

Methodological Issues with the Education Value-Added Assessment System (EVAAS) Value-Added Model (VAM): Validity, Reliability, and Bias

Tray Geiger, Ph.D. Student

Educational Policy and Evaluation, Mary Lou Fulton Teachers College, Arizona State University

Background

- Recently in the United States, many school teachers have been fired because they were determined to be “ineffective”; several fired teachers have filed lawsuits challenging the fairness of the terminations, which were based on the use of VAMs, specifically the EVAAS.
- Prior research has found VAMs, in general, to have methodological issues surrounding (1) validity, (2) reliability, and (3) bias.

Study Purpose

- Researchers wanted to empirically test the claim that, per the owners (SAS Institute, Inc.), the EVAAS is a “precise, reliable, and unbiased” way to measure teacher effectiveness and does a better job than other VAMs on the market.

Research Questions

- What are the EVAAS model’s levels of (1) validity, (2) reliability, and (3) bias?

Data

- Main dataset included over 11,000 teachers in Texas; data included teachers’ EVAAS scores, Instructional Practice (IP) (i.e., observational) scores, teacher-level variables (e.g., years of experience, subject taught, grade taught, etc.), and school-level variables (e.g., percentage of minority students, percentage of special education students, etc.).
- RQ 1 sample: 1,788 teachers with EVAAS scores and IP scores across 2012-13 (Year 1), 2013-14 (Year 2), and 2014-15 (Year 3).
- RQ 2 and RQ 3 sample: 2,036 teachers with EVAAS scores across 2012-13 (Year 1), 2013-14 (Year 2), and 2014-15 (Year 3).

Methods

- RQ 1 (Validity): Calculated correlation coefficients (Pearson’s r) between EVAAS and IP scores.
- RQ 2 (Reliability): Calculated correlation coefficients (Pearson’s r), analyzed amount of variation between EVAAS scores over time, and compared distributions of EVAAS scores per year using chi square analyses.
- RQ 3 (Bias): Utilized t tests and one-way fixed effects ANOVA to determine if EVAAS scores significantly differed by teacher- or school-level variables.

Table 1. Percentages of Variation in Teachers’ EVAAS and IP Scores Over Time

	EVAAS Scores	IP Scores
Year 1 to Year 2	62.1%	31.8%
Year 2 to Year 3	63.4%	26.4%

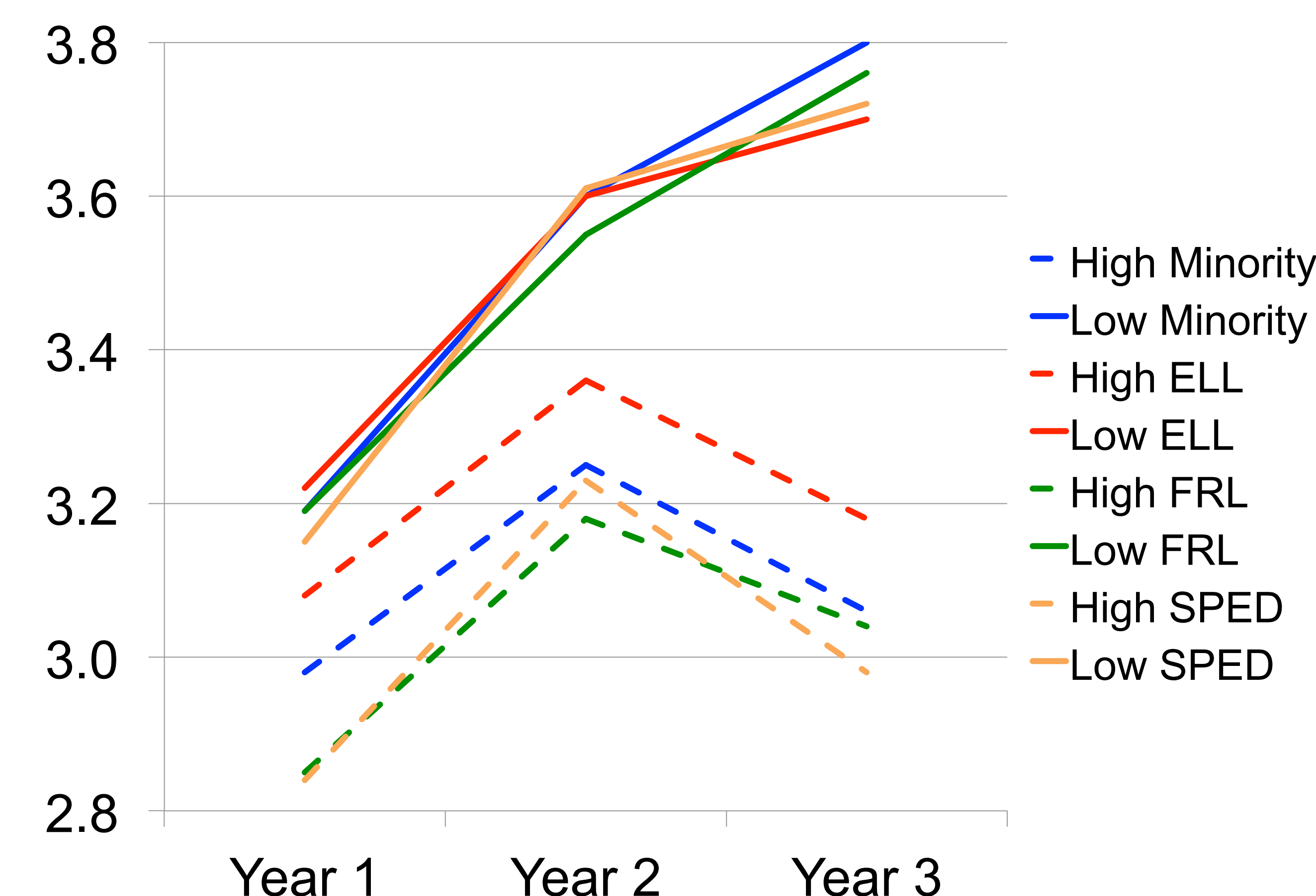


Figure 1. Teachers’ EVAAS Scores, per School Type

Results

Validity

- Correlations between EVAAS and IP scores across all years were weak ($.28 < r < .34$).
- Pattern held when disaggregating by teacher- and school-level variables.

Reliability

- Correlations between EVAAS scores across years were moderate (Year 1 and Year 2: $r = .52$; Year 2 and Year 3: $r = .49$).
- Over 65% of teachers received different EVAAS scores from year to year; there was significantly more variation in EVAAS scores than IP scores (see Table 1).
- Distributions of EVAAS scores across years significantly differed from each other.

Bias

- Teachers with the least amount of experience (i.e., two or fewer years) had significantly low EVAAS scores, while mathematics teachers tended to have higher EVAAS scores.
- Teachers in schools with the relatively lowest populations of English Language Learner (ELL) students, minority students, special education (SPED) students, and free and reduced lunch (FRL) students had significantly lower EVAAS scores than teachers in schools with the relatively highest populations, respectively (see Figure 1).

Conclusion

- Taken together, findings indicate the EVAAS model does not provide “precise, reliable, or unbiased” estimates of teacher effectiveness, per SAS Institute, Inc.’s claim; rather, the EVAAS is comparable to other VAMs.