Consortium for Educational Research and Evaluation– North Carolina

Comparing Value Added Models for Estimating Teacher Effectiveness

Technical Briefing

Roderick A. Rose
Gary T. Henry
Douglas L. Lauen
Carolina Institute for Public Policy

February 2012

Table of Contents

COMPARING VALUE ADDED MODELS FOR ESTIMATING TEACHER EFFECTIVENESS: TECHNICAL BRIEFING

1.0 Objective, Recommendations, and Summary

In the North Carolina Race to the Top proposal, the North Carolina Department of Public Instruction (NCDPI) committed to incorporate teacher effectiveness estimates into the existing teacher evaluation process by adding a criterion for each teacher's effectiveness in raising student test scores. The first step in adding a teacher effectiveness measure is to estimate the effectiveness of individual teachers who taught tested grades and subjects.

The objective of this technical briefing report is to: (1) identify commonly used *value added models* (VAM) for estimating the effectiveness of individual teachers; (2) identify criteria for judging the accuracy (including validity, reliability and consistency in classifying high and low performing teachers) of the VAMs for estimating teacher effectiveness; (3) present the assessment of alternative VAMs for estimating individual teacher effectiveness using both simulated and actual North Carolina data; (4) provide recommendations to NCDPI for them to consider in developing the request for applications (RFA) to estimate the effectiveness of individual teachers and evaluating potential contractors responsiveness to the RFA.

We identified eight primary VAMs (Section 2 and Appendix A) and nine criteria (Section 3) for this evaluation (see Appendix B for a description of the methods). We used both simulated data and actual data from North Carolina from 2005-06 through 2009-10, spanning 3rd through 8th grades. Simulating data allowed us to generate data for which we know each teacher's "true" effect in order to see how closely the alternative VAMs estimates were to the "true" effect. The actual NC data allowed us to assess the reliability, consistency, and percentage of NC teachers that can be expected to be identified as highly effective or ineffective based on the best available data for those assessments.

Based on our findings we recommend that DPI should request contractors to propose one or more of the following value-added models for estimating teachers' effectiveness:

- Three-level hierarchical linear model (HLM3): a 3-level rich covariate multilevel model (4th grade – 8th grades)
- Univariate response model (URM): an EVAAS model developed by the SAS Institute (5th grade – 8th grades)
- Student fixed effects model (SFE): an ordinary least squares model on a 3 year panel with student fixed effects (5th grade – 8th grades)

It is important to note that the HLM3 model allows for teachers from an additional grade level (4th grade) to be included in the teacher effectiveness estimates, which neither of the other higher performing models allow, even though the other higher performing models perform better on some criteria.

In sections 2 and 3, respectively, we describe the VAM models and criteria used to make these recommendations. In section 4, we provide a summary tabulation of the evidence supporting the recommendations. In the Appendices, we provide tables summarizing the key features of each VAM, explanations supporting the summary tabulation and recommendations, followed by tables developed from analysis of observed and simulated data.

2.0 The VAM Models Evaluated in this Study

After reviewing the research literature on alternative VAMs for estimating individual teacher's effectiveness and identifying the VAMs that have been used in other states or school districts, we identified eight primary VAMs for this analysis:

1. Two level hierarchical linear model (HLM2): a *random effects* model that accounts for the clustering of students with teachers in each year and grade level and can incorporate student and teacher/classroom characteristics to adjust effectiveness estimates. The teacher effect is captured by the teacher level residual or random effect, net of measured student characteristics, including background characteristics and the previous year's end-of-grade performance, and measured classroom characteristics that have been included in the model.
2. Three level hierarchical linear model (HLM3): a *random effects* model that accounts for the clustering of students with teachers in each year and grade level and of these teachers within schools and can incorporate student, teacher/classroom, and school characteristics in the models to adjust effectiveness estimates. The teacher effect is captured by the teacher level residual or random effect, net of measured student variables, including background characteristics and the previous year's end-of-grade performance, measured classroom characteristics, and measured school characteristics that have been included in the model.
3. Univariate response model (URM): an Education Value Added Assessment System (EVAAS) *random effects* model that accounts for the clustering of students with teachers and incorporates two previous years' end-of-grade performance but not student, classroom, or school characteristics. The teacher effect is captured by the teacher level residual or random effect, net of the student's previous end-of-grade test performances.
4. Multivariate response model (MRM): the original EVAAS model is a "multiple membership, multiple classification" *random effects* model that not only accounts for students clustering with teachers, but that each student and his or her peers cluster with different teachers in different years and may have multiple teachers in a given year for the same subject. The MRM accounts for the effects of all other past and future teachers on students. The teachers' effects for the teachers in any grade level are random effects that are adjusted for the effects of all other past and future teachers that a student has.
5. Student fixed effects model (SFE): a longitudinal within student (*fixed effects*) model that controls for all between-student variation by using each student as his or her own control over the duration of the panel. Only those measured student characteristics that change during the panel, including end-of-grade performance, can be used to further adjust the teachers' effect estimates. The teachers' effects are the means of the residuals of the

regression, net of all varying student effects that are included in the model, aggregated up to the teacher level.

