Teacher Effectiveness Index as an Aid to Determine Performing Teachers for Promoting Excellence in Education

A.C. Singh[1], E.C. Hedberg[1], T.B. Hoffer[1], and A. M. Kuyper[2]
[1]NORC at the University of Chicago, Chicago IL 60603
[2]Northwestern University, Evanston, IL
singh-avi@norc.org; ech@uchicago.edu; hoffer-tom@norc.org;
arendkuyper2014@u.northwestern.edu

Abstract

Direct measures of teacher effectiveneww such as average class test score or student evaluation score are subject to potential measurement biases induced by the teacher or school. As a supplement, a bias-free indirect measure of teacher effectiveness would be highly desirable. Value-added models (VAM) based on random coefficient regression modeling of student test scores provide an indirect measure of the teacher's effect for a given course (or subject) and grade by treating the teacher effect specific to the student as random. The effect varies in general with the subject, student characteristics, teacher characteristics and school policies. It is averaged to obtain an indirect measure of teacher's overall effect. However, there are several issues such as lack of a standard student population resulting in the parameter of interest being confounded with student, teacher, and school covariates, conceptual problems in interpreting student's gain in score as a causal effect due to nonrandom assignment, bias in test scores, and instability due to small sample size for estimation because of limited class size. We propose an alternative "teacher effectiveness index" (TEI) based on a new construct of latent teacher effectiveness (which may represent several behavioral characteristics) introduced explicitly in the model as a latent trait or unobserved random covariate (with mean 0 and variance 1) which is not dependent on student, teacher, and school level observed covariates. It is shown that TEI can overcome several limitations of the existing VAM measures. The teacher effect under a new VAM formulation turns out to be a product of TEI and a covariate-specific component which is a function of characteristics of school, teacher, and students taught by the teacher. The first (TEI) component is free from the subject and grade taught by the teacher and can be estimated more precisely because data from all students taught by the teacher in the same year regardless of grade and subject can be combined. The resulting stable estimate can be used to compare TEI across subjects, grades and schools. Moreover, the second component, being free from random effects, can then be estimated quite precisely although perhaps with limited utility as it remains subject to biases due to assignment and measurement error as in the case of the conventional VAM estimate.

Key Words: Common Factor or Latent Trait; Teacher-Induced and School-Induced Test Score Biases; Unobserved Covariates; Value Added Models

1. Introduction

The problem of improving the teaching force in America's public schools has moved to the forefront of federal, state, and local policy in recent years. The federal government's 'Race to the Top' education plan has established financial incentives to state and local educational authorities to base teacher performance evaluations in part on their students' test scores, typically using some measure of student gains from one year to the next to estimate the effect of the teacher. This has led to revisiting a long-standing research problem of measuring the effects or value-added (in student's gain in score) due to various factors such as the teacher, the school and the student. The goal is to isolate the effect of the teacher factor on students' performance from the effect of the school factor as well as from the effect of the student factor for a given course subject in a given grade or year. The teacher factor drives teacher effectiveness in educating students, while the school factor (school policies) drives school effectiveness in motivating and allocating appropriate resources to both teachers and students, and the student factor drives student effectiveness in learning. These factors of teacher, school, and student act as covariates and affect student performance outcome measures such as standardized test scores in various subjects. The value-added analyses seek to extract from test scores an overall teacher's value-added effect that can be used for identifying high- and low-performing teachers. For a review of value-added effects of schools and teachers, see Braun, Chudowsky, and Koenig (2010), Braun and Wainer (2007, Ch 27), and McCaffrey et al. (2004) among others.

Considerable progress has been made in the last twenty years or so on measuring value-added effects of teachers by extracting them from student test scores; see the special issue of Journal of Educational and Behavioral Statistics (2004). However, there still remains controversy among researchers about how to define and measure value-added effects. The main reason for this controversy is that the value-added modeling (VAM) philosophy holds schools and teachers accountable for student learning. However, due to nonrandom assignment of teachers to schools (reflecting local labor markets, linkages between teacher preparation programs and districts, and district rules affecting how teachers are assigned to schools) and, within schools, the students to teachers (usually reflecting the school administration's beliefs about the most appropriate matches between teachers and students), an estimated teacher effect on value-added is subject to selection or assignment bias. For this reason, it is difficult to give a causal interpretation of value-added effects of teachers without making strong assumptions about ignorability of the above two levels of assignment; see the important discussion by Raudenbush (2004).

The purpose of VAM measures is to provide indirect measures of teacher effectiveness as a supplement to direct measures such as principals' classroom observations, average class test score or pass rate, or average teacher evaluation score by students. The direct measures may not be adequate by themselves because of potential biases; see e.g., Barlevy and Neal (2009). These biases include teacher-induced measurement bias due to their disingenuous behavior in terms of the possibility of teaching to the test and/or diluted versions of the course content taught for favorable evaluation, and school-induced selection bias due to nonrandom assignment of teachers to schools and students to teachers. In view of these biases it is quite possible that a teacher might be basically effective but has a poor relative performance measure simply due to the disingenuous behavior of other teachers (e.g., teaching to the test) or less favorable assignment to a class consisting of largely low performing or less motivated students. As a supplement to the direct measure, it would therefore be desirable to have an indirect measure of teacher effectiveness that is not subject to the kinds of bias mentioned above.

There exists literature on VAM that attempts to measure the teacher's effect on a student test score in a given subject and grade by treating the teacher effect specific to a student as a random variable; this is introduced via random regression coefficients in the model for the mean test score under a linear mixed model with test score as the dependent variable. VAM uses student's past score and background variables for student, teacher, and school as independent variables or covariates. The teacher effect varies in general with the subject, grade, student characteristics, and school policies. It is averaged to obtain an indirect measure of the teacher's overall value-added effect. However, there are several issues with this indirect measure under VAM as listed below.

