

DRAFT November 3, 2008

Analytic Issues with VAM

Discussion of papers

by Dale Ballou and by Daniel McCaffrey and J.R. Lockwood

**Helen F. Ladd
Sanford Institute of Public Policy
Duke University
Durham, NC 27708
Hladd@duke.edu**

These comments were prepared for a workshop sponsored by the U.S. National Research Council and the US National Academy of Education on the use of value-added methods for instructional improvement, program evaluation and accountability, November 13 and 14, Washington, D.C. These comments draw heavily on Helen F. Ladd, “Teacher Effects: What Do We Know?”, a paper prepared for the Teacher Quality Conference at Northwestern University, May 1, 2008.

The two papers by Daniel McCaffrey and J.R. Lockwood and Dale Ballou summarize a number of analytical issues that arise in the estimation of value added models (VAM). McCaffrey and Lockwood (henceforth ML) define VAM as the use of longitudinal data to measure the effects on student achievement of students' current teachers and schools, separate from other inputs. In addition, they introduce the two categories of value added modeling identified by Douglass Harris (2008), namely VAM-A for accountability purposes and VAM-P for program evaluation. Much of their paper, however, deals with the measurement of teacher effects. Ballou provides no clear definition of value added modeling, but implicitly assumes that the goal of VAM (at least for the purposes of his paper) is to isolate the contributions of teachers to the learning of their students. Consistent with this focus of the two papers, most of my comments refer to the use of value added modeling specifically for determining the contribution of individual teachers to student achievement.

Let me begin by saying that the NAS committee has done well to select these three authors to discuss the analytical issues of VAM. As noted by ML, both economists and statisticians have contributed to this type of value added modeling. Ballou comes out of the economics tradition, but has also done work on this topic with William Sanders from the statistics tradition. McCaffrey and Lockwood are both statisticians, but as is evident from their discussion of the two approaches, they are deeply familiar with the strengths and weaknesses of both.

Ballou's paper is a straightforward discussion of specification and estimation issues, primarily from the perspective of an economist. Not surprisingly, he begins his paper by highlighting the issue of most concern to economists – namely the non-random matching of students and teachers, followed by a discussion of specification issues. ML covers similar issues but in a somewhat broader context and with more attention to specification issues. Together the two papers are quite comprehensive in the analytical issues they cover. It's easy to agree with the ML conclusion that statisticians need to pay attention more to the selection problems identified by economists and the economists need to relax some of the key assumptions of their models.

My comments are divided into two parts. First, I present a simplified model to highlight some of the assumptions underlying value added models, with particular attention to the issues included in the two papers. Second, I briefly raise a few additional practical and policy issues, some of which have been touched upon in both of these papers, and are also discussed in other in other papers for this conference. I spend no time on another issue raised in both papers, namely the question of test scaling, with the hope that my fellow discussant will comment on that topic.

Can teacher-specific effects be identified and measured?

As noted by ML, researchers have been using two main approaches to identify the effectiveness of individual teachers in raising student achievement. I refer to the first approach as value added modeling and include in that category both levels and gains models. The second approach, which ML refer to as the statistical approach, includes

mixed and layered models that directly model the full joint distribution of all student outcomes. Though for some purposes the mixed methods models are superior, they are computationally very demanding.

Value-Added Models

As is clear from both papers, and particularly from Ballou, a fundamental challenge in estimating teacher effects is the observation that teachers are not randomly assigned to teachers. For the moment, I set this issue aside to develop the conceptual foundations of the standard value added model, with particular attention to the assumptions underlying it.

Derivation of simple value added model. The starting point for this analysis is the observation that education is a cumulative process. In the context of a very simple model in which the only educational input that matters is teachers and in which all other determinants of achievement such as student background, ability, and motivation are set aside, we can write

$$A_{it} = f(T_{it}, T_{i,t-1}, \dots) + \varepsilon_{it} \quad (1)$$

where A_{it} refers to student i 's achievement, as measured by test scores, in year t , T_{it} refers to some measure of the quality of student i 's teacher in year t , and ε_{it} is an error term. This equation expresses student i 's achievement in year t as a function of her teacher in that year and in all previous school years plus a random error.

