Estimating the Impact of “Real World Outcomes” on Teacher Effectiveness Measurements: Initial Results from RealVAMs Project NSF - DRL #1336027

Jennifer Broatch *

Abstract
With evaluation systems increasingly reliant upon value added measures of teacher/program effectiveness, it is imperative to create the most accurate, reliable and valid estimates of contributions to student achievement. Value Added Models (VAMs) attempt to measure a teacher’s impact on student achievement beyond what is expected of a student given the student’s and his or her peers’ prior performance and demographic information. This paper discusses current modeling developments including initial results from the NSF Funded RealVAMS project that extends traditional VAM estimates by incorporating “real world” outcomes such as college entry.

Key Words: Value-added Modeling, accountability, mixed models, teacher effectiveness

1. Introduction

As a result of the Race to the Top (2011) competition (U.S. Department of Education, 2011), almost all states have implemented teacher evaluation systems that are highly reliant (50-100%) on some form of value-added model (VAM). All value added models (VAMS) currently used for accountability measure teacher effectiveness from standardized assessments with a vertically equated scale. They then use information from the test manufacturers for scaling, or other item response theory methods (Ballou et al., 2004; Martineau, 2006). Hence, districts and states are limited to specific standardized tests to measure teacher effectiveness that are equivalent/comparable year after year.

Recently, there has been a shift in the goals and standards for education in the United States, namely the development new standards to ensure that our high school graduates are “College and Career Ready” (U.S. Department of Education, 2014). What does College and Career Readiness look like? How can we measure a teacher’s impact on a student’s achievement toward the ultimate goal of College and Career Readiness? Often, the goal of a state, district or science, technology, engineering and mathematics (STEM) education project is not adequately measured by a test. Real-world outcomes, such as STEM career-persistence, college entry, etc..., are likely more relevant outcome measures. The RealVAMS model discussed here is an innovative approach to assess the impact of a teacher or intervention program on “real-world outcomes” in addition to traditional standardized test. In the RealVAMS project, we measure a teacher’s impact directly on whether the student enters or graduates from college in addition to a student’s achievement which yields multidimensional estimates of teacher effectivenes.

While a multidimensional VAM cannot capture the full complexity of student achievement, this methodology can provide a much better picture than relying solely on univariate test scores. The paper will discuss the RealVAMS model in section (2) and provide an example of the model using data from a large public school district

*Arizona State University at the West Campus, PO Box 37100, Phoenix, AZ 85069-7100

in section (3). The appendix (4) will present the open source package RealVAMS developed in R Statistical Computing software (R Development Core Team, 2006) to execute the multidimensional model.

2. Features of RealVAMS model

The RealVAMS model considers real world outcomes and outcomes that are not vertically equated by using a necessarily more complex mixed model to allow longitudinal non-equated continuous responses and real-world categorical responses. This model was initially developed and presented by Jennifer Broatch and Sharon Lohr (2012). This framework expands on the traditional multivariate value added model to estimate multidimensional effects. In the multidimensional model, the student responses may be different quantities; for example, the multivariate responses may include scores in different subjects (math and English-Language Arts (ELA)), scores from different assessment instruments (SAT, ACT), and categorical responses such as whether or not a student graduated or persisted in a STEM field. The RealVAMS model simultaneously estimates a different effect for each teacher for each response, rather than one overall estimate for the teacher.

Consistent with the value added modeling general framework, the student and classroom baseline characteristics, such as ethnicity, free lunch participation, sex, and measures of prior achievement (previous year's achievement score), will still remain as potential covariates in the model. The multivariate model is powerful for studying relationships among teachers and students in a multi-response (i.e. real) setting. RealVAMS will use a multivariate mixed model framework for student test achievement and real-world achievement. Although a multivariate approach can be computationally complex, it "exploit[s] the availability of tests in multiple subjects to improve the precision of estimation of teacher effects on any specific subject," "reduces confounding of teacher assignment with student background," and "should also increase robustness of results to non-ignorable missingness" (Raudenbush, 2004, pg. 127).