6. Teacher fixed effects (TFE): a longitudinal within teacher (*fixed effects*) model that captures between-teacher differences by incorporating an indicator variable for each teacher in the model, which is used as the teacher's effectiveness estimate. It is very similar to the HLM2 except that the teacher effects are recovered directly from the coefficient on the indicator variable associated with that teacher rather than random effects.
7. Student fixed effects instrumental variable model (SFEIV): an instrumental variable model uses a variable that is putatively unrelated to current student performance to adjust for unobserved effects on prior student test scores that may confound measurement of a teacher's effect. The *fixed effects* imply that each student is used as his or her own control. As with the SFE, only those characteristics that change can be used to further adjust the teachers' effect estimates. The teacher effect is the teacher level aggregate of the student residuals net of all varying student effects.
8. Teacher fixed effects instrumental variable model (TFEIV): same as the SFEIV, except that the *fixed effects* are directly estimated by teacher indicator variables in the model that are entered into the model. The teacher effect is the coefficient on the indicator variable associated with each teacher.

3.0 Evaluation Criteria

For this study, we developed four types of criteria: design criteria, which assess each VAMs design and its limitations; simulated data criteria, which assess each VAM using data that closely resembles NC data but for which the "true" teacher effect is known; simulated and actual data criteria, which assess each VAM using the simulated and actual data; and actual data criteria, which assess VAM performance using actual NC data only. The degree to which each VAM meet these criteria was assessed by examining the models' designs and testing its computational feasibility by running them using the University of North Carolina at Chapel Hill statistical computing platform.

3.1 Design Criteria

1. *Does the model limit the estimate of teachers' effectiveness to the year in which they taught the students or does it explicitly account for students' test score growth in subsequent years?*

   Teachers contribute to students' learning not only in the year in which the students are in their classes but may also contribute to their learning in subsequent years. An effect estimate for a teacher may consider only the year in which the students are in their classes or their cumulative effect on students' test scores. A model that explicitly includes teachers' effects in subsequent years was included.

2. *Can the VAM be estimated simultaneously for all teachers in the state who are teaching the same subject or grade – thus holding all teachers in the same grade and subject to the same*

statewide standard – or can the estimates only be computed one district at a time – thus establishing 115 district standards for North Carolina?

To make comparisons and judgments between teachers teaching a given subject and grade level consistent across the state, the model should accommodate the statewide population of teachers in that subject and grade level. The computing resources required for the VAM must be low enough to accommodate several hundred thousand students over multiple years.

3.2 Criteria Evaluated with Simulated Data

3. *Do the VAMs accurately estimate "true" teacher effects?*

   The central goal of each VAM is to estimate from student test scores each teacher's effect on learning. The relative performance of each VAM is assessed two ways. First, we show how well each VAM ranks teachers consistent with the teachers' "true" effects. Second, we demonstrate how well each VAM identifies teachers whose "true" effects place them in top or bottom 5% of all teachers.

4. *How accurately do the teacher effectiveness estimates (TEEs) from VAMs categorize a teacher as ineffective?*

   The negative consequences of incorrectly classifying a teacher who is not ineffective as ineffective are very serious, as teachers found ineffective will be subjected to a number of mandated actions. Given the stakes associated with committing this error, we focus this criterion on the incorrect classification of teachers who are actually *not ineffective* as ineffective. We compute the percentage of teachers who are in the middle of the distribution or higher – that is, teachers who are of average or higher effectiveness – who the VAM *incorrectly* identifies as ineffective.

5. *How sensitive are the VAM TEEs to the choice of threshold for establishing ineffectiveness?*

   Again, because of the potential for adverse consequences for identifying a teacher as ineffective, we investigate whether and how much the percentage of teachers incorrectly found to be ineffective under each VAM changes when different cutoff points are used for identifying ineffective teachers.

6. *Does the VAM produce estimates that are reliable when assumptions of the model are violated?*

   Each model provides for some control for student background factors; some of the models also control for school level variation. None of the VAMs tested using simulated data explicitly controls for peer effects. Effects that are not controlled or adjusted for have the potential to lead to incorrectly estimated teacher effects, as the effects of student, school and peer effects may be incorrectly attributed to the teacher. We examine the effect of these three characteristics – student background, school characteristics, and peer effects –on the relative performance of each VAM using the same standards as we did for criterion 3 (consistent rankings and percent agreement in the top or bottom 5%).

3.3 Criterion Evaluated with both NC Actual and Simulated Data

7. *How similar are Teacher Effectiveness Estimates from each VAM to the Teacher Effectiveness Estimates from each other VAM?*

   To examine the consistency between the VAMs, we use the standards used for criteria 3 and 6—consistent ranking and percent agreement in the top or bottom 5%—to compare each VAM's Teacher Effectiveness Estimates to those produced by the other VAMs.

3.4 Criteria Evaluated with NC Actual Data Only

8. *Does the VAM yield a reasonable number of high and low performing teachers?*

   We use the standard employed by some other jurisdictions – two standard deviations or more above or below the mean of teacher effectiveness to identify high and low performing teachers, respectively, and show the percentages identified for each VAM.

