Value-Added Assessment: Special Issue. *Journal of Educational and Behavioral Statistics*: of Teacher and School Performance, Spring 2004, vol. 29(1). Selected portions of the journal may be made available online at http://www.aera.net/pubs/jebs. For a free single copy of this issue, contact Tiffany Cain, NEA Research, Ext. 7430 (please note – this journal is highly technical).

Marcella Dianda, Student Achievement, and Denise McKeon, NEA Research, 5/2004

NEA in the Know is a series of occasional briefs published by NEA Research designed to provide NEA members and leaders with analyses of emerging research.

As states and districts wrestle with ways to hold teachers and schools accountable, many are looking at value-added assessment as a way of calculating the effect that schools and teachers have on student achievement – as measured by standardized test scores. Some statisticians and policymakers have asserted that value-added models are easily explained, understood, and defended against criticism. But, the papers on value-added models in the Spring 2004 special issue of the *Journal of Educational and Behavioral Statistics* make it clear that it may be impossible to find value-added models meeting these criteria. The statisticians writing in this special journal discuss the technical characteristics of different value-added models, focus on unresolved technical problems with the models, and even ask whether value-added models are an appropriate accountability tool. (For additional background information on value-added models, see two other *NEA in the Know* briefs – Daniel F. McCaffrey, J.R. Lockwood, Daniel M. Koretz, and Laura A. Hamilton. (2003). Evaluating Value-Added Models for Teacher Accountability http://connect.nea.org/edstats/ITKValue-Added.html and H. Kupermintz (Fall, 2003). Teacher Effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System http://connect.nea.org/edstats/tqTVAAS.html.

What the Research Examines

In this special issue of the Journal of Educational and Behavioral Statistics, several statisticians from education and related fields raise – and try to resolve – some of the

technical issues associated with using value-added models to measure teacher and school performance. The journal includes three papers and discussants' comments on the papers. In addition, in brief statements, a principal from Pennsylvania and a state legislator from Ohio explain why they embrace the use of value-added models in education.

What the Research Says

Paper #1. ***Controlling for Student Background in Value-Added Assessment of Teachers*** by Dale Ballou, William Sanders, and Paul Wright

Focus. Arguably the centerpiece of this special issue from the Association's perspective, this paper by Dale Ballou, Vanderbilt University, and William Sanders and Paul Wright, SAS Institute, addresses a key criticism of Sanders' value-added assessment model – its failure to control for socioeconomic status and demographic factors that affect student achievement. Sanders argues that, because his method measures achievement gains from a students' own starting point, it implicitly controls for socioeconomic status and other background characteristics.

Members of the research and policy community have raised questions about how Sanders' model can accurately measure the contributions of teachers and schools without controlling for these contextual variables, which the point out influence students' initial achievement scores on tests as well as the rate at which students learn.

In this paper, the authors add student background variables to the Tennessee Value-Added Assessment System (TVAAS) model. First, they predict gains in students' test scores as a function of individual students' eligibility for free or reduced-price lunch, and students' race and gender. Second, they predict gains in test scores as a function of two school-level factors – the percentage of students who are eligible for free or reduced-price lunch in a student's grade level and in the entire school. While these variables may seem meager, Ballou and his colleagues argue they are most likely to be available to a school district without the cost of additional data collection.

Findings. The authors report that including socioeconomic and demographic variables at the student level had little effect on the TVAAS. Correlations between initial teacher effects and those obtained after the introduction of these controls exceeded .9, using test data over a three-year period, in grades 3-8 in reading, math, and language arts. In addition, the adjusted and unadjusted TVAAS models agreed far more often than they disagreed on the identification of teachers who were significantly above or below average. In fact, the authors report that controlling for student-level factors had only a moderate impact on estimated teacher effects, even for teachers whose classes were entirely poor or entirely minority.

In contrast, however, controlling for socioeconomic status at the grade and school levels had a substantial impact on TVAAS estimates in some grades and for some subjects, but the researchers back away from this finding suggesting that the results are unreliable and additional study is needed.

