A Hierarchical IRT Model for Identifying Group-Level Aberrant Growth to Detect Cheating

Jennifer A. Brussow 1, William P. Skorupski 2, & W. Jake Thompson 3
1 Ascend Learning, LLC & The University of Kansas; 2 The University of Kansas; 3 The Achievement and Assessment Institution & The University of Kansas

ABSTRACT

As cheating on high-stakes tests continues to threaten the validity of score interpretations, approaches for detecting cheating proliferate. Most research focuses on individual scores, but recent events show group-level cheating is also occurring. The present Bayesian IRT simulation study extends the Bayesian Hierarchical Linear Model (BHLM) for detecting group-level aberrance. This preliminary study shows that the model reliably recovers individual ability as well as group-level increases. This model provides a valuable way for testing programs to analyze and detect potential cheating behaviors at the instructor, proctor, or administrator levels.

BACKGROUND

As cheating on high-stakes tests continues to be an issue for standardized testing, approaches for detecting cheating proliferate. However, the majority of research focuses on detecting cheating at the individual level. As recent events have shown (e.g., the Atlanta cheating scandal), cheating at the group level is also a threat to the validity of decisions made from scores on high-stakes standardized tests. Thiessen (2007) estimates that 25% of educators cheat on standardized tests in some way, and Jacob and Levitt (2004) found that 4% to 5% of classrooms in their study of a Chicago area school had incidences of classroom-level cheating. Such group-level cheating occurs when an instructor coaches students on the correct answers to one or more questions during the testing period, changes student answers, or provides an answer key to the group.

The present study adapts the Bayesian Hierarchical Linear Model (BHLM; Skorupski & Egan, 2013, 2014; Skorupski, Fitzpatrick, and Egan, 2016) to detect group-level aberrance. The BHLM models the change in individual scores nested within groups (schools or classrooms) over time. After group- and time-level effects are accounted for, additional aberrant growth is evidence that group-level cheating may have occurred. This simple, generalizable model focuses on each individual's pair of test scores. The current research adapts this model to an IRT framework. Since many testing companies use a latent trait model to estimate examinee ability, this method is more compatible with operational testing programs' current approach to scaling.

OBJECTIVES

- Adapt the Bayesian Hierarchical Linear Model for detecting group-level aberrance (BHLM; Skorupski & Egan, 2013, 2014; Skorupski, Fitzpatrick, and Egan, 2016) to an IRT framework.
- Evaluate convergence and parameter recovery to ascertain the model's viability.
- Use EAP estimates to evaluate classification accuracy of the model across the conditions examined.

METHODS

Data were simulated to emulate two years of standardized test scores for students nested within classrooms. Examinees were simulated within 300 total classrooms with group sizes ~U(5, 35) and with a mean increase in ability of 0.5 standard deviations. These conditions were chosen to mirror typical class sizes and growth rates observed in the American educational system and also to facilitate comparisons with the BHLM simulation in Skorupski, Fitzpatrick, and Egan (2016).

Variables to be manipulated included the size of the cheating effect (τ_g, either 0.5 or 1.0) and the percentage of groups simulated to be aberrant (1% or 5% of groups).

Year 2 scores were estimated using the following equations:

$$P(X_{ij} = 1|\theta_{2_i}, a_j, b_j) = \frac{e^{a_j(\theta_{2_i}-b_j)}}{1+e^{a_j(\theta_{2_i}-b_j)}} \qquad (1)$$

$$\theta_{2_i} = \rho\theta_{1_i} + \gamma + \tau_g + \varepsilon_i \qquad (2)$$

Where ρ is the correlation between the examinee's scores on the first and second years, γ is the mean increase in ability between years across all examinees (simulated to be normally distributed with a mean of 0.5 and a SD of 0.1), and τ_g is the mean difference in group scores between years (simulated to be normally distributed around 0 for non-cheating classrooms and around either 0.5 or 1.0 for cheating classrooms with a SD of 0.1).

As these equations show, group-level information is only included for the second year of assessment, which means that students do not need to remain in the same groups for both years. Individual scores for the first year only serve as the baseline for performance, so researchers do not have to collect information about past years' placement.

Student theta values were simulated to be ~N(0, 1) at time 1; *a*-parameters were drawn from a distribution ~U(0.5, 3.5); *b*-parameters were simulated to be ~N(0, 0.7) at time 1 and ~N(0.5, 0.7) at time 2. ρ was set to 0.7.