9. *For each VAM, are TEEs for individual teachers consistent or reliable from one year to the next?*

   Prior research indicates that teachers' effectiveness can change from year-to-year, especially during their first few years in the classroom. However, if a VAM produces quite different effectiveness estimates for individual teachers from one year to the next, this might suggest confounding effects, including teacher and student assignments to classes in any given year, are present and not sufficiently controlled by the VAM. We investigate the year-to-year stability of the Teacher Effectiveness Estimates by comparing teachers' placement in the quintile groupings (five performance categories of equal size) and their placement in the top or bottom 5% performance categories in one year to their placement in the following year.

4.0 Summary of Findings

	HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
1. Cumulative effects of teachers (+ indicates yes)?	-	-	-	+	-	-	-	-
2. Whole state (S) or one district (D) at a time?	S	S	S	D	S	S	S	S
3.1. Accuracy of the VAM in ranking teachers according to their "true" effect (+ indicates high).	-	+	+	*	+	-	-	-
3.2. Accurate identification of top 5% of teachers (+ indicates yes).	-	+	+	N/A	+	-	-	-
3.3. Accurate identification of bottom 5% of teachers (+ indicates yes).	-	+	+	N/A	+	-	-	-
4. Percent falsely identified as ineffective (+ indicates low).	-	+	+	*	+	-	-	-
5. Sensitivity of false identification to threshold (+ indicates low).	-	+	+	*	+	-	-	-
6.1. Sensitivity to student background (+ indicates low).	-	+	+	*	+	-	-	-
6.2. Sensitivity to school characteristics (+ indicates low).	-	+	-	*	+	-	-	+
6.3. Sensitivity to peer effects (+ indicates low).	-	-	-	*	-	-	-	-
7.1. Similarity of VAMs to each other (+ indicates similar).	+	+	+	N/A	-	+	-	-
7.2. Agreement on classifying teachers in the top 5% (+ indicates high agreement).	+	-	+	N/A	+	-	-	-

	HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
7.3. Agreement on classifying teachers in the bottom 5% (+ indicates high agreement).	+	-	+	N/A	-	+	-	-
8.1. Number of teachers 2 SD above mean (H = high, L = low, M = mix of high and low).	H	H	H	N/A	M	L	L	M
8.2. Number of teachers 2 SD below mean (H = high, L = low, M = mix of high and low).	H	H	H	N/A	M	L	M	M
9. Reliability of TEEs from year to year (+ indicates high).	+	-	+	N/A	+	+	-	-

Legend: **HLM2**: 2 level hierarchical linear model, students nested in teachers; **HLM3**: 3 level hierarchical linear model, students nested in teachers, nested in schools; **URM**: univariate response EVAAS model; **MRM**: multivariate response EVAAS model; **SFE**: student fixed effects model; **TFE**: teacher fixed effects model; **SFEIV**: student fixed effects instrumental variable model; **TFEIV**: teacher fixed effects instrumental variable model. *Model details follow on next page.* A "*" indicates the MRM was assessed on this criterion using small sample simulations only; N/A indicates the MRM was not tested on the criterion.

5.0 Conclusions and Study Limitations

Each of the three top performing models has numerous strengths and a few weaknesses.

The HLM3 model performs very well on numerous criteria, including overall accuracy, correct identification of top or bottom 5 percent, the infrequent identification of effective teachers as ineffective, and the identification of similar percentages of high and low performing teachers in reading and mathematics. The HLM3 performs less well than some of the other top performers in terms of the consistency with other models on the identification of top or bottom 5 percent of the teachers using actual data and year-to-year reliability. An advantage of the HLM3 is that it requires a single year of prior test scores which allows 4th grade teachers to be included in the VAM teacher effectiveness estimates.

The URM performs very well on all criteria, including overall accuracy, correct identification of top or bottom 5 percent, the infrequent identification of effective teachers as ineffective, and year-to-year reliability, except for the ability to correctly adjust for school effects on student test scores. The URM is the EVAAS model that does allow for statewide estimation of teachers' effectiveness but it excludes 4th grade teachers from the estimates.

The SFE is a top performer on almost all criteria, including overall accuracy, correct identification of top or bottom 5 percent, the infrequent identification of effective teachers as ineffective, and year-to-year reliability, except for agreement with the other top performing VAMs using both actual and simulated data. The SFE excludes 4th grade teachers from the estimates.

While each has some weakness, they are better than the other tested VAMs when assessed across all the criteria and standards.

Based on our findings we recommend that NCDPI should consider requesting potential contractors to propose one or more of the following value-added models for estimating teachers' effectiveness and providing them for use in teachers' evaluations:

- Three-level hierarchical linear model (HLM3): a 3-level rich covariate multilevel model (4th grade – 8th grade)
- Univariate response model (URM): an EVAAS model developed by the SAS Institute (5th grade – 8th grade)
- Student fixed effects model (SFE): an ordinary least squares model on a 3 year panel with student fixed effects (5th grade – 8th grade)

While there are efforts to apply value-added methods to untested grades and subjects, the reliability of the test measures (or lack of reliability compared to standardized end-of-grade exams) is likely to greatly affect the accuracy and reliability of those methods. This study did not address the performance of those methods nor did it assess the performance of the VAMs evaluated on high school tests, either end-of-course exams or high school graduation tests. Nor did we take into account feasibility for any particular contractor to accurately implement these models or manage the longitudinal datasets used for the analyses.