For simplicity, the RealVAMS model will be introduced where the responses for student i are all continuous, typically test scores. The RealVAMS model for the vector of t responses of student i presented in a traditional mixed model framework is:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{S}_i\boldsymbol{\eta} + \boldsymbol{\varepsilon}_i. \tag{1}$$

The $[t \times (p+1)]$ matrix $\mathbf{X}_i$ gives the p covariates of student i. These may include time-invariant covariates such as gender and ethnicity as well as time-varying covariates such as participation in free lunch programs. These covariates will vary by district and state (MET Researchers, 2012; White and Rowan, 2012). The $(t \times tm)$ matrix $\mathbf{S}_i$ indicates which teacher instructs student i, for each of the t responses. In the model, $\mathbf{S}$ can be expanded to allow for fractional instruction by different teachers in one time period. Overall, Model (1) considers t potentially different continuous responses that do not require time ordering or scaling. The latent teacher effect for teachers $j = 1, \ldots, m$ for responses $k = 1, \ldots, t$ is represented by the vector $\boldsymbol{\eta}$. For teacher j, $\boldsymbol{\eta}_j = (\eta_{j1}, \ldots, \eta_{jt})'$ is the t-vector of effects, where η_{jk} represents the effect of teacher j on response k. Thus, the vector $\boldsymbol{\eta} = [\boldsymbol{\eta}_1^{'}, \ldots, \boldsymbol{\eta}_m^{'}]'$ is the concatenation of the individual teacher vectors for all m teachers.

We would, however, expect the components of the teacher effects, $\boldsymbol{\eta}_j$ to be correlated. This addresses the concern with the EVAAS model (Wright et al., 2010). That is, when modeling math and science scores for example, we would expect a teacher's effect on a student's math score and his or her effect on a student's science score to be related. Similarly, we would hypothesize that a teacher's effect on a student's achievement on a standardized test would be related to their effect on a real-world outcome such as college entry. Therefore, we allow for this flexibility similar to that of Mariano et al. (2010) by setting $\text{cov}(\boldsymbol{\eta}_j) = \mathbf{G}_j$ where $\mathbf{G}_j$ is a nonnegative definite matrix. The VAMs in the current literature, with the exception of Mariano et al. (2010), assume a teacher's effects are independent across responses, so $\text{cov}(\eta_{kl}, \eta_{kl'}) = 0$ for $l \neq l'$. Finally, $\boldsymbol{\varepsilon}_i$ is assumed to follow a normal distribution with mean 0 and variance $\mathbf{R}_i$, where $\mathbf{R}_i$ is unrestricted and allows the student responses to be correlated over the t responses. Additionally, all $\boldsymbol{\eta}_j$ and $\boldsymbol{\varepsilon}_i$ are assumed uncorrelated.

The model also assumes teachers are independent so that $\boldsymbol{\eta}$ is normally distributed with mean 0 and variance $\mathbf{G} = \text{blockdiag}(\mathbf{G}_1, \ldots, \mathbf{G}_m)$ where

$$\mathbf{G}_j = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1t} \\ \vdots & & & \vdots \\ g_{1t} & g_{2t} & \cdots & g_{tt} \end{bmatrix}$$

and all $\mathbf{G}_j$ are initially assumed equal.

The model in (1) assumes that all responses are continuous and normally distributed. In order to include real-world or categorical responses, we employ a generalized linear mixed model (GLMM) to allow binary responses in addition to the continuous response. The generalized linear mixed model can be viewed as a simple extension of the continuous model presented. The critical difference stems from the calculation of the VAM estimates. The method of estimation is admittedly more complex and computationally intensive.

Nonetheless, for binary responses, we adopt the continuous response model (1) for an unobservable latent trait $\tilde{\mathbf{y}}$. The binary response is defined to be $y_{ij} = 1$ if the latent variable $\tilde{y}_{ij} > 0$. For example, suppose STEM field graduation is the binary response of interest, then the response would equal 1 if the student graduated in a STEM field and 0 otherwise. A response $y_{ij} = 1$ is then equivalent to the student's underlying latent STEM graduation trait exceed some threshold, $\tilde{y}_{ij} > 0$. Thus,

$$\tilde{\mathbf{y}}_i \quad = \quad \mathbf{X}_i\boldsymbol{\beta} + \mathbf{S}_i\boldsymbol{\eta} + \tilde{\boldsymbol{\varepsilon}}_i. \tag{2}$$

where $\boldsymbol{\eta} \sim N(0, \mathbf{G})$ and $\tilde{\boldsymbol{\varepsilon}} \sim N(0, \mathbf{R})$. To maintain the identifiability of the parameters, we take $\mathbf{R}_i$ to be a correlation matrix. The other terms in the model are defined as in (1).

The GLMM contains the linear mixed model inside the inverse link function:

$$E[\mathbf{y}_i \mid \boldsymbol{\eta}] = g^{-1}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{S}_i\boldsymbol{\eta}) \tag{3}$$

where $g(\cdot)$ is a multivariate probit link function for a binary response, following the recommendation of McCulloch (1994) and Rabe-Hesketh and Skrondal (2001). We can include the typical student achievement tests and real-world response by using the identity link for the continuous responses (for example, math test score) and the probit link for the binary responses (for example, STEM field graduation). Thus, we can simultaneously model teachers' effects on the likelihood of college entry or graduation as well as their effects on math standardized test scores.