First, the underlying teacher effect parameter being estimated by this indirect measure may not be particularly meaningful in practice. The reason for this is that a standard population of students representing different background and school characteristics that could be assigned to the teacher is needed over which the parameter can be defined as the average effect on student test scores for a given subject and grade taught by the teacher. However, this is not possible due to practical limitations in teacher and student assignment. Moreover, it is difficult in general to estimate this parameter with precision because of class sizes not being large enough. In addition, the parameter definition is confounded in general with the observed covariates used in the model for student, teacher, and school corresponding to a given subject and grade; thus making it difficult to use this measure for comparison across teachers from different schools, subjects, and grades. Second, due to nonrandom assignment of students to teachers, and teachers to students, there is also the conceptual problem in interpreting a student's gain in score as a causal effect of the teacher. Third, even if we can assume that given the covariates, the assignment mechanism of students to teachers, and teachers to students is ignorable (i.e., not subject to selection bias) for VAMs, all the estimated fixed parameters in the VAM mean function and variance components are affected by the presence of potential measurement bias in the dependent variable as mentioned earlier which implies that the VAM indirect measure of teacher effect is also susceptible to measurement bias like the direct measure.

In light of the above concerns, we propose a "Teacher Effectiveness Index," or TEI, based on a new construct of latent teacher effectiveness. This measure is an index as it may represent several indicators or behavioral characteristics of the teacher. It is introduced explicitly in a new VAM formulation as an unobserved random covariate standardized with mean 0 and variance 1. The proposed index uses the framework of common factor or latent trait models (also a special case of linear mixed models; see Skrondal and Rabe-Hesketh, 2004, pp.66) to capture the unobserved heterogeneity; and it turns out that the teacher effect measure is now obtained as a product of the proposed TEI measure and an adjustment factor depending on the observed characteristics of student, teacher, and school. We observe that the new indirect measure or index by definition is not subject to measurement biases in location and scale of the dependent variable because of pre-specified values of mean and variance. It is also not subject to problems in causal interpretation of gain in score because it is designed to measure the teacher effectiveness construct and not the value-added effect. Moreover, since it is not specific to a given subject and grade, it can be estimated more reliably by using test scores of all students in different subjects and grades taught by the same teacher in a given year. In addition, since by construction it is free from observed characteristics of student, teacher, and school, it is suitable for comparison across teachers within and between schools. Thus, the proposed TEI measure may be of considerable interest independently of teacher's value added effect and can serve as an effective supplement to the direct measure.

The organization of this paper is as follows. A simple version of the current VAM formulation based on random coefficient regression models with a single covariate is first reviewed in Section 2 followed by the proposed new formulation in terms of latent trait or common factor models in Section 3. Further comparison of the two formulations is considered in Section 4 followed by their generalizations to complex models with multiple covariates at student, teacher, and school levels in Section 5. Finally, summary and remarks are presented in Section 6.

2. Review of Current VAM Formulation using Simple Models with a Single Covariate

Let $y_{ijk(g)}^{(a)}$ denote the standard test score for student *i* in grade *g* under teacher *j* in school *k* for subject '*a*' which we will also denote simply by y_{ijk} after dropping the superscript '*a*' and subscript *g* whenever it is clear from the context. Let $y_{ij'k'(g')}^{(a)}$ denote the test score for the same subject at the end of previous grade g' under possibly a different teacher j' and a different school k' which will also be denoted by x_{ijk} and that student's grade for the same subject in the previous year-end (i.e., Spring time) is available as a pre-test score. We will assume for simplicity that this is the only covariate available. Note that use of the prior score $y_{ij'k'(g')}^{(a)}$ as a covariate is commonly done for longitudinal data; see Diggle, Liang, and Zeger (1996). Now a linear regression model incorporating teacher and school factor effects to capture the extra variability due to these factors (besides heterogeneity captured by the observed covariates) can be defined by including fixed teacher-specific (*u*-factor) and school-specific (*v*-factor) effects in the regression coefficients of intercept and slope as follows.

$$\begin{aligned} y_{ijk} &= (\gamma_0 + u_{0jk} + v_{0k}) + (\gamma_1 + u_{1jk} + v_{1k})x_{ijk} + \varepsilon_{ijk} \qquad (1) \\ &= (\gamma_0 + \gamma_1 x_{ijk}) + (u_{0jk} + u_{1jk}x_{ijk}) + (v_{0k} + v_{1k}\, x_{ijk}) + \varepsilon_{ijk} \end{aligned}$$

where $\varepsilon_{ijk} \sim_{iid} N(0, \sigma_\varepsilon^2)$, and for parameter identifiability, the constraints $\sum_j u_{0jk} = 0 = \sum_j u_{0jk}$, and $\sum_k v_{0k} = 0 = \sum_k v_{1k}$ are imposed. The parameters γ_0 and γ_1 are the usual fixed regression parameters of intercept and slope. With many teachers and schools and small class sizes in general, the number of regression parameters in model (1) is too many for the purpose of precise estimation. A common way out is to introduce extra information by making *u*- and *v*- parameters (or effects) random (see Skrondal and Rabe-Hesketh, 2004, pp.50) with prior distributions given by

$$\begin{pmatrix} u_{0jk} \\ u_{1jk} \end{pmatrix} \sim_{iid} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \varphi_{00} & \varphi_{01} \\ \varphi_{01} & \varphi_{11} \end{pmatrix}\right), \ \begin{pmatrix} v_{0k} \\ v_{1k} \end{pmatrix} \sim_{iid} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{00} & \psi_{01} \\ \psi_{01} & \psi_{11} \end{pmatrix}\right), \qquad (2)$$