Appendix A: Features of Value-Added Models

	Method	Model	Teacher Effect	DV: Status, Change Score or Mean-Centered	Data Structure / LHS # Periods	# Pretests	Earliest Grade Possible	Source of Teacher Effect
1	HLM2	$Y_{it} = X_{it}\beta_t + \beta Y_{i,w=1} + u_t + e_{it}$	shrinkage estimate from variance component u_t	Status	Cross-sectional / 1	1	4	Between-student differences
2	HLM3	$Y_{its} = X_{its}\beta_{ts} + X_s\beta_s + \beta Y_{i,w=1} + u_s + u_{ts} + e_{its}$	shrinkage estimate from variance component u_{ts}	Status	Cross-sectional / 1	1	4	Between-student differences
3	URM	$y_{it} = \beta_0 + \beta_1 C_i + u_t + e_{it}$	shrinkage estimate from variance component u_{ts}	Status	Cross-sectional / 1	2	5	Between-student differences
4	MRM	$y_{ijkl} = \mu_{jkl} + \left(\sum_{k^* \le k} \sum_{t=1}^{Tijk^*l^*} p_{ijk^*l^*t} \, x \, \tau_{ijk^*l^*t}\right) + e_{ijkl}$	shrinkage estimate from variance component $\tau_{ijk^*l^*t}$	Status	Panel / 3	0	5	Between-student differences adjusted for previous teachers' contributions to learning
5	SFE	$(Y_i - \bar{Y}_i) = (\mu_i - \bar{\mu}_i) + (\alpha_i - \bar{\alpha}_i) + (e_i - \bar{e}_i)$	Mean of the compound error across students within each teacher	Mean-centered	Panel / 3	0	5	Change in student outcomes and characteristics
6	TFE	$Y_{it} = X_{it}\beta_t + \alpha_t + e_{it}$	teacher dummy variables α_t	Status	Cross-sectional / 1	1	4	Between-student differences
7	SFE-IV	$\Delta Y_{i,w=1} = \Delta X_i\beta + Y_{i,w=2} + r_i$ $\Delta Y_i = \Delta X_i\beta + \Delta\hat{Y}_{i,w=1} + e_i$	Mean of e_i within each teacher	Change score	Cross-sectional / 1	2	5	Change in student outcomes and characteristics
8	TFE-IV	$\Delta Y_{i,w=1} = \Delta X_i\beta + Y_{i,w=2} + \alpha_t + r_i$ $\Delta Y_i = \Delta X_i\beta + \Delta\ddot{Y}_{i,w=1} + \alpha_t + e_i$	teacher dummy variables α_t	Change score	Cross-sectional / 1	2	5	Change in student outcomes and characteristics

Appendix B: Detailed Explanation of Findings for Each Question

1. Cumulative teachers' effects?

Does the model limit estimates of teachers' effectiveness to the year in which they taught the students or explicitly account for students' test score growth in subsequent years?

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
-	-	-	+	-	-	-	-

- This concept is referred to in the VAM literature as teacher effect *layering*.
- If the model *explicitly and completely accounts for* the accumulation of teacher effects, and estimates the teacher effect for the current tested subject from only that portion of the student's learning that is unique to the current teacher, the cell is labeled +. If it is not explicitly or completely accounted for, the cell is marked with a -.
- Findings:
  - The MRM is the only model that explicitly and completely accounts for cumulative teacher effects. It does this by incorporating all information about all current and previous teachers a student has had over all grade levels and years for which the student has been tested. Accordingly, it is a very dense and computing-intensive model to estimate.
  - The other models do not explicitly account for cumulative teacher effects and cannot completely adjust for this accumulation in the estimation of the teacher effects. Instead, they use either student pre-tests and covariates, or students or teachers as their own controls, both of which can at best result in a partial teacher effect adjustment for the accumulation of learning. The extent to which these controls adjust for accumulation cannot be known.

2. Individual teachers compared other teachers across the whole state or within one district at a time?

Can the VAM be estimated simultaneously for all teachers in the state teaching the same subject or grade, or can it only be estimated one district at a time?

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
S	S	S	D	S	S	S	S

- If the VAM can be estimated on the statewide population of teachers, then the cell is labeled S. If the VAM can only be estimated on the population of teachers in a single district, then the cell is labeled D.

- More complex models with a greater number of records and variables to account for, and more complex effects estimation will demand more intensive computing resources and make whole-state estimation less likely.
- Findings:
  - For most of these models, modest computing resources are needed to estimate effects for the statewide population of teachers.
  - For the MRM, a statewide estimate is not possible, though within-district estimates should be possible for most districts.

3. Accuracy: Are the TEEs more or less precise than those from other models?

Three criteria are used to answer this question:

3.1 Accuracy of the VAM in ranking teachers according to their "true" effect.

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
-	+	+	*	+	-	-	-

- This criterion is assessed using a coefficient representing the correlation between each teacher's rank on the VAM effect and the rank on the "true" effect.
- VAMs with high correlations with the "true" effect (approximately .90 and up) are labeled +.
- Findings:
  - HLM3, URM and SFE, had the highest 3 correlations with the "true" effect (ranging from .892 to .934).
  - The HLM2, SFEIV and TFEIV were lower, ranging from .65 to .86.
- The MRM (*) was not tested on the full sample simulation that the other VAMs were subjected to, but it was tested in an earlier phase of testing using small-sample simulations and it was not found to perform as well as the HLM3, URM or SFE.