Two additional assumptions permit this relationship to be transformed into one that can be used to estimate the effect of the student's teacher in year t on the student's achievement in that same year, controlling for the effects of teacher quality in all prior years. One assumption is that the relationship is linear and that the teacher quality measure has a constant marginal impact on student achievement. The second is that student achievement, or knowledge, decays from one year to the next at a constant rate. As a result, the rate at which a student's knowledge persists from one year to the next is also constant. Letting β be the effect of T and α the rate at which knowledge persists, we can rewrite equation 1 as

$$A_{it} = \beta T_{it} + \alpha \beta T_{i,t-1} + \alpha^2 \beta T_{i,t-2} + \alpha^3 \beta T_{i,t-3} + \dots + \varepsilon_{it} \quad (2)$$

and, after rearranging terms, as

$$A_{it} = \beta T_{it} + \alpha (\beta T_{i,t-1} + \alpha \beta T_{i,t-2} + \alpha^2 \beta T_{i,t-3} + \dots) + \varepsilon_{it} \quad (3)$$

Noting that the expression within the parentheses is simply $A_{i,t-1}$ and changing the order of the terms, we end up with

$$A_{it} = \alpha A_{i,t-1} + \beta T_{it} + u_{it}, \quad (4)$$

where the error term, u_{it} , is equal to $\varepsilon_{it} - \alpha \varepsilon_{i,t-1}$

Thus, the effects on current achievement of the student's prior teachers are captured by the lagged achievement term. If a student's knowledge does not persist from year to year the persistence parameter, α , would be zero.

Models of this form (but with additional explanatory variables as discussed below) are typically referred to as value-added models and are commonly used in the literature to estimate β , namely the effect of current teachers on current achievement. The popularity of such models comes largely from their simplicity and intuitive appeal. Logically, it makes sense to control statistically for the achievement, or knowledge, that the student brings to the classroom at the beginning of the year when estimating the effect of her current teacher. In addition, the value-added model is flexible in that it does not impose a specific assumption about the rate at which knowledge persists over time; instead it allows that rate to be estimated. Nonetheless, the model is valid only if the underlying assumptions about the constancy of effects are valid. Further, as pointed out by Ballou, such models raise statistical concerns because of the inclusion on the right hand side of the equation of the lagged achievement term, which in the presence of serial correlation would be correlated with the error term.

Gains model. This last statistical problem can be avoided by assuming there is no decay in knowledge so that the persistent parameter, α , equals 1 and moving the lagged achievement term to the left hand side of the equation. This procedure generates the gains model:

$$A_{it} - A_{i,t-1} = \beta T_{it} + \varepsilon_{it}. \quad (5)$$

In this case, the parameter, β , refers to the effect of a student's teacher on her gain in achievement. If the assumptions underlying the initial value-added model are correct, however, and the decay rate is not zero, the gains model is incorrectly specified. The reason is that the term $(\alpha-1)A_{i,t-1}$ is now missing from the right hand side of the equation. To the extent that prior year achievement is positively correlated with teacher effects, the estimated teacher effects would be biased downward. Thus, within the framework of education as a cumulative process, the shift to the gains model solves one statistical problem but introduces a new one.

Full value added (or gains) model with student fixed effects. In fact, most researchers estimate a richer form of the simple model in equation 1, one that includes time-varying student characteristics, classroom or school characteristics, and student fixed effects. As noted by Ballou, this full model can only be estimated with longitudinal data on individual students and multiple cohorts of students. If data are available for only a single cohort of students, no classroom characteristics such as class size or the composition of the students, can be included in the equation because teachers and their classrooms are indistinguishable.

$$A_{it} = \alpha A_{i,t-1} + \beta T_{it} + \gamma X_{it} + \delta C_{it} + \theta_i + \eta_{it} \quad (6)$$

where A_{it} , $A_{i,t-1}$ and T are as defined above and
 X_{it} are time varying student variables
 C_{it} are classroom and school characteristics in year t

θ_i are student fixed effects
 η_{it} is an error term

To be consistent with the cumulative model of the education process, this model requires the same assumptions underlying the simple value added model in equation 4..

The student fixed effects are a crucial part of this enriched model. They control for the time-invariant characteristics of students – both those that are measurable and those that are not -- and under certain assumptions address the fundamental problem highlighted by Ballou, namely that the teachers are not randomly assigned to students. The inclusion of student fixed effects means that the teacher effects are derived from the within-student variation in student achievement. The key assumption needed for student fixed effects to address fully the concern about nonrandom sorting is that students are assigned to teachers based on their permanent or average characteristics rather than on any time-varying unmeasurable characteristics. Most value added studies of teacher effects either implicitly or explicitly make this assumption. I return below to Jesse Rothstein’s recent challenge to the validity of this assumption.