where $\varphi -$ and $\psi -$parameters are variance-covariances of random effects *u* and *v* which are assumed to be independent of each other. The resulting model is a random coefficient regression (rcr) model which can also be expressed as a multi-level model (Raudenbush and Bryk, 2002, p.85) in terms of new $\alpha -$ and $\beta -$ random parameters as defined below:

$$\textit{Level 1: } y_{ijk} = \alpha_{0jk} + \alpha_{1jk}x_{ijk} + \varepsilon_{ijk} \qquad (3a)$$

$$\begin{aligned} \textit{Level 2: } \alpha_{0jk} &= \beta_{0k} + u_{0jk} \\ \alpha_{1jk} &= \beta_{1k} + u_{1jk} \qquad (3b) \end{aligned}$$

$$\textit{Level 3: } \beta_{0k} = \gamma_0 + v_{0k}$$
$$\beta_{1k} = \gamma_1 + v_{1k} \qquad (3c)$$

where the four original *u*- and *v*- random parameters are related to the new random parameters via equations $u_{0jk} = \alpha_{0jk} - \beta_{0k}$, $u_{1jk} = \alpha_{1jk} - \beta_{1k}$, $v_{0k} = \beta_{0k} - \gamma_0$, and $v_{1k} = \beta_{1k} - \gamma_1$.

For students taught by the teacher *j* in school *k*, the VAM teacher effect under the above rcr model is defined as

$$\omega_{jk,teacher}^{rcr.vam} = \sum_i (u_{0jk} + u_{1jk} x_{ijk}) / m_{jk}\ , \qquad (4a)$$

and the VAM school effect for all students and teachers in school *k* can also be defined similarly although it is the teacher effect that is of primary interest, as

$$\omega_{k,school}^{rcr.vam} = \sum_{ij} (v_{0k} + v_{1k} x_{ijk}) / m_k \qquad (4b).$$

where m_{jk} is the class size, and m_k is the school size. Besides depending on fixed covariates, the above measure depends on unobserved covariates or random effects which in general are specific to student, teacher, and school as well as grade and subject. It follows from Appendix I that although random effects can be estimated using best linear unbiased prediction (BLUP) theory, their mean square error (MSE) is typically not small because of many parameters and not enough observations per parameter. The fixed first order parameters (γ) and second order parameters (variance and covariance components) can be estimated consistently using maximum likelihood or restricted maximum likelihood although the latter is usually preferable in practice. Main limitations of VAM based on rcr models are listed below.

(i) There is a conceptual problem in the teacher effect parameter definition under VAM because of lack of a standard class; the measure depends on the assignment (or sorting) of students to teachers and teachers to schools. As a result, it is difficult to employ such a measure to compare teachers across subjects within the same school or across schools within the same subject.

(ii) The nonrandom assignment of students to teachers and teachers to schools makes it difficult to draw any causal inference about teacher's effect on student's gain in score.

(iii) It is subject to measurement biases (teacher-or school-induced) as mentioned in the Introduction. The location bias only affects the fixed model intercept but the scale bias in the *y*-variable affects other regression coefficients and random effect variances. Also any measurement bias in the *x*-variable (in particular the previous test score) affects all random effects appearing in intercepts and slopes.

(iv)There is instability in VAM estimates due to small class size per teacher which may vary considerably over years.

As explained in the next Section, the proposed VAM formulation using latent traits or common factors (Skrondal and Rabe-Hesketh, 2004, pp. 9, 66) as unobserved covariates for teacher ability or effectiveness provides an alternative way of capturing heterogeneity in the data. It is shown that the new measure of teacher effectiveness index (TEI) under the proposed VAM formulation termed Behavioral Latent Status Index (blsi) can overcome several limitations of the current VAM formulation.

3. Proposed VAM Formulation using Latent Traits

The proposed VAM formulation starts with introducing explicitly in the model a hypothetical construct (or latent trait) of teacher ability or effectiveness and another construct of school effectiveness as primary parameters of interest. Unlike the teacher effect under VAM, it is not a function of student, teacher, and school-specific observed covariates. Instead, we define ability or effectiveness index as an unknown function of unobserved covariates such as indicators or behavioral characteristics of teacher in terms of advance preparation, attitude in interaction with students, and providing inspiration and encouragement; and school in terms of providing constructive policies for learning and teaching. In principle, there could also be a latent trait for each student representing behavioral characteristics such as diligence in homework, discipline, and eagerness to learn; see Section 5 for a general formulation. The new constructs are termed behavioral latent status indices (BLSI) because they are single measures (or indices) for teachers and schools to be interpreted as composite functions of unobserved behavioral characteristics.

The blsi model for VAM can now be defined. Let θ_{jk} and η_k denote respectively standardized teacher and school effectiveness indices (TEI and SEI for short) with independent standard normal distributions; i.e.,

$$\theta_{jk} \sim_{iid} N(0,1)\ ,\ \eta_k \sim_{iid} N(0,1) \qquad (5)$$

Then the blsi model is given by

$$y_{ijk} = \gamma_0 + \gamma_1 x_{ijk} + \tau_0\, \theta_{jk} + \tau_1 x_{ijk}\theta_{jk} + \lambda_0\, \eta_k + \lambda_1 x_{ijk}\, \eta_k + \varepsilon_{ijk} \qquad (6)$$

where $\tau_0 > 0$ and $\lambda_0 > 0$ are scale adjustment coefficients for standardizing TEI θ_{jk} and SEI η_k respectively;i.e., they are square roots of the corresponding variances and can be interpreted as effects of unobserved covariates –TEI and SEI. The coefficients τ_1 and λ_1, on the other hand, are not scale adjustments because θ_{jk} and η_k are already scaled by $\tau_0 > 0$ and $\lambda_0 > 0$, but are interaction effects between observed and unobserved covariates. Equivalently, τ_1/τ_0 can be interpreted as the regression coefficient for the interaction between $\tau_0\, \theta_{jk}$ and x_{ijk} , and similar interpretation for λ_1/λ_0.