3.2 Correct identification of the top 5% of teachers.

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
-	+	+	N/A	+	-	-	-

- This criterion is assessed by identifying the percentage of teachers whose "true" effect and VAM effect *agree* in categorizing the teacher in the top 5% of teachers.
- Models with the highest levels of agreement (96% or higher) are labeled with a +.

- Findings:
  - The HLM3, URM, and SFE always outperformed the other models with agreement above 96%. The SFE was best at 97%.
  - All of the models except the TFE performed similarly well on this criterion, with agreement above 95%. The TFE was at 93-94%.
- The MRM (*) was not assessed on this criterion.

3.3 Correct identification of the bottom 5% of teachers.

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
-	+	+	N/A	+	-	-	-

- This criterion is assessed by identifying the percentage of teachers whose "true" effect and VAM effect *agree* in categorizing the teacher in the bottom 5% of teachers.
- Models with high level of agreement (96% or higher) are labeled with a +.
- Findings:
  - The HLM3, URM, and SFE always outperformed the other models with agreement above 96%. The SFE was best at 97%.
  - All of the models except the TFE performed similarly well on this criterion, with agreement above 95%. The TFE was at 93-94%.
- The MRM (*) was not assessed on this criterion.

4. What percentage of teachers is falsely identified as being ineffective based on an ineffectiveness threshold of two standard deviations below the mean?

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
-	+	+	*	+	-	-	-

- This question is answered using the population of teachers *above* two standard deviations below the mean on their "true" effect (thus *not ineffective*). Those who were subsequently found to be *below* two standard deviations from the mean on their VAM estimate were considered to be falsely identified as ineffective.
- Models that demonstrate higher false identification of ineffectiveness are labeled with a -.
- In the first test of this criterion, the threshold for the VAM effect was also two standard deviations below the mean, the same as the true effect threshold.
- Findings:
  - The HLM3, URM and SFE perform similarly well in misclassifying less than one percent (0.8-0.9%) of not-ineffective teachers as ineffective.

  - A small gap (less than ½ of a percent) separates these three from most of the other four models. However, the TFE falsely identified up to an additional 1% of teachers as ineffective.

- In the second version of this criterion, the threshold defining VAM ineffectiveness was one quarter of a standard deviation lower (-2.25 SD) than the threshold defining ineffectiveness on the "true" teacher effects. This provides a .25 SD margin of error for falsely identifying teachers as ineffective when they should not be.
- Findings:
  - The number of teachers misclassified was slightly lower (falling by up to ½ of a percent), and the differences between the top 3 performers (HLM3, URM and SFE) and the other four models widened.
- The MRM is marked with a "*" because large-sample simulations were not performed on the MRM. Small-sample simulations suggested that the MRM performed as well as the HLM3, URM, or SFE in false identification, but the findings were strongly affected by the sample size.

5. How sensitive is the false identification of ineffectiveness for each VAM to the choice of a threshold for establishing "true" ineffectiveness?

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
-	+	+	*	+	-	-	-

- To determine the sensitivity of false identification of ineffectiveness to the threshold used to establish ineffectiveness (set at 2 standard deviations below the mean for criterion 4), the threshold for ineffectiveness was varied in a range from 2.5 standard deviations below the mean teacher effect to 1.5 standard deviations below the mean teacher effect.
- For assessing sensitivity, we required a margin of error of .25; thus the threshold on the VAM is .25 SD lower than the threshold on the "true" effect.
- VAMs that are less sensitive to the threshold are marked with a +.
- Findings:
  - When the threshold was closer to the mean, the teacher was more likely to be misclassified relative to when the threshold was farther away from the mean. This tendency occurs for all of the VAMs.
  - However, the SFE, HLM3, and URM perform relatively better, with misclassification rising by about one percentage point over the range of thresholds from -2.5 to -1.5 SDs.
  - In contrast, the HLM2, SFEIV and TFEIV models rose by 1.5 to 2 percentage points. The TFE rose by more than 3 percentage points.

- The MRM is marked with a * because large-sample simulations were not performed on the MRM; small-sample simulations, however, suggested that the MRM did not perform as well as the HLM3, URM, or SFE in accuracy or sensitivity to the threshold.

6. Does the VAM produce estimates that are reliable when assumptions of the model are violated?

Three standards are used to answer this question:

6.1 Sensitivity of the VAM estimate to student background characteristics.

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
-	+	+	*	+	-	-	-

- The sensitivity of the VAMs to student background was assessed by comparing the performance of each VAM on each of criteria 3-5 using both an unadjusted teacher effect and a teacher effect adjusted for its correlation with a student covariate representing pre-enrollment characteristics.
- Models that were less sensitive to student background are marked with a +.
- The covariate was entered into each model if appropriate (this is always the case; not just for this criterion). The covariate could not be included in the URM or MRM, and it would have no effect on the estimation of the SFE, SFEIV or TFEIV models because it is differenced to zero.
- Findings:
  - While the models that incorporate the covariate (HLM2, HLM3 and TFE) are the most sensitive to the adjustment of the teacher effect for its correlation with the student covariate, the differences across all of the models and all criteria were negligible.
- The MRM is marked with a * because large-sample simulations were not tested on the MRM.