Conclusions. Not surprisingly, Ballou, Sanders and Wright conclude that critics' concerns that the TVAAS model does not adequately account for socioeconomic and demographic factors that affect student achievement are unfounded. They conclude that in TVAAS, the longitudinal history of a student's performance serves as a substitute for

student socioeconomic status and demographic controls. They note, however, that critics' concerns may apply when "less sophisticated models are used to estimate teacher effects." That is, the authors maintain that it makes considerable difference whether simple models, such as the one Tekwe and her colleagues examine in Paper #2 in this special journal issue, include student-level controls for socioeconomic status and demographics.

But Ballou and his colleagues admit that applying TVAAS methodology to schools may be problematic. In fact, they say "we cannot be confident that the TVAAS controls for contextual variables [i.e., demographic and socioeconomic variables at the school level] in the same way that it controls for the influence of student-level SES and demographics."

Paper #2: ***An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance*** by Carmen D. Tekwe, Randy L. Carter, Chang-Xing Ma, James Algina, Mel Lucas, Jeffrey Roth, Mario Ariet, Thomas Fisher, and Michael B. Resnick.

Focus. In this paper, Carmen Tekwe, a statistician from the Center on Aging and Health at Johns Hopkins University, and eight colleagues examine whether different value-added models provide different results. Their special interest is whether models that use simple value-added measures are "just as good" as models that are based on more complex measures, such as TVAAS.

Simple value-added measures, which compare changes in the average achievement of students in a school to changes in average student achievement in the district, are easy to calculate and understand. In contrast, more complex models, which may or may not include "covariates" (i.e., students' socioeconomic status; last year's test score) are complex and produce value-added measures that are not easily understood.

The researchers report that this is the first comparative study of four prominent value-added models using a common set of student data (i.e., test scores students in grades 3-5 in a medium size Florida school district over two years).

Findings. Tekwe and her colleagues found that the simplest value-added model produced results that were highly correlated (.91) with results from models that used more complex value-added measures. And, in a second finding, they report that the three complex models they studied, which included TVAAS, all produced different value-added measures. This was expected because the more complex models improve on the simple model in different ways; hence, different results.

Conclusions. Surprisingly the authors conclude that the choice among value-added models must not be based primarily on empirical or statistical considerations, and they do *not* recommend the simple value-added model over more complex ones. They suggest, instead, that the major consideration in model selection is a policy question that is at the heart of the debate around using value-added assessment for teacher and school

accountability: Should schools be held accountable for significant effects of student background characteristics?

Whether variables like socioeconomic status and race/ethnicity are included or excluded, they affect the value-added estimate. The authors argue that if they are included, in effect, schools are excused from responsibility for their effect on achievement. But schools, they argue, are probably partially, if not wholly, responsible for ameliorating the effects of SES and key student background factors, so they should not be totally excused. On the other hand, neither should they be totally responsible.

As a practical matter, the researchers report that when they excluded variables that tapped student and school-level socioeconomic and demographic characteristics from the models, the resulting value-added measures were biased against schools with an over representation of poor and minority students. But including these variables resulted in value-added measures that were biased against schools with an under representation of poor and minority students.

Paper #3: ***Models for Value-Added Modeling of Teacher Effects*** by Daniel F. McCaffrey, J.R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton

Focus. The authors develop and apply a general value-added model to explore issues raised by using value-added modeling of student achievement to evaluate teachers and schools. All of the currently prominent value-added models, the authors say, can be seen as restricted cases of this general model. The general model, then, provides a framework for comparing current value-added models, including the TVAAS model. Like the other papers in this special issue, the authors also consider the amount of bias in estimates of teacher and school effects when student background characteristics are omitted from the models.

Findings. While Ballou, Sanders and Wright report in Paper #1 that adjusting for socio-demographic characteristics at the student level makes little different in value-added estimates of teacher effects, McCaffrey and his colleagues find that the issue is more complex. Their findings relate especially to the omission of socio-economic variables at the school level, an issue raised but not fully addressed by Ballou and his colleagues. In an example in the paper in which they apply their general model to charter schools serving heterogeneous students, they found that the average number of students at the schools who were eligible for free or reduced-price lunch was correlated with student test scores and with gains in test scores. In addition, free and reduced lunch eligibility at the school level predicted test scores even after the authors controlled for individual students' lunch status.