In the context of these models, the teacher variables are typically entered as 0-1 indicator variables, either for each teacher or for each teacher by year. Thus teacher effects are estimated by the method of teacher fixed effects (in contrast to the method of random effects), an approach that seems reasonable given the goal of determining the effectiveness of a group of specific teachers.

Measurement error. Note that teacher-by-year fixed effects are identified by the number of students taught by the teacher in that year, a number that could well be quite small, especially at the elementary school level. Even when teacher effects are based on multiple years of data, the number of students taught will differ across teachers, which means that the coefficients of the teacher indicator variables are estimated with different degrees of precision. Had they been estimated by random effects rather than by fixed effects, estimates for individual teachers would have been shrunken toward the mean, with the amount of shrinkage greater for the teacher effects that are estimated with less precision. Letting β_t^* represent the predicted teacher effect for teacher t that emerges from a fixed effects specification, β_t the true value and ε a random error, we can express the predicted teacher effect that emerges from a fixed effect specification as a function of the true effect plus an error term as follows:

$$\beta_t^* = \beta_t + \varepsilon \quad (7)$$

One might then calculate the true teacher effect for any given teacher as a weighted average of the estimated teacher effect for that teacher and the mean teacher effect for the sample as a whole:

$$\lambda \beta_t^* + (1-\lambda) \text{mean } B_t^* \quad (8)$$

where $\lambda = \text{Var} \beta_t / (\text{Var} \beta_t + \text{Var} \varepsilon)$.

Thus, the larger is the random error of the estimate, the smaller is λ and the greater is the weight placed on the mean teacher effect. Though such an adjustment is conceptually straightforward, estimating λ directly from the variances can be tricky to implement in practice. One implication of this shrinkage procedure is that teachers who teach small numbers of students are unlikely to be identified as either particularly effective or particularly ineffective teachers. Although the outcome on the low side may be appropriate since it would protect decent teachers with small classes from being unjustly sanctioned, the shrinkage procedure could also keep some very effective teachers from being recognized.

Additional considerations. Though much more could be said about this standard value added (or gains model), I add here only two additional considerations. The first refers to the role of parents. As pointed out by Todd and Wolpin (2003) compensating behavior by parents could potentially mute the estimated differences in teacher effectiveness. That outcome would occur if parents spend more productive time working on school work with their children when their children have ineffective teachers than when they have effective teachers.

Another is whether to include school fixed effects in the model. Often they are not included, particularly if student fixed effects are in the model, as in equation 6. In the absence of student fixed effects, the addition of school fixed effects, along with a rich set of student level control variables, can help mitigate the problem caused by the non-random assignment of teachers to students. Their inclusion in the model means, however, that teacher effects are identified solely by differences in teacher quality within schools and thereby changes the question being asked. As correctly noted by Ballou, the use of school fixed effects is just one example of how the introduction of various types of fixed effects – introduced as solutions to particular statistical problems -- constrains the analysis.

How stable are teacher effects? In most cases one would expect that a teacher who is very effective (or ineffective) in one year would be similarly effective (or ineffective) in the following year. Hence, one way to evaluate the validity of the teacher effects that emerge from value added models would be to examine their stability from one year to the next. The more unstable they are they less useful they are likely to be for making high stakes decisions about teachers.

Only a few studies have explored the stability of teacher effects (Ballou, 2005, Aaronson et al (2007) and Koedel and Betts (2007)). In all cases, the studies find that teacher effects are quite unstable. The most complete study of the stability of teacher effects is by Lockwood, McCaffrey and Sass (2008) for Florida. Because this issue is discussed not only in ML but also in other sessions, I will not pursue it further, except to highlight that instability is a big problem.

The Rothstein challenge. Both Ballou and ML correctly highlight Jesse Rothstein's recent challenge to the value-added approach (Rothstein 2008). It is well known that the use of student fixed effects in longitudinal models solves the problem of

non-random matching of students to teachers only when such matching is based on the time invariant characteristics of the students, such as their basic ability or motivation. Rothstein refers to such matching as “static tracking” and contrasts it to the “dynamic tracking” that occurs when school administrators sort students into classrooms and teachers in a non-random way in based in part on the student’s current performance. As described by ML, Rothstein uses North Carolina data to provide evidence of dynamic tracking by showing that a student’s fifth grade teachers appear to affect the student’s fourth grade learning in reading.