6.2 Sensitivity of the VAM estimates to school characteristics.

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
-	+	-	*	+	-	-	+

- The sensitivity of the models to school characteristics was assessed by comparing the performance of each model in ranking according to the "true" simulated ranks, while adjusting the proportion of the student test score that was explained by between school variation. Between school variation was made intentionally heterogeneous by categorizing some schools as "high socioeconomic status" and others as low on this characteristic. The proportion of variability between high socioeconomic schools was

allowed to vary in scenarios ranging from 6% to 30%, and the rankings of the VAMs with the simulated "true" effects were monitored. The coefficient for ranking (see 3.1) was used in this assessment.

- Models that were less sensitive to school characteristics are marked with a +.
- Findings:
  - The best performing models were the SFE, HLM3, and TFEIV, with nearly constant ranking coefficients (between .85 and .90).
  - The MRM is marked with a * because large-sample simulations were not performed on the MRM; small-sample simulations, however, suggested that the MRM performed as well as the TFEIV though not as well as the SFE or HLM3.

6.3 Sensitivity of the VAM estimate to peer effects.

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
-	-	-	*	-	-	-	-

- The sensitivity of the models to peer (or "classroom") effects was assessed by allowing peer effects to explain a portion of the students' test scores, adjusting the teacher effect by half of this portion (the other half from the student portion of the effect), and re-examining the comparisons conducted for questions 3-5.
- The portion of students' test scores attributed to the "true" peer effect was varied from 0% (the models used up to this point) to 8% in 2% increments.
- Because none of the models include a control for classroom or peer effect nesting, the potential for mis-estimation when a "true" peer effect was present was acute and this potential increased with the size of the "true" peer effect.
- While we did not test a peer or classroom-specific covariate in the simulations, analysis of actual data models that include such a variable suggest that including a peer variable in the simulation might reduce the VAMs' sensitivity to the presence of a peer effect; however, the analysis also suggest that this depends upon the extent to which the peer effect and true teacher effect are correlated. This represents an avenue for further work. While this issue is particularly pertinent to the HLM2 and HLM3 models, which correctly specify student clustering with teachers and can incorporate classroom effects, the TFE model could also be affected.
- Models that were less sensitive to peer effects are marked with a +.
- Findings:
  - The results confirmed that, across the tests used for questions 3-5, the performance of the VAMs relative to the "true" effects worsened as the peer effects went up.
  - However, while they all worsened, the extent to which the performance worsened depended on both the VAM and the criterion used.

  - The most sensitive criterion was the rank correlation (from 3.1), which saw the HLM3 drop from .89 to .73; the URM from .90 to .73; and the SFE from .93 to .77.
  - The % agreement tests (from 3.2 and 3.3) were the least sensitive to the increasing peer effect, dropping by at most 3 percentage points as the peer effect rose from 0% to 8% of the student outcome. This effect was observed across all VAMs.
  - On the false identification of ineffective teachers (from question 4), the increasing peer effect caused the performance of the three best models (HLM3, URM and SFE) to become more similar to the next best three (HLM2, SFEIV, TFEIV), largely due to a more substantial decline in the performance of the best three. The overall increase in false identification was at worst approximately ½ of a percent of not ineffective teachers.
  - The relative performance of the VAM estimates depends on the level of the peer effect. As the level of the peer effect rises, the VAMs look more similar in the proportion of teachers identified as ineffective. Despite this contraction of the differences in the effects, the HLM3, URM and SFE models were still the best-performing models.

- The MRM is marked with a * because large-sample simulations were not performed on the MRM; small-sample simulations, however, suggested that the MRM did not perform as well as the HLM3, URM, or SFE in sensitivity to the violations studied.

7. Are the TEEs from each VAM similar to estimates from each other VAM?

Three criteria are used to answer this question:

7.1 Similarity between VAMs on ranking.

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
+ HLM3, URM, TFE	+ HLM2, URM, TFE	+ HLM2, HLM3, TFE	N/A	-	+ HLM2, HLM3, URM	- TFEIV	- SFEIV

- To assess this criterion, both actual North Carolina data and simulated data were used. In both, the teachers were ranked according to their effects as estimated by the VAMs, and a coefficient representing the correlation between the rankings of pairs of VAMs was estimated. This was repeated for every possible pairing of VAMs and was assessed under a number of scenarios regarding the magnitude of the teacher effect.
- The actual North Carolina data included 5th through 8th grade for both math and reading end-of-grade exams.
- VAMs with a tendency to correlate highly with the other VAMs are labeled +; those that tended to have lower correlations are labeled -. The VAMs with which each VAM is correlated is listed below the label.