The blsi model can also be expressed as a multi-level model as follows;

$$\begin{aligned}
&\textit{Level 1: } y_{ijk} = \alpha_{0jk} + \alpha_{1jk} x_{ijk} + \varepsilon_{ijk} && (7a)\\
&\textit{Level 2: } \alpha_{0jk} = \beta_{0k} + \tau_0\, \theta_{jk} && (7b)\\
&\qquad\quad \alpha_{1jk} = \beta_{1k} + \tau_1\, \theta_{jk} &&\\
&\textit{Level 3: } \beta_{0k} = \gamma_0 + \lambda_0\, \eta_k && (7c)\\
&\qquad\quad \beta_{1k} = \gamma_1 + \lambda_1\, \eta_k &&
\end{aligned}$$

where $$\begin{pmatrix}\tau_0\theta_{jk}\\ \tau_1\,\theta_{jk}\end{pmatrix} \sim_{iid} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\tau_0^2 & \tau_0\tau_1\\ \tau_0\tau_1 & \tau_1^2\end{pmatrix}\right), \begin{pmatrix}\lambda_0\,\eta_k\\ \lambda_1\,\eta_k\end{pmatrix} \sim_{iid} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\lambda_0^2 & \lambda_0\lambda_1\\ \lambda_0\lambda_1 & \lambda_1^2\end{pmatrix}\right) \qquad (8)$$

Observe that unlike the covariance structure (2) of rcr models, it follows from (8) that under blsi models, the correlation between random errors in the intercept and slope for level 2 is one because of the common latent factor θ_{jk} and similarly for errors at level 3. This happens naturally by definition of the common factor or latent trait because given all the key covariates (only one in our simple example) are included in the model, the only covariates left out are the unobserved TEI θ_{jk} at level 2 and SEI η_k at level 3. Therefore the corresponding correlations between random errors should be one as expected.

The blsi-vam teacher effect for all students under teacher jk is given by

$$\omega_{jk,TEI}^{blsi.vam} = \theta_{jk} \sum_i (\tau_0 + \tau_1 x_{ijk})/m_{jk} \tag{9a}$$

and the blsi-vam school effect for all students and teachers under school k is given by

$$\omega_{k,SEI}^{blsi.vam} = \eta_k \sum_{ij} (\lambda_0 + \lambda_1 x_{ijk})/m_k. \tag{9b}$$

The blsi-vam measures of teacher and school effects are products of two separate components. For example, for the teacher effect, it is a product of TEI and an adjustment factor depending on observed covariates and fixed parameters. Thus, unlike rcr models, the adjustment factor is free from random effects and can be estimated consistently although it is still subject to bias as in conventional rcr-vam measures. However, the estimate of the first component can be made very stable by pooling test score data over all subjects and grades taught by the same teacher in the same year rather than just one subject and one grade. Moreover, using a time series model such as state space, TEI can be connected over years and therefore even more efficient estimates of TEI can be obtained.

Although the new blsi-vam measure of teacher effect can be estimated more precisely than the conventional rcr-vam measure, it also suffers from measurement and assignment biases like rcr-vam does due to the presence of the second component in (9a). However, the first component (or TEI) itself provides a new evaluation measure for comparing teachers with respect to their teaching ability and not their value-added effect on gain in score. TEI, by definition, is free from nonrandom assignment bias and does not require a standard student class for its definition in order to be comparable across teachers and schools. The problem of causal interpretation of teacher's effect on student's gain in score also does not arise because TEI is not designed to measure value-added effect. It simply measures the latent teaching ability and therefore provides a fair means to compare teachers in terms of their ability or effectiveness. TEI is also free from measurement scale bias in test scores (e.g., teach-to-test) because this bias only affects the second component in (9a) or the adjustment factor. In addition, it is free from bias in the covariate (such as the previous test score) because it is standardized and so any bias present will only affect the scale adjustment factors. Finally, as mentioned earlier, the TEI estimate can be made more precise by combining test score data over subjects and grades in a given year and over years through longitudinal models.

4. Further Comparison of blsi-vam and rcr-vam Formulations

It is observed that both formulations assume that the postulated model is correct. It would be of interest to see what happens if the model assumed is not correct due to missing covariates. In this case, rcr-vam does not measure the teacher effect as desired because the missing covariate affects random intercepts and slopes. For example, consider a second covariate z_{ijk} and assume for simplicity that the model has only teacher factor and no school factor effects. That is, the rcr-vam model is given by

$$y_{ijk} = (\gamma_0 + \gamma_1 x_{ijk} + \gamma_2 z_{ijk} + \gamma_3 x_{ijk} z_{ijk}) + (u_{0jk} + u_{1jk} x_{ijk} + u_{2jk} z_{ijk} + u_{3jk} x_{ijk} z_{ijk}) + \varepsilon_{ijk}. \qquad (10)$$

Now suppose the missing covariate is z_{ijk} and let $\mu_{ijk(z)}$ denote the conditional mean $E(z_{ijk}|x_{ijk})$ and $\tilde{z}_{ijk}$ denotes the deviation $z_{ijk} - \mu_{ijk(z)}$; i.e., the centered covariate. Then, the revised rcr model (10) in the presence of the missing covariate is equivalently given by

$$y_{ijk} = (\gamma_0 + \gamma_2 \mu_{ijk(z)}) + (\gamma_1 + \gamma_3 \mu_{ijk(z)}) x_{ijk} + (u_{0jk} + \gamma_2 \tilde{z}_{ijk}) + (u_{1jk} + \gamma_3 \tilde{z}_{ijk}) x_{ijk} + \tilde{\varepsilon}_{ijk}, \qquad (11)$$

where $\tilde{\varepsilon}_{ijk} = u_{2jk} z_{ijk} + u_{3jk} x_{ijk} z_{ijk} + \varepsilon_{ijk}$.