The model was estimated using fully Bayesian estimation via the *rstan* package in R 3.4.1. Parameters for θ_1, *b*, and ρ were estimated using a ~N(0, 1) prior; parameters for *a* were estimated using a lognormal (0, 1) prior; parameters for τ_g were estimated using a ~N(0, 3) prior; and parameters for θ_2 were estimated using a ~N(μ, 1) prior, where μ is the linear combination of parameters expressed in Equation 2. Results were evaluated for convergence using $\hat{R}$.

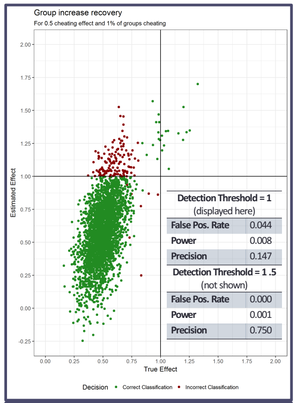
For this preliminary study, 10 replications per condition were conducted.

RESULTS

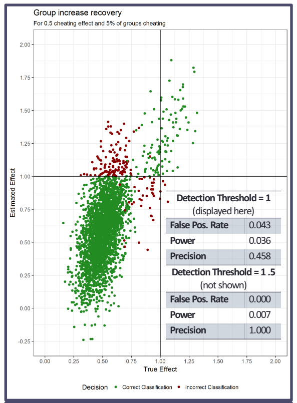
- The proposed method is able to accurately estimate student-level thetas simultaneously with group-level increases ($\hat{R}$ used to investigate convergence).
- The model performed better in terms of identifying true positives when the true aberrant group increase was 1 than when it was 0.5.
- The model over-identified potentially cheating classrooms, suggesting a decision threshold greater than 1.0 should be used.
  - This effect was greater when only 1% of groups were simulated to be cheating, since more of the non-cheating classrooms naturally fell above the decision threshold.
  - Using a decision threshold of 1.5 dramatically reduced the number of false positives, though it also increased the number of false negatives.
- The proportion of groups that cheated did not appear to affect classification accuracy.

RESULTS

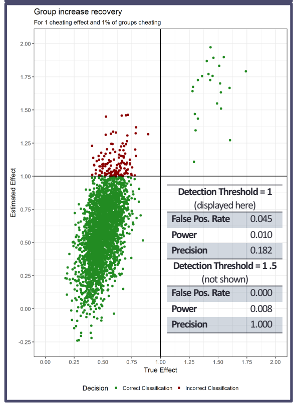
These plots show classification accuracy across the four conditions. A decision threshold of a group increase of 1 was compared to whether the group was simulated to be cheating. Classification accuracy with a decision threshold of 1.5 was also calculated, and those plots are also available upon request.


Group increase recovery
For 0.5 cheating effect and 1% of groups cheating

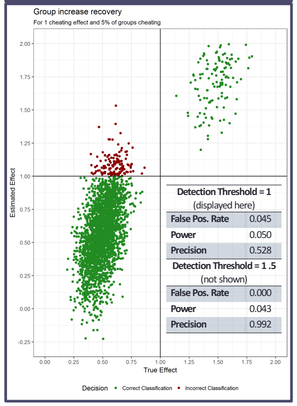
Detection Threshold = 1 (displayed here)	
False Pos. Rate	0.044
Power	0.008
Precision	0.147
Detection Threshold = 1 .5 (not shown)	
False Pos. Rate	0.000
Power	0.001
Precision	0.750


Group increase recovery
For 0.5 cheating effect and 5% of groups cheating

Detection Threshold = 1 (displayed here)	
False Pos. Rate	0.043
Power	0.036
Precision	0.458
Detection Threshold = 1 .5 (not shown)	
False Pos. Rate	0.000
Power	0.007
Precision	1.000


Group increase recovery
For 1 cheating effect and 1% of groups cheating

Detection Threshold = 1 (displayed here)	
False Pos. Rate	0.045
Power	0.010
Precision	0.182
Detection Threshold = 1 .5 (not shown)	
False Pos. Rate	0.000
Power	0.008
Precision	1.000


Group increase recovery
For 1 cheating effect and 5% of groups cheating

Detection Threshold = 1 (displayed here)	
False Pos. Rate	0.045
Power	0.050
Precision	0.528
Detection Threshold = 1 .5 (not shown)	
False Pos. Rate	0.000
Power	0.043
Precision	0.992

CONCLUSIONS

- This model provides a valuable way for testing programs to analyze potential cheating behaviors at the group level.
- The growth aberrance statistic provides a straightforward means of conceptualizing group-level effects and detecting aberrant growth that may indicate cheating.
- Decision thresholds should be set high to minimize false positives, increase power, and improve precision.
- Future research will examine the posterior probability of cheating: the proportion of posterior draws for growth aberrance above a given threshold.