In summary, the RealVAMS model (also presented in Broatch and Lohr (2012)) is an extension of a standard mixed model with the following key distinctions:

- Unrestricted non-equated responses.
- Multidimensional teacher effects: $\boldsymbol{\eta}_j = [\eta_{j1}, \ldots, \eta_{jt}]'$
- Assumes teachers are independent so that $\boldsymbol{\eta} \sim N(0, \mathbf{G})$ with $\mathbf{G} = \text{diag}(\mathbf{G}_1, \ldots, \mathbf{G}_m)$ where all $\mathbf{G}_j$ are assumed equal:

$$\mathbf{G}_j = \begin{bmatrix} g_{11} & g_{12} & \cdots & g_{1t} \\ \vdots & & & \vdots \\ g_{1t} & g_{2t} & \cdots & g_{tt} \end{bmatrix}. \tag{4}$$

- Allows all effects of the same teacher to be correlated, even if they are teaching different subjects or the response is measured on a different scale.
- Dramatically differs from the typical VAMS that estimate one teacher effect, $\boldsymbol{\eta}_j$ for each teacher $j = 1, \ldots m$, $\boldsymbol{\eta} \sim N(0, \sigma^2_\eta)$

Computational Issues Implementing the Model

There are computational issues that arise from the complexity of the model, yet an important part of this research is making the methodology accessible for use by educational researchers and practitioners. The primary issue with the RealVAMS model presented (2) is the computational difficulty in estimating the teacher effects and correlations, specifically the the teacher effects are allowed to be correlated (the off-diagonal of $\mathbf{G}$ are not assumed to be 0). This is because finding maximum likelihood estimators directly using the likelihood is a challenging computational problem. Similarly, because of the complexity of the structure of the covariance matrix of the value added estimates, quadrature methods such as Gauss-Hermite integration are impractical. Broatch and Lohr (2012) use the pseudo-likelihood approach described to perform computations and obtain approximations to the maximum likelihood estimators; they then adopt the penalized quasi-likelihood approach used in SAS PROC GLIMMIX (SAS Institute Inc., 2008) to approximate the maximum likelihood estimates (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) (Syntax for SAS is presented in Broatch and Lohr (2012)).

Value added estimates calculated in GLIMMIX presented in Broatch and Lohr (2012) are limited. In a simple model that modeled one continuous response and one binary response, SAS had "insufficient memory" to estimate the teacher effects for more than 30 teachers. In fact, 24 teacher effect estimates took nearly three days to compute. Karl, Yang, and Lohr (2013) overcome these computational challenges by using the EM algorithm as an alternative to the pseudo-likelihood approach for a VAM with continuous responses similar to model (1) and $\mathbf{G}_{ij} = 0$ for $i \neq j$ (the off-diagonal of $\mathbf{G}$ are assumed to be 0). The method can be implemented in the GPvam package (2012) in the open source software R (R Development Core Team, 2006). The RealVAMS package presented in section (4) expand GPvam package to fit the RealVAMS model that estimates a multidimensional teacher effect. The RealVAMS packages utilizes an efficient EM algorithm to effectively estimate the random teacher effects for teachers in a large public school district. In fact, the model presented in section (3) for 84 teachers converges in 120 seconds!

3. Application to Large Public School Data

The RealVAMS model and R package were applied to data from a large school district. This section will demonstrate the model and the multidimensional estimates that can result from using this VAM. The goal was to estimate multidimensional teacher effects on (1) student achievement measured by the state assessment (scale score on state assessment) and (2) whether or not a student entered college (College entry (Y/N)). Students in the 11th grade that took the required state assessment and had information recorded by the Clearinghouse on student college entry information were included in this application. This dataset includes 912 students and 86 teachers (students with missing data were excluded in this analysis). The model controlled for the following covariates:

- Student Gender
- Student Ethnicity
- Math/Reading PLAN scale scores (national assessment - required in district)
- Gifted status (Y/N)
- Special education status (Y/N)
- ELL status (Y/N)

Results

The covariate effects are fairly typical and are not vastly different from the standard model. Note that the covariates were allowed to vary for each response. The RealVAM fixed effects estimates are:

Covariate	State Test	SE	College Entry	SE
Intercept	-0.702	6.123	-0.791	0.3
Male	0.872	1.842	-0.134	0.102
Native American	-17.307	13.625	-1.161	0.728
African American	-3.496	3.859	-0.046	0.2
Asian American	11.05	4.602	0.566	0.304
Hispanic American	-9.849	3.789	-0.433	0.192
Two or more	-1.629	4.619	0.021	0.248
White (Reference)	–	–	–	–
Special Needs Status -YES	-15	4.103	-0.303	0.204
ESL - YES	-19.513	13.084	-0.026	0.664
Gifted- YES	6.78	2.473	-0.188	0.139
PLAN- Math	4.972	0.279	0.045	0.015
PLAN - Reading	0.777	0.26	0.03	0.014

Figure (1) displays the relationship between the teacher effect estimate on the state assessment and the teacher effect estimate on the probability of college entry. This shows a significant relationship between the two estimates ($r_G = 0.79$).

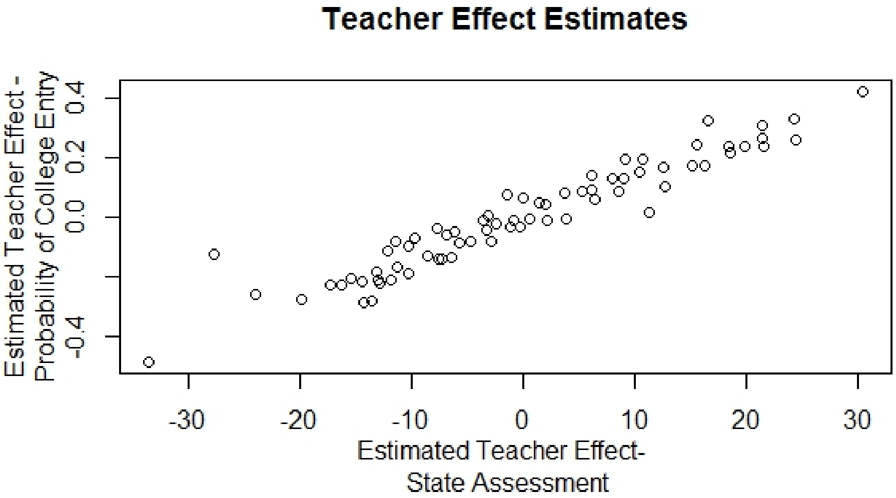

Figure 1: Teacher Effectiveness Estimates - Random Effects from model

Similar to other VAM applications the variance component due to the students is much larger than the variance component due to the teachers with $r_R = 0.15$ (Schochet & Chiang, 2010). The multidimensional teacher effects were also highly correlated, $r_G = 0.79$. As a result, the g_{12} (off-diagonal) is significantly different from 0 ($p = 0.007$).

Teachers: $\hat{\mathbf{G}}_j$=

$$\begin{bmatrix} 264.75 & 3.3765 \\ 3.3765 & 0.0691 \end{bmatrix}$$

Students: $\hat{\mathbf{R}}_i$=

$$\begin{bmatrix} 602.00 & 3.62 \\ 3.62 & 1.00 \end{bmatrix}$$

,

Multidimensional Teacher Effects

There was an outlying teacher that the RealVAMS multidimensional teacher effects highlights, see figure (2). This teacher is near the bottom for VAM estimates for the state standardized assessment, yet the highlighted teacher is not near the bottom for the VAM estimate toward student college entry.

	All Teachers	Teacher #22
Percentage of College Entry:	70.9%	76.0%
State Assessment:	121.90	55.20

This teacher is an example of where the multidimensional effectiveness measurements provide conflicting information. This teacher highlights a real scenario where a teacher would be labelled as “highly” ineffective when measured by student achievement on a standardized test, yet the teacher would not be labelled highly ineffective on the teacher’s impact towards their students’ college entry. If VAMS are to used for teacher accountability, it seems prudent to measure multiple dimensions of teacher effectiveness as teachers may contribute to different aspects of student “achievement.”