Conclusions. McCaffrey and his colleagues conclude that controlling for student-level socioeconomic and demographic factors alone will not be sufficient to remove the effects of background characteristics in all school systems, especially systems that serve heterogeneous students – a conclusion that cautions against the use of the Sanders' model for make estimates of the value teachers and schools add to gains in student test scores.

In diverse schools, "contextual effects," including the distribution of socio-demographic factors across students and the assignment of teachers to students, influence estimates of teacher and school effects.

The authors also note that separating contextual effects from teacher effects poses particular technical challenges. It may require collecting detailed information on teachers and teacher assignments that generally is not currently available. But such methods are important because heterogeneous populations are likely to be common in many school systems, especially large school systems.

Finally, the authors conclude that adequately addressing the complexities they found by applying their general model will require an extremely complex value-added model.

Discussants' Comments

The journal includes three discussions of the papers. As a group, they raise key questions about using value-added models to estimate teacher and school effects.

Review #1. **What Are Value-Added Models Estimating and What Does This Imply for Statistical Practice?** by Stephen Raudenbush

In his comments, Stephen Raudenbush from the University of Michigan asks about the kind of information that value-added models can really provide. At best, he says, they can provide information parents can use in selecting a school by identifying how their child might fare in different schools. That is, the models are best suited for estimating what Raudenbush labels as a "Type A" effect, a combination of school practice (including instruction and school management), the school environment (e.g., the neighborhood in which it is located), and school demographics.

In contrast, value-added models cannot provide information that holds teachers and schools accountable for their contributions to student achievement, which he labels as a "Type B" effect. Raudenbush argues that policymakers, who use value-added models to hold teachers and schools accountable, want to hold them accountable for their practice. But value-added models simply cannot isolate practice at the school or classroom levels from other factors—demographics and school environment. This issue was a prime focus of all three papers in this special journal issue.

Review #2. **The Real World is More Complicated than We Would Like** by Mark Reckase

Mark Reckase from Michigan State University raises an even more fundamental issue in his comments – the nature of the tests that value-added models use to derive estimates of teacher and school effects. All the models look at student test scores from grade to grade and use some measure of gain as an indicator of growth, but this may be too simple, says Reckase. He points out that the papers ignore a key fact -- what is tested shifts within across grade levels. In mathematics, for example, 3rd grade tests predominantly measure

arithmetic skills while by 8th grade the tests shift to problem solving, pre-algebra, and algebra skills. Reckase complains that the way the results are reported in the papers, implies that the tests are measuring the same thing when they are not.

This is an important point because a maximum change in test scores occurs if the pattern of instruction matches the shift in test content (within and across grades). If there is a mismatch between the shifts in test content and teachers' instruction, value-added estimates will be affected. Reckase concludes that the complex statistical procedures described in the papers in this special journal issue may be giving a glossy finish to misleading assessment results. He recommends taking a closer look at the tests before putting a lot of confidence in the results of value-added analyses.

Review #3. **A Potential Outcomes View of Value-Added Assessment in Education** by Donald B. Rubin, Elizabeth A. Stuart, and Elaine L. Zanutto

In their review, Donald Rubin and his colleagues remind us that value-added models are causal models. That is, a value-added estimate of a teacher's contribution to a student's test score is interpreted as follows: Johnny did well on a test because he had Mrs. Jones as a teacher. Whatever Mrs. Jones did (i.e., the value she added) caused his improved performance. Therefore, Mrs. Jones is a good teacher. And the opposite scenario also is true. If Johnny's test scores declined, Mrs. Jones is also the cause of the decline.

Rubin and his colleagues say that value-added models do not support causal inferences. It is even difficult to estimate causal effects even in randomized experiments, the 'gold standard' for making such estimates. And value-added models operate in the real world of schools where randomization is not possible. The authors conclude that the three papers in this special journal issue – and value-added models in general – cannot estimate causal effects at all.

They go on to ask, "If causal inference here is so difficult, how are we to guide policy and think about the benefits of value-added assessment?" Take the current value-added models at face value and consider them as descriptive measures, they suggest. Then, focus on the following question: "Do these descriptive measures, and proposed reward systems based on them, improve education?"

In addition, Rubin and his colleagues argue that the three papers in the journal do not focus sufficiently on a widely recognized problem when using longitudinal data: Missing test information (a real challenge in schools and districts where student mobility is high). All the papers restrict their analyses to students who have complete test data from year to year – an option that is not available in the real world of schools.

