Susan Gillmor & William Skorupski
University of Kansas

Utility of cognitive achievement assessments for educator evaluation

Educational policy mandates the use of statistical growth models for teacher evaluation despite a paucity of validity research from the measurement community. While cognitive assessments are intended to measure student achievement, this study explores the appropriateness of their use for educator evaluation by systematically comparing the validity of the two most commonly used model-types: 1) value-added model (VAM), and 2) Student Growth Percentiles (SGP)[1] with a Monte Carlo simulation study. Two research questions guided the analysis: 1) how much variance is shared among the true teacher effects and the teacher effectiveness estimates from the two models? and 2) how much error variance surrounds the estimates? To answer these questions we simulated longitudinal student achievement scores with known teacher influence. The simulated student data has four levels: time points nested within students, nested in classrooms, within schools. Each student's score is a function of prior achievement, a grade level effect, a school effect ($\delta \sim N(0, .20)$), a teacher effect, and a normally distributed random error component. The independent variable is the size of the teacher effect, with three levels: small $\tau \sim N(0, .10)$, moderate $\tau \sim N(0, .20)$, and large $\tau \sim N(0, .30)$. These values are chosen based on published research on the size of teacher effects.[2] 1000 replications were run for each of the three conditions. We then applied both the value-added model and the Student Growth Percentiles to all 3000 datasets.

Figures 1-3 below shows the sampling distributions of the correlations between true effects and VAM and SGP estimates under the three conditions.

Figure 1:

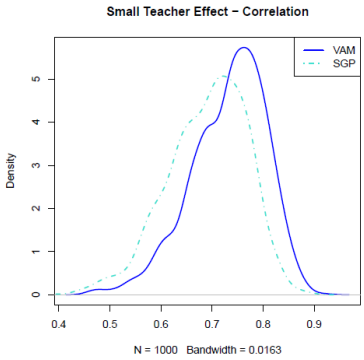


Figure 2:

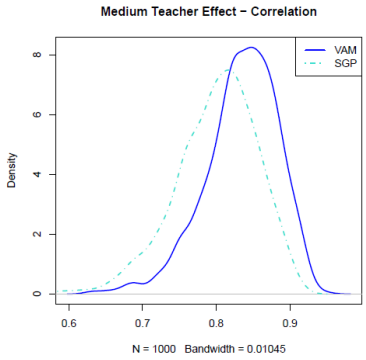
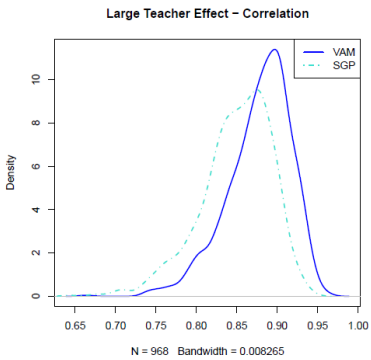


Figure 3:



The VAM generally performs better, accounting for more of the variance in true effects than the SGP. As the influence of the teachers increases, so does the ability of both models to detect the effects. The squared correlation coefficients can be thought of as reliability coefficients because they represent is the proportion of shared variance between true and observed scores. The average reliability coefficients produced from the conditions in this study ranged from .477 to .770. Even when teacher effects are large, the reliability of both of the tested growth models at estimating those effects is unacceptably low.

Figures 4, 5, and 6 below show the sampling distributions of the mean squared errors when the true teacher effects are regressed on the estimates from each model for the three conditions. Again, the Value-added model exhibits more desirable results, estimating the true effects with less random error.

Figure 4:

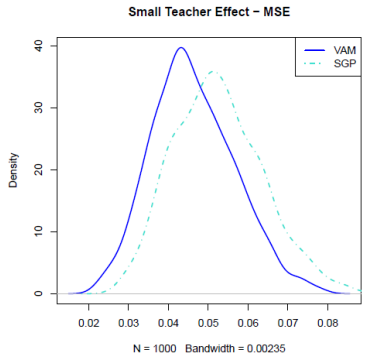


Figure 5:

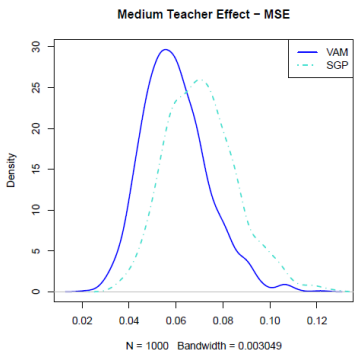
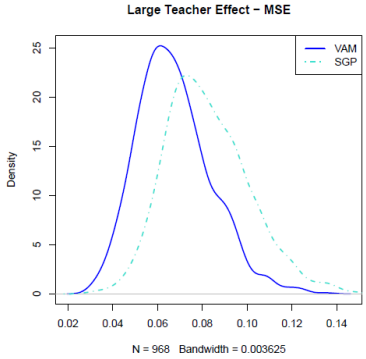


Figure 6:



Please contact scgillmor@ku.edu for more information, thank you.

[1]Betebenner, D. W. (2007). Estimation of student growth percentiles for the Colorado student assessment program. *National Center for the Improvement of Educational Assessment. Available online: http://www.cde.state.co.us/cdedocs/Research/PDF/technicalsgppaper_betebenner.pdf*.

[2]Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237-257.