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 Institute & 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 (BHL) 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.

METHODS

Data were simulated to emulate two years of standardized test scores for students nested within classrooms. Examinees were simulated within 100 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 study in Skorupski, Fitzpatrick, and Egan (2016).

Variables to be manipulated included the size of the cheating effect ($r_2$, 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:

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

(1)

$$\theta_{ij} = \rho(\theta_1 + \gamma_1 + \tau_2 + e_i)$$

(2)

Where $\rho$ is the correlation between the examinee’s scores on the first and second years, $\gamma$ 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 $\tau_2$ 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 $\sim U(0.5, 3.5);$ $b$-parameters were simulated to be $\sim N(0, 0.7)$ at time 1 and $\sim N(0.5, 0.7)$ at time 2; $\rho$ was set to 0.7.

The model was estimated using fully Bayesian estimation via the rstan package in R 3.4.1. Parameters for $\theta_1$, $\theta_2$, and $\rho$ were estimated using a $N(0, 1)$ prior; parameters for $e_i$ were estimated using a lognormal (0, 1) prior; parameters for $\tau_2$ were estimated using a $N(0, 3)$ prior; and parameters for $\gamma$ were estimated using a $N(0, 1)$ prior, where $\gamma$ is the linear combination of parameters expressed in Equation 2. Results were evaluated for convergence using $R$.

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

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.

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 and improve precision.
- Future research will examine the posterior probability of cheating: the proportion of posterior draws for growth aberrance above a given threshold.

- The proposed method is able to accurately estimate student-level thetas simultaneously with group-level increases ($\hat{R}$ used to evaluate 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 below 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, though power improved when more groups were simulated as cheaters.