- Findings:
  - The NC actual data shows a clear pattern distinguishing the HLM2, HLM3, URM and TFE from the other 3 models.
  - The HLM2, HLM3, URM and TFE were more highly correlated with each other, in a range from 0.69 to 0.94.
  - The SFE, SFEIV, and TFEIV had lower correlations in rankings with the other models, typically 10 or more points lower. These models were similarly ranked among themselves. The correlation between the SFEIV and TFEIV stands out, at above 95%.
  - The simulation data and NC actual data yield slightly different findings regarding the SFE and TFE models. The rankings from the SFE were not as highly correlated with those from the URM and HLM3 using the actual NC data as they were when using the simulated data. Alternatively, the simulated TFE model was not highly correlated with any other models.
  - The correlation between the URM and HLM3 was similar across the simulated and actual NC data; however, the HLM2 stands out as being highly correlated with both when using the actual NC data when it was not highly correlated with the URM or HLM3 when using the simulate data.
  - In the analysis of the actual NC data, the rankings from the models for reading were much lower than those for math.
- The MRM was not assessed on this criterion.

7.2 Agreement between VAMs on classifying teachers in the top 5%.

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
+ HLM3, URM, SFE	- HLM2, URM	+ HLM2, HLM3, SFE	N/A	+ HLM2, URM, TFE, SFEIV, TFEIV	- SFE, SFEIV, TFEIV	- SFE, TFE, TFEIV	- SFE, TFE, SFEIV

- This was tested by identifying the percentage of teachers whose effect under each pair of VAMs *agrees* in categorizing the teacher in the top 5% of teachers. Agreement was assessed for every possibly pair of VAMs, and was assessed under a number of scenarios regarding the magnitude of the teacher effect.
- VAMs that tended to have high agreement are labeled +; those that tended to have low agreement are labeled -. The models that each VAM had high agreement with are listed below the label.
- The actual North Carolina data included 5th through 8th grade for both math and reading end-of-grade exams.

- Findings:
  - On the actual NC data, agreement on math tended to be higher than agreement on reading. However, the patterns, such as the top 2 correlations for each VAM, exhibited very stable tendencies across exam. On the other hand, they were not stable across grade level.
  - On the actual NC data, there was a tendency for the HLM2, HLM3 and URM to correlate highly with each other and lower with the others; and for the SFE, TFE, SFEIV and TFEIV to correlate highly with each other and lower with the others. This is less consistent for the SFE, which tended to correlated highly with the HLM2 and URM on math.
  - On the simulated data, the TFE was relatively low compared to the other models; otherwise the models had very similar levels of agreement.
- The MRM was not assessed on this criterion.

7.3 Agreement between VAMs on classifying teachers in the bottom 5%.

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
+ HLM3, URM, TFE	- HLM2, URM	+ HLM2, HLM3, TFE	N/A	- TFE, SFEIV, TFEIV	+ HLM2, URM, SFE, SFEIV, TFEIV	- SFE, TFE, TFEIV	- SFE, TFE, SFEIV

- This was tested by identifying the percentage of teachers whose effect under each pair of VAMs *agrees* in categorizing the teacher in the bottom 5% of teachers. Agreement was assessed for every possibly pair of VAMs, and was assessed under a number of scenarios regarding the magnitude of the teacher effect.
- VAMs that have high agreement are labeled +; those that had low agreement are labeled -.
- Findings:
  - On the actual NC data, agreement on math tended to be higher than agreement on reading. Further, the patterns, such as the top 2 correlations for each VAM, exhibited very stable tendencies across exam and grade level.
  - On the actual NC data, there was a tendency for the HLM2, HLM3 and URM to correlate highly with each other and lower with the others; and for the SFE, TFE, SFEIV and TFEIV to correlate highly with each other and lower with the others. The TFE was also frequently highly correlated with the HLM2 and URM on math.
  - On the simulated data, the TFE was relatively low compared to the other models; otherwise the models had very similar levels of agreement.
- The MRM was not assessed on this criterion.

8. Does the model yield a reasonable number of high and low performing teachers?

Two criteria were used to answer this question:

8.1 Percentage of teachers two standard deviations or higher above the mean effect.

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
H	H	H	N/A	M	L	L	M

- This was assessed by calculating each teacher's standardized score and identifying those teachers with a score greater than or equal to 2 above or below -2.
- VAMs with higher percentages of teachers across both subjects identified as high performing were labeled "H", and those with lower numbers labeled "L". Those with a mix of high and low were labeled "M". Whether any number was low depended on whether math or reading was being examined, and the grade level.
- Findings:
  - Most of the VAMs are in the 1-4% range across grade level and subject.
  - For math, the percentage identified as high performing falls as grade level increases. This tendency is not observed for reading.
  - For math, the HLM2, HLM3, URM, SFE, and TFEIV models identify approximately 3% or fewer teachers as high performing across grade levels. The TFE and SFEIV usually identify 2% or less.
  - For reading, the HLM2, HLM3, and URM models identify approximately 3% or fewer teachers as high performing across grade levels. The SFE, TFE, SFEIV, and TFEIV identify approximately 2% or less.
- The MRM was not assessed on this criterion.