It easily follows that the rcr-vam measure (4a) does not measure the teacher value-added effect because of contamination of random intercept and slope effects of the teacher factor by the missing random covariate $\tilde{z}_{ijk}$.

Interestingly, the proposed blsi-vam model also does not measure TEI-the new parameter of interest without bias because of contamination by the missing random covariate $\tilde{z}_{ijk}$. In particular, $\tau_0 \theta_{jk}$ is replaced by $\tau_0 \theta_{jk} + \gamma_2 \tilde{z}_{ijk}$ and $\tau_1 \theta_{jk}$ is replaced by $\tau_1 \theta_{jk} + \gamma_3 \tilde{z}_{ijk}$, so that the correlation between the two is no longer 1 as was the case in the covariance structure given by equation (8). Thus both rcr and blsi formulations require that all important covariates are included so that any effect due to unobserved covariates is mainly due to teacher and school effects. In practice, a natural question arises as to which model to use in a particular application. Suppose we have an adequate model in that all important covariates are included and the goal is to extract teaching ability or TEI. It follows that such a parameter can be explicitly built in the model as a latent trait and then the blsi formulation naturally arises. In practice, model diagnostics can be performed to check sensitivity of blsi measures if some covariates are dropped based on their relative insignificance or subjective considerations. These diagnostics should be performed in addition to usual ones for covariate selection and residual analysis.

5. Generalization to Complex Models with Multiple Covariates

Here we consider more comprehensive models under the blsi-vam approach incorporating covariates representing background characteristics of student, teacher, and school. For simplicity, we consider modeling of test scores for a single subject (e.g., mathematics or reading) in a given grade. Generalizations of the model to include scores from multiple subjects and grades can also be made. Here, the proposed model is fit with two years of data although the structure allows for longitudinal data over several years. Now, for each triplet (*i*, *j*, *k*) corresponding to the *i*th student, *j*th teacher, and *k*th school for grade g, let $x_{ijk(g)}$, $x_{jk(g)}$, and $x_{k(g)}$ denote respectively the covariates representing background characteristics of student, teacher, and school. Examples of teacher's background variables $x_{jk(g)}$ are years of experience, length of service, position (FTE), race, gender; and student background variables $x_{ijk(g)}$ are race/ethnicity, gender, family structure, and socio-economic status; and school and school district background variables $x_{k(g)}$ are characteristics such as pass rate, poverty rate, SAT average, homeless count, staff type count, and average teacher salary by experience for each grade g . Here we work with only single covariates representing each of student, teacher, and school background variables for simplicity. Also let ζ_{ijk} denote PEI (pupil or student effectiveness index) besides θ_{jk} and η_k denoting TEI and SEI respectively. The outcome variable (i.e., the test scores) $y_{ijk(g)}^{(a)}$ for a given subject is expected to depend on unobserved covariates or latent traits (ζ_{ijk}, θ_{jk} and η_k), observed covariates ($x_{ijk(g)}$, $x_{jk(g)}$, and $x_{k(g)}$) for each triplet (i, j, k) and their interactions with observed and unobserved covariates, as well as on the previous test score $y_{ij'k'(g')}^{(a)}$. We also assume for simplicity that each class is taught by a single teacher for a given subject.

First consider a model consisting of only TEI θ_{jk}. Later we will introduce SEI η_k and PEI ζ_{ijk} also. The BLSI-vam model with TEI analogous to equation (6) is given by

$$y_{ijk(g)}^{(a)} = \nu_{ijk(g)}^{(a)} + \theta_{jk} h_{ijk(g)}^{(a),TEI} + \varepsilon_{ijk(g)}^{(a)}, \tag{12a}$$

where $\nu_{ijk(g)}^{(a)} = \gamma_{0(g)}^{(a)} + \gamma_{1(g)}^{(a)} y_{ij'k'(g')}^{(a)} + \gamma_{2(g)}^{(a)} x_{ijk(g)} + \gamma_{3(g)}^{(a)} x_{jk(g)} + \gamma_{4(g)}^{(a)} x_{k(g)}$,
and $h_{ijk(g)}^{(a),TEI} = \tau_{0(g)}^{(a)} + \tau_{1(g)}^{(a)} y_{ij'k'(g')}^{(a)} + \tau_{2(g)}^{(a)} x_{ijk(g)} + \tau_{3(g)}^{(a)} x_{jk(g)} + \tau_{4(g)}^{(a)} x_{k(g)}$.

Introducing SEI η_k, we obtain an enlarged model as

$$y_{ijk(g)}^{(a)} = \nu_{ijk(g)}^{(a)} + \theta_{jk} h_{ijk(g)}^{(a),TEI} + \eta_k h_{ijk(g)}^{(a),SEI} + \varepsilon_{ijk(g)}^{(a)}, \tag{12b}$$

where $h_{ijk(g)}^{(a),SEI} = \lambda_{0(g)}^{(a)} + \lambda_{1(g)}^{(a)} y_{ij'k'(g')}^{(a)} + \lambda_{2(g)}^{(a)} x_{ijk(g)} + \lambda_{3(g)}^{(a)} x_{jk(g)} + \lambda_{4(g)}^{(a)} x_{k(g)}$