Another Perspective

The journal includes statements from an Ohio state legislator and a principal from Pennsylvania in which they explain why they support the use of value-added models in their states. The statements stand in stark contrast to the papers included in the journal

and to the reviews of those papers. The statisticians not only wrestle with the technical adequacy of value-added models, they also express concern about the validity of using these models to estimate teacher and school effects on student achievement. In contrast, this policymaker and school administrator express no such reservations.

They say that adding a value-added component to their accountability systems: 1) enables them to determine if every student is making adequate yearly progress as required by ESEA/NCLB; 2) allows parents, taxpayers, and educational decision makers to see more clearly whether schools are addressing each student's needs; 3) helps states and districts measure the effectiveness of teachers training and professional development; 4) helps schools that are below the state average to demonstrate gains in student achievement; 5) provides a "true measure" of an individual student's progress over time; and 6) eliminates such factors as socioeconomic status by looking at change in achievement.

And while the state legislator provides an unqualified endorsement of the value-added approach, the principal raises two issues that probably mirror concerns other educators have with these models. Echoing Reckase in his review, the principal expresses concern about the quality of the tests on which value-added models are grounded. "Are they good enough so that we can trust the change scores to mean something?" She also expresses concern that value-added estimates might be affected by high student mobility rates. The authors in this special journal issue do not address this issue, but it is one reason why individual students may have missing test data, and the issue of missing data is addressed by McCaffrey and his colleagues in a monograph reviewed in another issue of *NEA in the Know* (http://connect.nea.org/edstats/ITKValue-Added.html).

What the Research Means for NEA

This special issue of the *Journal of Educational and Behavioral Statistics* provides the Association with two important perspectives on value-added models – that of statisticians and that of policymakers/stakeholders. These two perspectives are rarely found in the same publication – both groups come at the issue of value-added methods from very different perspectives and with very different goals in mind.

First we have the perspective of statisticians who are debating among themselves about whether the models can, in fact, provide reliable, unbiased, and fair estimates of teacher and school effects on student achievement. In their paper, Ballou, Sanders and Wright clearly advocate this application based on their experience with the TVAAS model, but they are the only authors to do so.

Other paper authors and reviewers express strong reservations about using value-added models to estimate the effects of schools or teachers on student achievement. In fact, one of the reviewers concludes that based on the evidence presented in the papers in this journal, and other papers on value-added models, using these models in test-based accountability systems is probably "ill-advised at this time."

The use of value-added models is ill-advised because of the following key problems discussed the papers:

1. Given current statistical know-how, it is impossible to isolate the effects of teachers' and schools' practice from demographic and economic factors that also affect student achievement. All the papers agree on this point, even the paper co-authored by Bill Sanders, the developer of the TVAAS, which expresses reservations about using his model to estimate school effects.

2. When variables that tap student and school-level socioeconomic and demographic characteristics are excluded from the models, value-added estimates are biased against schools with an over representation of poor and minority students. But including these variables results in value-added measures that are biased against schools with an under representation of poor and minority students.

3. While value-added measures include complex statistical procedures, they are only as valid as the tests on which they are based. Those tests vary in quality and in the degree to which they match the instruction teachers and schools provide.

4. For a variety of technical and logical reasons, value-added models cannot work for the purpose for which they are intended – drawing causal inferences (i.e., that a student's 6th grade teacher cause increases in that student's test scores on the state test). Value-added models do not meet the basic requirements of causal models.

The second important perspective included in the journal is that of a state legislator and school principal who advocate the models' use in accountability systems. There is a striking contrast between their confidence in the models and the concerns of most of the statisticians writing in the journal. This is especially true of the state legislator who is either not aware of, or not concerned about, the technical issues the statisticians raise. He is interested in "shifting to a more student-oriented" system and sees value-added models as a way to do that.

Although the jury is still out on the technical merits and applicability of value-added models, there is little doubt that there is tremendous interest in them from the policy perspective. If Association leaders and staff can share some of the concerns and cautions raised in this journal with school and district leaders - and state policy makers - more appropriate value-added models may ultimately emerge. Until then, the Association must guard against the inappropriate use of value-added systems.