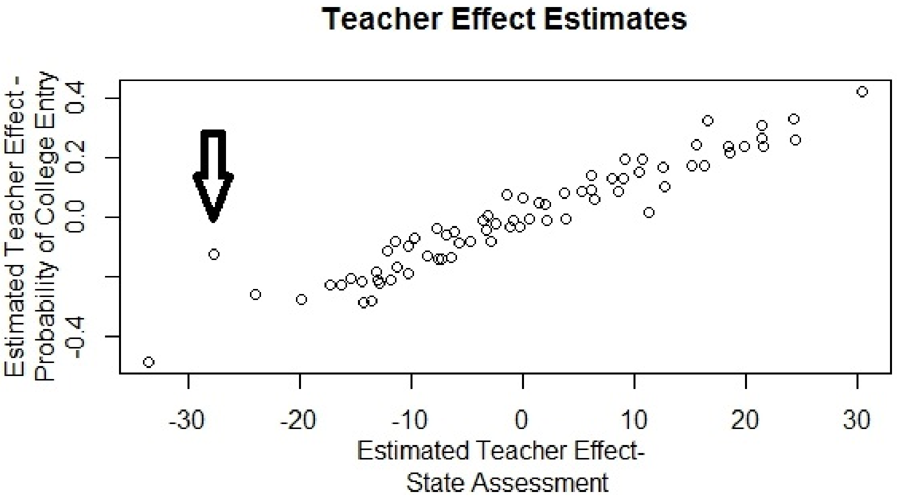

Figure 2: Outlying Teacher Highlighted

Acknowledgments

I also would like to thank fellow panelists Jennifer Green (Montana State University), Pamela Fellers (University of Nebraska-Lincoln), and Eric Parsons (University of Missouri) for their valuable prospectives of VAMS. This material is based upon work supported by the National Science Foundation under Grant No.1336027. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Ballou, D., Sanders, W., and Wright, P. (2004). Controlling for student background in value-added assesment of teachers. *Journal of Educational and Behavorial Statistics*, 29:37–65.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.

Broatch, J. E. and Lohr, S. (2012). Multidimensional assessment of value added by teachers to real-world outcomes. *Journal of Educational and Behavioral Statistics*, 37:256–277.

Karl, A., Yang, Y., and Lohr, S. (2012). GPvam: Maximum Likelihood Estimation of Multiple Membership Linear Mixed Models Used in Value Added Modeling. *R-Core Development*, pages Retrived online: `http://cran.r--project.org/web/packages/GPvam/index.html`.

Karl, A., Yang, Y., and Lohr, S. (2013). Efficient maximum likelihood estimation of multiple memebership linear mixed models, with an application to educational value-added assessments. *Computational Statistics and Data Analysis*, 59:13–27.

Mariano, L. T., McCaffrey, D. F., and Lockwood, J. R. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics*, 35:253–279.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavorial Statistics*, 31:35–62.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, 89:330–335.

MET Researchers (2012). Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains. *MET Project Research Paper-Bill and Melinda Gates Foundation*, Initial Year 2 Findings from the MET Project:Retrived online: `http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf`.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabe-Hesketh, S. and Skrondal, A. (2001). Parameterization of multivariate random effects models for categorical data. *Biometrics*, 57:1256–1264.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29:121–129.

SAS Institute Inc. (2008). *SAS/STAT 9.2 User's Guide*. SAS Institute Inc., Cary, NC.

U.S. Department of Education (2011). Race to the Top. Retrived Online: `http://www2.ed.gov/programs/racetothetop/index.html`.

U.S. Department of Education (2014). College and career standards and assessments. Retrived Online: `https://www2.ed.gov/policy/elsec/leg/blueprint/faq/college-career.pdf`.

White, M. and Rowan, B. (2012). Measures of effective teaching (MET) longitudinal database (LDB): A user guide to the “core study” data files available to met early career grantees. *Produced for Inter-University Consortium for Political and Social Research*, page Retrived online: `http://www.naeducation.org/MET_User_Guide_Data_Files.pdf`.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:233–243.

Wright, S. P., White, J. T., Sanders, W. L., and Rivers, J. C. (2010). Sas evaas statistical models. *SAS Institute - Cary, N.C.*, pages Retrived online: `http://www.sas.com/resources/asset/SAS--EVAAS--Statistical--Models.pdf`.

4. Appendix: RealVAMS Package

RealVAMS R Package: “RealVAMS” function arguments:

- **score.data:** data frame that includes - continuous response, unique student IDs, teacher IDs and time period indicator
- **outcome.data:** data frame that includes - binary response, unique student IDs, teacher IDs and time period indicator
- **persistence:** choices are “CP” or “VP”, for complete and variable persistence of the teacher score effects. Note that the binary response is limited to “CP”
- **score.fixed.effects, outcome.fixed effects** - covariates
- **max.iter.EM:** the maximum number of EM iterations during each pseudo-likelihood iteration
- **tol1:** Convergence tolerance for EM algorithm during each interior pseudo-likelihood iteration.
- **pconv:** Convergence criterion for outer pseudo-likelihood iterations. Compare to the PCONV option of SAS PROC GLIMMIX.