Finally, if we also introduce PEI ζ_{ijk}, we get a further enlarged model as

$$y_{ijk(g)}^{(a)} = \nu_{ijk(g)}^{(a)} + \theta_{jk} h_{ijk(g)}^{(a),TEI} + \eta_k h_{ijk(g)}^{(a),SEI} + \zeta_{ijk} h_{ijk(g)}^{(a),PEI} + \varepsilon_{ijk(g)}^{(a)}, \tag{12c}$$

where $h_{ijk(g)}^{(a),PEI} = \sigma_{0(g)}^{(a)} + \sigma_{1(g)}^{(a)} y_{ij'k'(g')}^{(a)} + \sigma_{2(g)}^{(a)} x_{ijk(g)} + \sigma_{3(g)}^{(a)} x_{jk(g)} + \sigma_{4(g)}^{(a)} x_{k(g)}$,

and where γ's in $\nu_{ijk(g)}^{(a)}$ are regression coefficients different from those in (6), the dimension being the number of observed covariates x's, the random factors ζ_{ijk}, θ_{jk} and η_k, as before, are independent of each other with mean 0 and variance 1, and the model errors $\varepsilon_{ijk(g)}^{(a)}$'s are independent with mean 0 and constant variance σ_ε^2 for a given g. Observe that the model allows for effects of unobserved covariates (BLSI's) ζ_{ijk}, θ_{jk} and η_k on the outcome variable $y_{ijk(g)}^{(a)}$ to vary with the covariates x's as captured by the $\sigma, \lambda, \tau -$coeffcients. Similarly, the $\gamma -$coefficients capture the varying effects of observed covariates x's on the outcome variable. Using suitable priors and hyperpriors, a hierarchical Bayes approach can be easily used to estimate all parameters –fixed and random. However, estimates of all random effects lack precision in general, the estimate of the factor PEI (ζ_{ijk}) will be especially unreliable since there is in general only one observation or test score per student for a given subject and grade. Incidentally, if we were to extend the model to data with several subjects and past years, there would be more observations per student for estimation of ζ_{ijk}. In this longitudinal set-up, we could still assume that conditional on the previous score $y^{(a)}_{ij'k'(g')}$, $\varepsilon^{(a)}_{ijk(g)}$'s are uncorrelated across years (i.e., across grades g), but the unobserved covariates or random effects ζ_{ijk}, θ_{jk} and η_k are expected to be dependent over years. Modeling dependence of these effects over years could be done, for example, using state space models and estimation by Kalman filtering. However, for the modeling problem considered here with two successive years of data, PEI cannot be estimated well but, in fact, it could be dropped as discussed below.

In practice, the factor PEI is typically not of direct interest and could be subsumed within the model error. However, with this modified model error $\delta^{(a)}_{ijk(g)}$(denoting $\zeta_{ijk} h^{(a),PEI}_{ijk(g)} + \varepsilon^{(a)}_{ijk(g)}$), the new model errors in (12c) will no longer be independent of $y^{(a)}_{ij'k'(g')}$ due to correlation between covariates ζ_{ijk} for the same student for two successive years. This implies that bias in estimation of parameters might arise due to correlation between the covariate $y^{(a)}_{ij'k'(g')}$ and the new model error $\delta^{(a)}_{ijk(g)}$ which would require more complex estimation strategies. This problem in estimation can be avoided under the assumption of a Markovian dependence (likely to hold in practice) between ζ_{ijk}'s over years. In addition, the $\sigma-$parameters representing interactions of the student latent trait or SEI with observed covariates are expected to be very small in the presence of teacher and school latent traits, and therefore variance of the new model error $\delta^{(a)}_{ijk(g)}$ could still be assumed to be approximately constant.

The blsi-vam measure from the above model (12c) has a more general form than (9) because of more covariates, and is given by

$$\theta^{(a),blsi-vam}_{jk(g),TEI} = \theta_{jk}\left[\frac{1}{m^{(a)}_{jk(g)}}\sum_i h^{(a),TEI}_{ijk(g)}\right] \qquad (13)$$

where the multiplicative factor θ_{jk} is free from the confounding effects of student, teacher, and school covariates and is common for all subjects and grades taught by the teacher in a given year, and $m^{(a)}_{jk(g)}$ is the class size.

Next we consider a few existing rcr-vam measures under complex models such as covariate adjustment models due to Rowan et al. (2002) and variable persistence teacher effect models due to McCaffrey et al. (2004), and Mariano et al. (2010) which are generalizations of the original constant persistence model of Sanders et al. (1997). The model proposed by Rowan et al. also conditions on the prior score as in the BLSI model and works with two years of score data although it could also be defined for multiple years. However, it considers only student-level covariates and random intercept (not slope) consisting of random teacher $u^{(a)}_{0jk(g)}$ and school $v^{(a)}_{0k(g)}$ effects. More specifically, the model of Rowan et al. can be written as

$$y^{(a)}_{ijk(g)} = \gamma^{(a)}_{0(g)} + \gamma^{(a)}_{1(g)} x_{ijk(g)} + \gamma^{(a)}_{2(g)} y^{(a)}_{ij'k'(g')} + u^{(a)}_{0jk(g)} + v^{(a)}_{0k(g)} + \varepsilon^{(a)}_{ijk(g)} \qquad (14)$$

where the $\gamma-$parameters are different from those in (12), and the constructs of teacher and school effectiveness indices or unobserved covariates θ_{jk} and η_k in the BLSI approach are replaced by teacher and school random effects. It follows as mentioned earlier that the rcr-vam measure of teacher effect (which in this case is simply $u^{(a)}_{0jk(g)}$) is not suitable for comparison across teachers of different subjects or in different grades although it does not depend on observed covariates due to simplified model assumptions.

