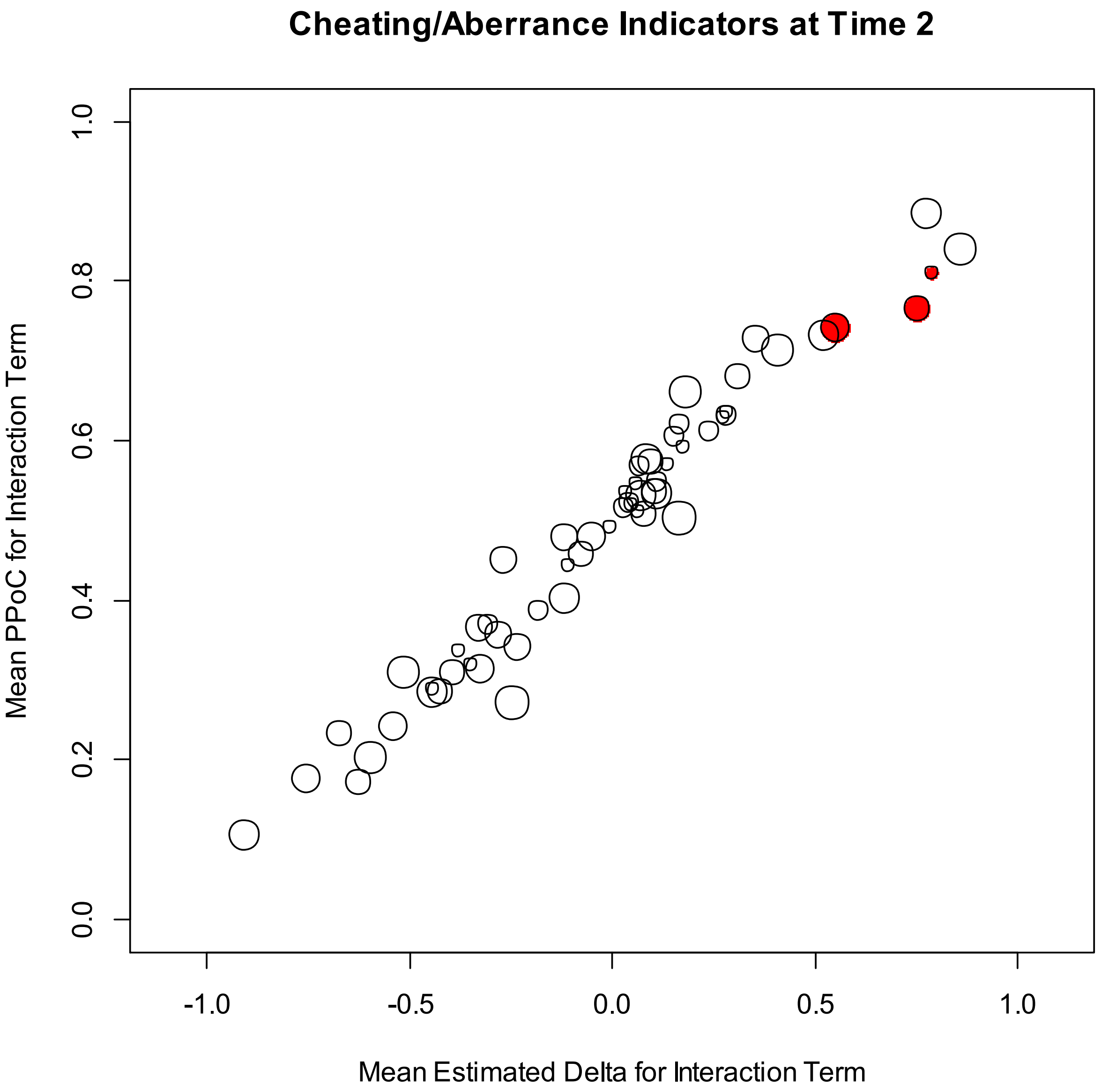


Figure 9. *Scatterplot of mean estimated delta by mean PPoC value for interaction terms at Time 2. Circle sizes are proportional to group size. Circles shaded in red indicate groups simulated with cheating/aberrant behavior.*