I agree with Ballou and ML that Rothstein’s evidence on dynamic tracking represents a serious challenge to the validity of the standard value approach. The argument on the face of it seems quite compelling. At the same time, it suggests that teacher effects may be largely spurious, which conflicts with the conclusion from a range of other studies showing that teachers matter (Hanushek, Kain and Rivkin 2005; Rowan, Correnti and Miller, 2002; and Nye, Konstanopoulos and Hedges, 2004). Hence additional research on the validity of the static tracking model is clearly needed. A first step would be to reestimate the Rothstein models with multiple cohorts, and to examine results for math in addition to reading. The use of multiple cohorts would permit the researcher to separate teacher effects from contextual effects, which as discussed below, have emerged as a cause of concern with respect to the estimation of teacher effects in more complex models. Although Rothstein believes that the use of multiple cohorts will not change the results (personal communication with the author, April 2008), it would be useful to have that confirmed empirically. A second step would be to explore the student-assignment process used by school principals. Some preliminary informal investigation by this author in a few North Carolina elementary schools provides little support for the hypothesis of dynamic tracking in some schools, but the observations were limited. Clearly more investigation is needed. In addition, it might be productive to remove the school fixed effects from the Rothstein’s equation to estimate teacher effects relative to teachers throughout the district rather than to those within each school.

Mixed Methods or Layered Models (Multivariate Modeling)

These models are far more complicated than the simple value added models in that they specify a joint distribution for the entire multivariate vector of test scores. Included among these models are the Tennessee Value Added Assessment System (TV AAS) developed by William Sanders for Tennessee, the cross classified models of Rowan, Correnti, and Miller (2002) and Raudenbush and Bryk (2002) and the persistence models of McCaffrey et al. (2003).

The key element of such models is that a student’s performance in any year is modeled not only as a function of her teacher in the current year, but also of her teachers in prior years. Moreover, in such models teacher effects are typically estimated using random rather than fixed effects. A major advantage of multivariate models relative to the simpler value added models with fixed effects is that the models use more information to identify teacher effects. In particular they make use of the fact that student scores in later years hold information about the effectiveness of teachers in the past. Another advantage is that they are very flexible. The primary disadvantage of such models is their

tremendous computational demands. Until computational methods are developed to make it easier to estimate such models, it is likely that the more standard value-added models will be the basis of much of the ongoing research in this field.

I focus here on the TV AAS layered model because it has received significant attention in the literature. Implicit in this specific model is the assumption that any teacher effects in a prior year persist undiminished in future years. No student covariates are included. Instead, the complex correlations among the errors from the repeated test scores substitute for student specific covariates.

Kupermintz (2003) provides some useful insights into the TV AAS model. First he notes that the resulting estimates rank teachers within each school system. Hence, a weak teacher in a system with many other weak teachers may receive a more favorable ranking than a similar teacher in a stronger system. Second, the teacher effects are “shrunk” towards the system average for reasons similar to those discussed above. Thus, once again, it is difficult to get accurate estimates of the effectiveness of teachers who are working with small numbers of students. In addition, and perhaps most significantly, Kupermintz questions the validity of the estimated teacher effects given that they emerge from a model that includes no student level or classroom level covariates. Though he acknowledges that the model uses prior achievement as a covariate or “blocking variable,” which means that each child serves as his or her own control, he notes that such “blocking” procedures were developed in the context of controlled experiments not in the context of observational studies. In contrast to controlled experiments in which treatments can be randomly assigned, students are not randomly assigned to teachers (Kupermintz, 2003, p. 292). As a result, the estimated teacher effects may be confounded by the effects of correlated student level characteristics that are omitted from the model. Further, he argues that for the TV AAS procedure to be valid, the prior year achievement variable would have to serve as a proxy for a variety of contextual factors including, for example, the socio-economic or achievement mix of students in the classroom.