The variable persistence model of McCaffrey et al. (2004) assumes that the past teacher effects continue to persist over years but get dampened (hence the qualifier 'variable') and does not consider prior test scores as covariates. Because of the persistence assumption, it requires all past scores under a longitudinal set-up but only student-level covariates are included. More specifically, it can be written as

$$\begin{aligned} y_{ijk(g)}^{(a)} = \gamma_{0(g)}^{(a)} + \gamma_{1(g)}^{(a)} x_{ijk(g)} + \left(u_{0jk(g)}^{(a)} + \xi_{gg'} u_{0j'k'(g')}^{(a)} + \xi_{gg''} u_{0j''k''(g'')}^{(a)} + \cdots \right) \\ + \left(v_{0k(g)}^{(a)} + \chi_{gg'} v_{0k'(g')}^{(a)} + \chi_{gg''} v_{0k''(g'')}^{(a)} + \cdots \right) + \varepsilon_{ijk(g)}^{(a)} \qquad (15) \end{aligned}$$

where $\xi_{gg'}$, $\xi_{gg''}$, ... represent discount factors for past teacher effects for grades g', g'', .. and similarly $\chi_{gg'}$, $\chi_{gg''}$, … represent discount factors for past school effects for prior grades. McCaffrey et al. do mention that teacher and school level covariates could be included with random or fixed coefficients. With random coefficients, however, the model gets extremely complex. As in the case of Rowan et al., this model also does not permit a fair comparison of teachers across different subjects using the rcr-vam measure $u_{0jk(g)}^{(a)}$ of teacher effect unlike BLSI measures. The above variable persistence model is a generalization of the earlier constant persistence model of Sanders et al. (1997), well known as the Tennessee Value-Added Assessment system (TVAAS), which assumes all the discount factors as unity and does not include any covariates. A general persistence model where a teacher's effect could vary arbitrarily over years without being connected via discount factors is considered by Mariano et al. (2010).

6. Summary and Remarks

We quote the words of the famous statistician John Tukey, "All models are wrong, but some are useful." The proposed blsi-formulation based on latent trait or common factor models that directly extracts teacher's pure ability or effectiveness index (TEI) can be viewed in this spirit. It was assumed that the model has a reasonable set of observed covariates and that no important covariate is missing so that essentially the main unobserved covariates are TEI and SEI, and thus their estimates will not be seriously contaminated by missing covariates. As part of model diagnostics, it was suggested that a sensitivity analysis could be performed to check the impact on TEI estimates when some observed covariates are dropped on purpose. The proposed TEI measure is free from student, teacher and school specific covariates and is common for all subjects and grades taught by the teacher in a given year. Thus it can provide a fair means of comparison of teaching ability of teachers across different subjects and grades within a school and across schools in a given year. However, TEI does not measure the value-added effect on student's gain in score. In fact, the blsi-vam measure of the teacher effect has TEI as one of the components but the other component suffers from limitations of measurement and assignment biases as the existing rcr-vam measure. Unlike TEI for blsi-models, it seems difficult to explicitly build in a value-added parameter in the rcr-model to measure teacher's effect on student's gain in score because of confounding effects of observed covariates specific to student, teacher, and school. Moreover, it is not possible to obtain reliable estimates of teacher's value-added effect due to small class size.

Above considerations suggest the need for a new formulation which departs from the conventional VAM formulation. The proposed BLSI approach is an attempt in this direction by identifying teachers with teaching ability or effectiveness who will have favorable impact on students in the long run over several years rather than identifying teachers with tangible effects on student's gain in test score in a given year. It may be of interest to note that the proposed effectiveness indices at different levels in a hierarchy (student, teacher, and school) appear to be in line with the framework of multi-level hierarchical linear models advocated in an important early paper by Raudenbush and Bryk (1988). In reviewing the reconceptualization of school and classroom effects, they made the following statement which provides a useful perspective on the proposed BLSI approach: "*These theoretical perspectives emphasize the hierarchical, multilevel character of educational decision-making: decisions made at the school level, for example, constrain options available to teachers. Teacher's decisions, in turn, influence how opportunities for learning will be distributed across children….an adequate conceptualization for the effects of schooling will include not only a model for how schools differentially allocate resources for instructions and opportunities for learning, but also a model for how students might differentially respond given the available opportunities.*"

In practice, both direct measures in terms of average teacher evaluation score by students and average class test score and indirect measures of TEI and blsi-vam can be used to identify performing teachers. Using TEI estimates for a contemporary group of teachers available in the data, we can construct a normative distribution of TEI for a given year. Each teacher's effectiveness in any given year can be ranked in terms of percentiles of the normative distribution of TEI. Performance incentives for teachers can be determined in a fair manner based on patterns in the behavior of above ranks over years given that the direct measures of their teaching effect are above a minimal threshold. Incidentally, it is the relative rank of TEI compared to other teachers and not the absolute value that are important for determining performance incentives. Clearly, this is an area that requires further investigation. Finally, we note that it is sufficient to have a large sample data on schools, teachers, and students to fit BLSI models to obtain efficient estimates. Once the model parameters are estimated, TEI for any teacher can be estimated by providing to the model predictor input about background variables for the teacher, school, and students taught by the teacher along with prior and current test scores.

Acknowledgments

Thanks are due to Norman Bradburn for several useful discussions during the early stages of this research. Thanks are also due to Michele Zimowski for very helpful comments, and Kirk Wolter and Dan Kasprzyk for their support and encouragement.

References

Barlevy, G. and Neal, D. (2009). Pay for percentile. *Federal Reserve Bank of Chicago Working Paper* No. 2009-09.

Braun, H. and Wainer, H. (2007). Value-added modeling. In Handbook of Statistics--Psychometrics, Eds: C.R. Rao and S. Sinharay, Chapter 27, pp. 867-982.