8.2 Percentage of teachers two standard deviations or lower below the mean effect.

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
H	H	H	N/A	M	L	M	M

- This was assessed by calculating each teacher's standardized score and identifying those teachers with a score greater than or equal to 2.
- VAMs with higher percentages of teachers across both subjects identified as low performing were labeled "H", and those with lower numbers labeled "L". Those with a mix of high and low were labeled "M". Whether any number was low depended on whether math or reading was being examined, and the grade level.
- Findings:
  - Most of the VAMs are in the 1-4% range across grade level and subject.

  - For math, the percentage identified as low performing increases as grade level increases. This tendency is not observed for reading.
  - For math, the HLM2, HLM3, URM, SFE, SFEIV and TFEIV models identify approximately 2% or fewer teachers as low performing across grade levels, though this increases in some cases (HLM3, SFE, and SFEIV) to 3% as grade level increases. The TFE identifies 4% or less (usually 3% or less).
  - For reading, the SFE usually identifies 4% as low performing; the HLM2, HLM3, URM, TFE, SFEIV and TFEIV identify 3% or fewer as low performing.
- The MRM was not assessed on this criterion.

9. Are TEEs for individual teachers consistent or reliable from one year to the next?

Two criteria were used to answer this question:

9.1 Are teachers in each quintile of performance in one year in the same quintile in the other years?

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
+	-	+	N/A	+	+	-	-

- This was assessed by calculating separate models for each year of three years (2007-08, 08-09, and 09-10) for each VAM using NC actual data; estimating the teacher effects using the results of these VAMs; identifying teachers in each quintile of the distribution of teacher effects in each year and then comparing across years. If teachers' effectiveness was actually the same from year to year and the VAM were perfectly reliable, for example, teachers in the first quintile in any year would be in the first quintile in every other year.
- This criterion was assessed two ways. First, we observed the percentage of teachers in the same quintile in each year. If the teachers' effectiveness were actually the same and VAMs were perfectly reliable, the percentage would be 100%. Second, we looked at the percentage of "switchers", who were in the top quintile in one year and then in the bottom quintile in the next. This percentage should be zero if teachers don't move from first to worst or vice versa from one year to the next.
- It is difficult to know how much teachers' effectiveness actually changes from year-to-year and therefore, how to interpret the changes. However, switching from highest to lowest or lowest to highest seems likely to be relatively low.
- VAMs with less change from year to year in teacher quintile classification were labeled +.
- Findings:
  - The URM is the only consistent performer among all models on both agreement and switching. The HLM2, SFE and TFE perform well depending on the subject and

grade level. The IV models (SFEIV and TFEIV) are persistently poor performers. The HLM2 always outperforms the HLM3 model.

  - As measured by the agreement between two years' quintiles on reading, reliability is not very high across all models. The percentage in the same quintile in each year is between 15 and 40%; in rare cases it goes as high as 54%. When measured by "switching" confirms the poor reliability; up to 27% of the teachers appear to switch from highest to lowest or lowest to highest.
  - On math, alternatively, reliability was much higher, with % agreement approaching 70% in some cases, and except in rare cases, generally higher than 20%. Switching was very low, rarely exceeding 10%; for the best models (the URM, HLM2 and sometimes the TFE) it was rarely higher than 5%.

- The MRM was not assessed on this criterion.

9.2 Are teachers in the top 5% (or bottom 5%) of performance in one year also in the top 5% (or bottom 5%) in the other years?

HLM2	HLM3	URM	MRM	SFE	TFE	SFEIV	TFEIV
+	-	+	N/A	+	+	-	-

- This was assessed by calculating separate models for each year of three years (2007-08, 08-09, and 09-10) for each VAM using NC actual data; estimating the teacher effects using the results of these VAMs; identifying teachers in the top and bottom 5% of the distribution of teacher effects in each year and then comparing across years. If teachers' effectiveness was actually the same from year to year and the VAM were perfectly reliable on this criterion, teachers in the top 5% in any year would be in the top 5% in every other year.
- To assess this criterion we observe two trends: first, the percentage of teachers in the top 5%, bottom 5% and middle 90% in both periods; second, the percentage of teachers in the top 5% in one period and then bottom 5% in the other (or vice/versa). We label the first "agreeing" and the second "switching".
- Findings:
  - There are many combinations of grade level and year, particularly in math, when there are no switching teachers. There are more than twice as many combinations of grade level and year in which teachers switch in reading. It usually a very low proportion, no more than 5% in most cases (there are a handful above 10%).
  - The number in the "middle 90%" who are in agreement is very high in all cases, usually above 90% of the teachers in that category.
  - For math, the number in agreement in the top 5% and bottom 5% is often low, usually less than 40% of the top/bottom 5 in any year; in fact, most of the teachers in the top 5% or bottom 5% in any year find themselves in the middle in the other, demonstrating some evidence of regression to the mean (because the rankings are

based on single-year estimates, they are much more sensitive to extreme values in any given year).

- For reading, the agreement numbers are always lower, by as much as 20%.
- The smaller numbers of teachers in the top and bottom 5% make the trends less stable than we observed for the quintiles, but we still conclude that the URM is a relatively high performer throughout, with the HLM2, SFE and TFE doing well depending on the grade level and subject.
- The MRM was not assessed on this criterion.