The extent to which the absence of covariates, at either the student level or at the classroom level, distorts the results has been examined at by Lockwood and McCaffrey (2007) in the context of a general multivariate model (also see McCaffrey et al, 2004). Despite concerns that the use of random effects models can lead to inconsistent estimates when unobserved individual effects are correlated with other variables in the model, Lockwood and McCaffrey (2007) demonstrate through analysis and simulation that the mixed method approach does not generate much bias in practical applications, especially when the number of tests scores or individual students is relatively large. The authors’ simulations support the claim of William Sanders that the joint estimation of multiple test scores for individual students, along with other elements of the TV AAS approach, effectively purges the results of any bias that would otherwise arise as a result of the variation in student backgrounds (Lockwood and McCaffrey, 2007, p. 244.) At the same time, however, the mixed methods approach cannot control for bias when the student population is stratified. A stratified student population is one “in which there are disjoint groups of students such that students within a group share teachers but students in different groups never share any teachers ” (McCaffrey and Lockwood, 2007, p. 245).

Ballou, Sanders and Wright (2004) reinforce these conclusions empirically in the context of the TV AAS model. To examine the effects of student level covariates, the authors add them to the TV AAS model in a two- stage approach. They begin with a first-stage equation in which student achievement gains are estimated as a function of student characteristics and standard teacher fixed effects (not the teacher fixed effects that emerge from the TV AAS model). The inclusion of the teacher fixed effects ensures that the estimated coefficients of the student characteristics are uncorrelated with any time invariant component of teacher quality. They then use the estimated coefficients of the student characteristics to adjust the gain scores for each student and rerun the TV AAS model with the adjusted student gain scores. Consistent with the findings of Lockwood and McCaffrey (2007), the authors conclude that the use of the adjusted gain scores does not significantly change the estimates of teacher effects and hence that the unadjusted TV AAS model does an acceptable job of controlling for student-level covariates.

The results differ, however, when Ballou, Sanders and Wright (2004) make similar adjustments for contextual factors (such as the percent of students in a grade or school eligible for free and reduced price lunches). In that case, the TV AAS results change significantly, are implausibly large in some grades, and are sensitive to minor changes in model specification. Thus, consistent with the findings of Lockwood and McCaffrey, they conclude that the stratification of students across schools renders the TV AAS model problematic.

Practical and Policy Issues Related to the Use of VAM

I assume that Ballou and ML were specifically asked to focus primarily on the analytical issues alone, and to leave for others papers discussion of the practical and policy implications. In that spirit, I will make just a few quick points about some of these other aspects of value added modeling, while sticking generally to the analytical issues that are the subject of this session.

Not mentioned in either of the papers is the purely practical challenge of making sure that each student is matched to her correct teacher for the specific subject. Though this issue is mundane, it is crucial for this type of analysis. Currently most state data systems do not provide direct information on which students are taught by which teachers. Until recently, for example, those of us using the North Carolina data have had to make inferences about a student's teacher from the identity of the proctor of the relevant test and a wealth of other information from school activity reports. In my own work, I have been able to match between 60-80 percent of students to their teachers at the elementary and high school levels but far lower percentages at the middle school level (Clotfelter, Ladd and Vigdor, 2007 and 2008). Moreover, even if states start providing more complete data of this type, a number of conceptual issues still complicate the situation. These include, for example, how to deal with students who are pulled out of their regular classes for part of the day, how to deal with team-taught courses, and how to deal with students who transfer into or out of a class in the middle of the year. For the

purposes of research, a 60-80 percent match rate may suffice. If value added models are used as part of a teacher evaluation system, however, it may well not suffice.

More generally, what is acceptable for research on teacher effects or for program evaluation (VAM-P) may differ from what is useful for a teacher accountability and incentive system (VAM-A). For example, for the purposes of research, it may be useful to include school fixed effects in a value added model in order to separate the effects of teachers from other school level characteristics, including the quality of school leadership. From the perspective of a district or state school administrator, in contrast, the measures of teacher effectiveness that emerge from such a model which, in effect, compare teachers to other teachers within the same school, are likely to be far less useful than measures that compare teachers to other teachers throughout the district or state. Similarly, in order to isolate the effects of a teacher from the context of her classroom, analysts will need to pool data for individual teachers over time, thereby ruling out the estimation of the year-specific measures for each teacher that would be desirable for a teacher improvement program. Further, the methods of dealing with measurement error, including the shrinkage strategy mentioned above, imply that the value added modeling system provides less information on the effectiveness of some teachers than of others. Finally, although the instability of estimated teacher effects may not matter for some purposes, it is likely to be unacceptable as the basis for an accountability program.