Diggle, P.J., Liang, K.-Y., and Zeger, S.L.(1996). *Analysis of Longitudinal Data*. Oxford University Press: New York.

Hanushek, E.A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *The Journal of Human Resources*, 14, 3, 351-388.

Henry Braun, Naomi Chudowsky, and Judith Koenig (*Editors)* (2010). *Getting Value Out of Value-Added; Report of a Workshop***,** National Academic Press, 96 pages (available at http://www.nap.edu/catalog/12820.html)

Journal of Educational and Behavioral Statistics (2004). *Value-Added Assessment: Special Issue*, No. 1, Vol. 29.

Mariano et al. (2010). General persistence models, *Journal of Educational and Behavioral Statistics*, 35, 253-279.

McCaffrey, D.F., Lockwood,J.R., Koretz, D., Louis, T.A., and Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statisitics*, 29, pp. 67-101.

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

Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods.* 2nd Ed. Sage publications Inc., California: Thousand Oaks

Raudenbush, S.W. and Bryk, A.S. (1988). Methodological advances in analysing effects of schools and classrooms on student learning. *Review of Research in Education*, 15, 423-479.

Rowan, B., Correnti, R., Miller, R.J. (2002). What large scale survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record*, 104, 1525-1567.

Sanders, W.L., Saxton, A., and Horn, B.(1997). The Tennessee Value-Added Assessment System: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is student achievement a valid evaluation measure?* , Thousand Oaks, CA: Corwin Press, Inc. pp. 137-162.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Boca Raton, FL: Chapman and Hall/CRC

Appendix I (Review of BLUP Estimation and their MSE)

Using the general notation of linear mixed models (Skrondal and Rabe-Hesketh, 2004), consider best linear unbiased predictor (BLUP) estimation of a realized value of the random effect vector $\boldsymbol{\eta}_{q\times 1}$ and best linear unbiased estimation (BLUE) of fixed parameter $\boldsymbol{\beta}_{p\times 1}$under the model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\eta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y}$ is the vector of student test scores, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are covariate matrices corresponding to observed and unobserved covariates , $\boldsymbol{\beta}$ is a *p*-vector of fixed regression parameters, $\boldsymbol{\eta}$ is a *q*-vector of random effects with mean **0** and covariance matrix $\boldsymbol{\Gamma}_\eta$, $\boldsymbol{\varepsilon}$ has mean **0** and covariance matrix $\boldsymbol{V}$ and is independent of $\boldsymbol{\eta}$, and both have typically normal distributions although not necessarily required for BLUP estimation.

The BLUP of $\boldsymbol{\eta}$ is given by

$$\begin{aligned}\widehat{\boldsymbol{\eta}}_{BLUP} &= \left(\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^{-1} + \boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{BLUE})\\ &= \boldsymbol{\Gamma}_\eta \boldsymbol{Z}'\boldsymbol{W}^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{BLUE})\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_{BLUE}$ = $(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{y}$, $\boldsymbol{W} = \boldsymbol{V} + \boldsymbol{Z}\boldsymbol{\Gamma}_\eta\boldsymbol{Z}'$.

Moreover, the mean square error (MSE) of estimates about their true values are given by

MSE $(\widehat{\boldsymbol{\beta}}_{BLUE} - \boldsymbol{\beta})$ = $(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}$ **,** and

MSE $(\widehat{\boldsymbol{\eta}}_{BLUP} - \boldsymbol{\eta}) = \left(\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^{-1} + \boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z}\right)^{-1} + \boldsymbol{\Gamma}_\eta\boldsymbol{Z}'\boldsymbol{W}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{Z}\boldsymbol{\Gamma}_\eta$

where the leading or the first term $\left(\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^{-1} + \boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z}\right)^{-1}$ can also be expressed alternatively as

$$(\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z})^{-1}\left(\boldsymbol{\Gamma}_\eta + (\boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z})^{-1}\right)^{-1}\boldsymbol{\Gamma}_\eta \quad \text{or} \quad (\boldsymbol{\Gamma}_\eta - \boldsymbol{\Gamma}_\eta\boldsymbol{Z}'\boldsymbol{W}^{-1}\boldsymbol{Z}\boldsymbol{\Gamma}_\eta).$$

As expected, it is seen that $\boldsymbol{\beta}$ can be consistently estimated if the total number of observations (or the number of rows of $\boldsymbol{X}$**)** is much larger than *p* (number of $\boldsymbol{\beta}$-parameters or the number of columns of $\boldsymbol{X}$) or more precisely when the eigenvalues of its MSE matrix or $(\boldsymbol{X}'\boldsymbol{W}^{-1}\boldsymbol{X})^{-1}$ can be quite small as measured by the sum of eigenvalues or trace of the matrix. However, this will not be the case with $\widehat{\boldsymbol{\eta}}_{BLUP}$ because the total number of observations is not expected to be much larger than *q*—the number of $\boldsymbol{\eta}$-parameters or the number of columns of $\boldsymbol{Z}$; i.e., looking at the leading or the first term of its MSE , the eigenvalues of $\left(\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^{-1} + \boldsymbol{Z}'\boldsymbol{V}^{-1}\boldsymbol{Z}\right)^{-1}$ are not expected to be very small. The second term in MSE $(\widehat{\boldsymbol{\eta}}_{BLUP} - \boldsymbol{\eta})$ reflects the additional variability due to the estimation of $\boldsymbol{\beta}$ and is of much smaller order than the first term. If the rows of $\boldsymbol{Z}$ could be increased by including more observations corresponding to each random effect, then the corresponding BLUP estimators would naturally be more stable. The stability of the BLUP MSE matrix can be measured by its trace (i.e., the sum of its diagonal elements which is the same as the sum of eigenvalues) and the condition number defined as the ratio of maximum and minimum eigenvalues.