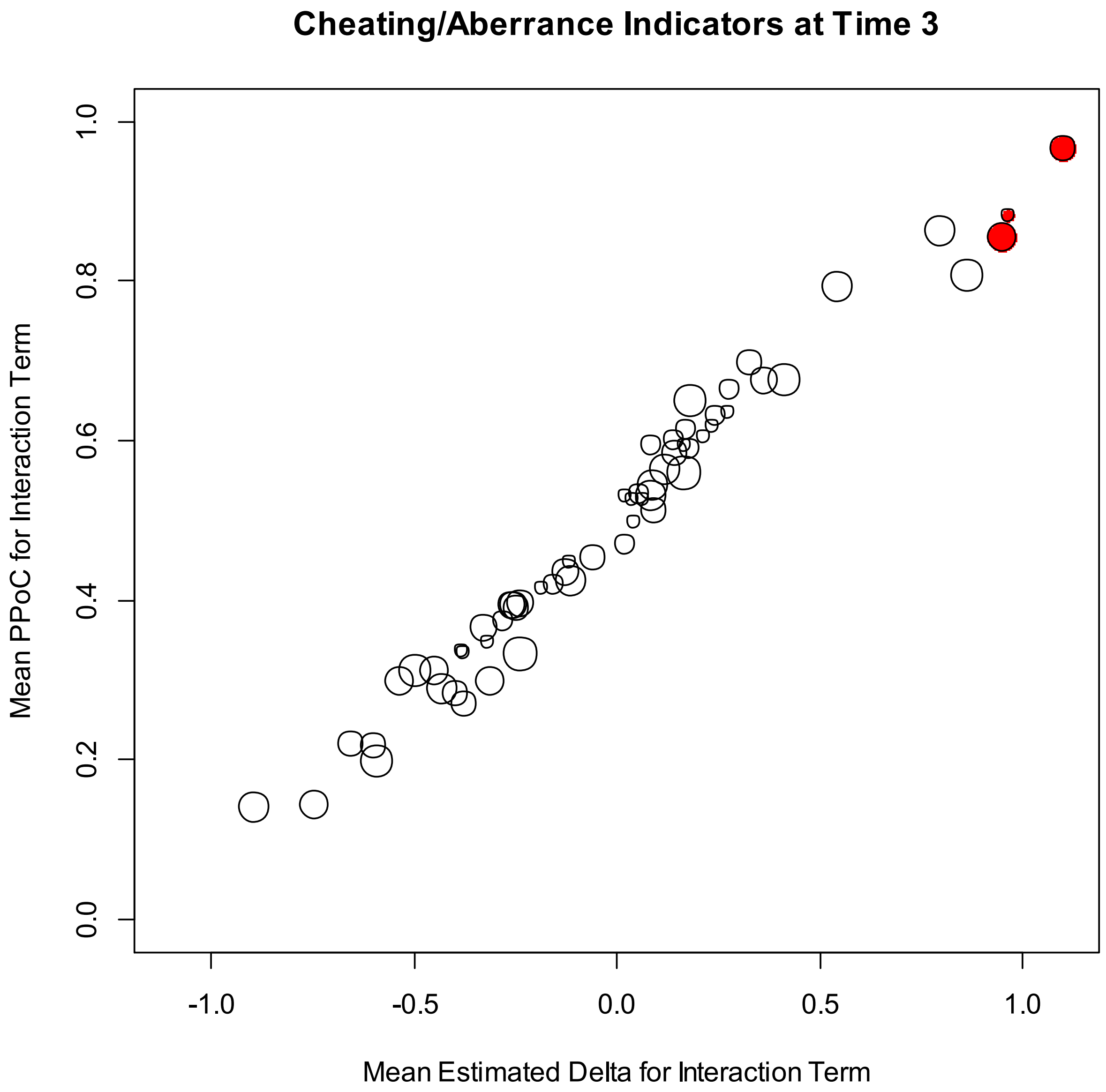


Figure 10. Scatterplot of mean estimated delta by mean PPoC value for interaction terms at Time 3. Circle sizes are proportional to group size. Circles shaded in red indicate groups simulated with cheating/aberrant behavior.

Conclusion

Overall, this Bayesian HLM methodology appears to have great promise for detecting groups of individual demonstrating cheating or aberrant behavior. The group-level main and interaction effects, once standardized in δ statistics, provide a straightforward way to conceptualize group-level aberrance. The PPoC provides additional insight into the probability of this aberrance being a non-zero effect. The combination of δ and PPoC values seems to be a reasonable way of achieving good detection power while maintaining very reasonable Type I error rate. Using some reasonable criteria, we have demonstrated how these statistics might be used to flag groups as potential cheaters. Detection of aberrance was very good for Times 2 and 3, but quite low for Time 1. It appears that this is due to the nature of the simulation of aberrance. Aberrance at Time 1 made the growth trajectory overall look fairly flat, which tended to change the estimate of the group mean, but fail to detect the interaction effect. It is at least encouraging that these interaction effects were among the largest of those estimated, but overall the interaction effects at Time 1 were the smallest. Ultimately, if the baseline (Time 1) for the group mean has been affected, it may be very difficult to detect cheating and aberrance. Additional research into this phenomenon seems warranted. However, it also appears that if cheating or aberrance occurs at a time point after an established baseline, this methodology is very effective at flagging potentially cheating groups. For real data, one would expect these interaction effects to all be close to zero, if group- and time-level main effects were adequate to explain performance. However, for any analysis like this one, a statistical procedure can only produce a flag that *implies* aberrant performance, it can never *prove* that, for example, teachers in a school, are cheating. Thus, this procedure, like any other, will still need to be validated by personally investigating those schools which are flagged. The Bayesian HLM approach should be useful for providing more insight into the groups for which such investigations are warranted.

References

Cohen. J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Drasgow, F., Levine, M.V., Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38,* 67 – 86.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Levine, M. V. & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269-290.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325-337.

Skorupski, W. P. & Egan, K. (2011, April). *Detecting cheating through the use of hierarchical growth models*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Skorupski, W. P., & Egan, K. (2010). *A review of methods for detecting cheating* (Internal Report). Monterrey, CA: CTB/McGraw-Hill.

Thiessen, B. (2007). Case study—policies to address educator cheating. Retrieved from: http://homepage.mac.com/bradthiessen/pubs/format.pdf

Wainer, H., Wang, X., Skorupski, W. P., & Bradlow, E. T. (2005). A Bayesian method for evaluating passing scores: The PPoP Curve. *Journal of Educational Measurement, 42*(3), 271-282.

Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement, 21,* 307 – 320.

Wollack, J. A. (2003). Comparison of answer copying indices with real data. *Journal of Educational Measurement, 40,* 189 – 205.