A related issue is the tradeoff between analytical complexity vs. transparency, an issue raised by ML. Once again the purpose of the value added modeling is important. If the purpose is research or program evaluation, it may be appropriate to get the best measure possible, even if it is analytically complex. In contrast, if the value added modeling is to be used for the purposes of accountability and incentive programs transparency may become far more important. Assuming a goal of such a program is to change teachers' behavior, teachers need to understand what goes into the outcome measure, what they can do to change the outcome, and to have confidence that the measure is consistently and fairly calculated. Though these considerations push in the direction of simplicity and transparency, an offsetting consideration is that the system is likely to be most effective if teachers believe the measure treats them fairly in the sense of holding them accountable or rewarding them for things that are under their control. To achieve that end some analytical complexity is often needed.

In light of these countervailing pressures, it would be useful to have more research into the various hybrid or second best systems that might be designed to promote both goals. I am reminded here of how the state of South Carolina had to modify its first gains-based school accountability system back in the early 1980s. State policy makers were committed to measuring school effectiveness with achievement gains on the ground that the use of gains would represent a big improvement over a levels approach, while at the same time being transparent. Once state policy makers understood that the approach was biased toward schools serving more advantaged students, however, they divided schools into divisions so that could continue to use the transparent approach but at the same time let schools compete for bonuses with schools similar to themselves.

To that end, more experimentation and evaluation of pilot projects based on value added models that balance practical and political concerns with statistical considerations would undoubtedly be useful. At the same time, the types of analytical concerns discussed by Ballou and ML call for great caution with respect to the use of teacher-specific value added measures for high stakes decisions about teachers.

References.

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics* 25:95-135.

Ballou, Dale. 2005. "Value-added assessment: Lessons from Tennessee." In R. Lissetz (ed.) *Value Added Models in Education: Theory and Applications*. Maple Grove, MN: JAM Press.

Ballou, Dale, William Sanders and Paul Wright. 2004. "Controlling for Student Background in Value-Added Assessment of Teachers." *Journal of Educational and Behavioral Statistics*, 29 (1), pp. 37-66.

Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L. (2006). "Teacher-student matching and the assessment of teacher effectiveness." *Journal of Human Resources*, XLI (4), 778-820.

Clotfelter, C. T., H. F. Ladd, and J. L. Vigdor. (2007a). "Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects." *Economics of Education Review*. December.

Clotfelter, C.T., H.F. Ladd, & J.L. Vigdor (2007b) "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects." National Bureau of Economic Research Working Paper 13617. Also available on the CALDER web site (Caldercenter. Org) .

Hanushek, Eric A., John F. Kain, Daniel M. O'Brien, Steven G. Rivkin. 2005. "The Market for Teacher Quality." National Bureau of Economic Research Working Paper 11154.

Harris, D. N. (forthcoming). "The policy uses and "policy validity" of value-added and other teacher quality measures." In D.H. Gitomer (ed), *Measurement issues and the assessment of teacher quality*. Thousand Oaks, CA: SAGE Publications.

Koedel, Cory and Julian R. Betts. 2007. "Re-examining the Role of Teacher Quality in the Educational Production Function." Working Paper #2007-03. Nashville, TN: National Center on Performance Initiatives.

Kupermintz, Haggai. 2003. "Teacher Effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System," *Educational Evaluation and Policy Analysis*. Vol 25, no. 3 (Fall), pp. 287-298.

Lockwood, J.R. and Daniel F. McCaffrey. 2007. "Controlling for individual heterogeneity in longitudinal models, with applications to student achievement," *Electronic Journal of Statistics*. Vol 1, pp. 223-252.

Lockwood, J.R., Daniel F. McCaffrey, and Tim R. Sass. 2008. "The Intertemporal Stability of Teacher Effect Estimates." Paper prepared for the Value Added Conference at the University of Wisconsin, April 23 and 24, 2008.

Nye, Barbara, Spyros Konstantopoulos, and Larry Hedges. 2004. "How Large Are Teacher Effects?," *Educational Evaluation and Policy Analysis*. Vol 26, no. 3, pp. 237-257.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement". *Econometrica*, Vol 73, no. 2 (March), pp. 417-458.

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

Wright, S.P., S. P. Horn and W.L. Sanders. 1997. "Teacher and classroom context effects on student achievement: Implications for teacher evaluation." *Journal of Personnel Evaluation in Education*. 11, pp. 57-67.