Methodological Concerns About the Education Value-Added Assessment System

Audrey Amrein-Beardsley

EDUCATIONAL RESEARCHER 2008; 37; 65
DOI: 10.3102/0013189X08316420

Methodological Concerns About the Education Value-Added Assessment System

Audrey Amrein-Beardsley

Value-added models help to evaluate the knowledge that school districts, schools, and teachers add to student learning as students progress through school. In this article, the well-known Education Value-Added Assessment System (EVAAS) is examined. The author presents a practical investigation of the methodological issues associated with the model. Specifically, she argues that, although EVAAS is probably the most sophisticated value-added model, it has flaws that must be addressed before widespread adoption. She explores in depth the shortage of external reviews and validity studies of the model, its insufficient user-friendliness, and methodological issues about missing data, regression to the mean, and student background variables. She also examines a paradigm case in which the model was used to advance unfounded assertions.

Keywords: academic achievement; accountability; achievement gains; educational legislation; evaluation methods; K–12 education; measurement techniques; teacher effectiveness

The No Child Left Behind Act of 2001 (NCLB) mandates that all states measure student academic achievement using standardized tests and that they report on progress using Adequate Yearly Progress measures. Since the law's implementation in 2002, test researchers and statisticians have been exploring alternative analytical methods to incorporate more valid measures of student learning to document students' academic progress over time. These methods of analyzing gains, growth in scores, or the amount of knowledge added from year to year as students progress through school have been appropriately termed *value-added models.* In theory, value-added methodologies allow richer analyses of test score data. Groups of students are followed to examine and assess their learning trajectories as they progress over time through different classrooms taught by different teachers in different schools and districts.

It is more defensible, for example, to examine a teacher's effectiveness on the basis of how much the teacher's students learned from the time they entered the classroom to the time they left than by simply relying on a traditional "snapshot" measure—a measure capturing the level at which students exited the classroom independent of their level when entering. Using value-added models, teachers are not given inappropriate credit for having a stellar set of students or penalized for having a difficult-to-teach class (Ballou, 2002). Teachers, schools, and districts are simply evaluated on the value that they have added to student learning.

This approach is particularly advantageous in schools whose students traditionally have posted composite test scores below state or district averages but whose leaders know that the progress their students have made during the year is above average. Students in such schools may be categorized as below average at the end of the year, yet they may have learned more during that time than the students to whom they are compared.

In the 2007 Phi Delta Kappa/Gallup Poll, members of the general public were asked the following question:

> One way to measure a school's performance is to base it on the percentage of students passing the test mandated by the state at the end of the school year. Another way is to measure the improvement students in the school made during the year. In your opinion, which is the best way to measure the school's performance—the percentage passing the test or the improvement shown by the students? (Rose & Gallup, 2007, p. 35)

Eighty-two percent of respondents stated that the best way to measure school performance is to measure the gains posted by students longitudinally—to measure the value that the district, school, or teachers added to students' learning over time (Rose & Gallup, 2007). The public's response is indicative of the overall trend in educational measurement and evaluation. Such value-added methods are becoming increasingly popular among educators and policy makers (Olson, 2004a), testing vendors (Olson, 2004b), and the U.S. Department of Education.

To examine how current Adequate Yearly Progress measures required by NCLB might include measures of growth, the U.S. Department of Education (2006a, 2006b) funded two growth model projects in Tennessee and North Carolina and recently funded three more such projects in Arkansas, Delaware, and Florida. The projects were intended to pilot statewide initiatives to integrate value-added analyses into statewide accountability systems. Five more states were to be awarded growth model project grants by the end of 2007 (U.S. Department of Education, 2006a).

The goal of these pilot growth model projects is to inform the reauthorization of the accountability provisions written into NCLB and to incorporate best methodological practices, as all states are required to integrate value-added models into their accountability procedures (Battelle for Kids, 2007a; Olson, 2004a). The federal government is poised to spend $100 million

Educational Researcher, Vol. 37, No. 2, pp. 65–75
DOI: 10.3102/0013189X08316420

per year over the next 4 years to help states warehouse data and execute value-added assessments of their student achievement data (Hoff, 2007).

The Education Value-Added Assessment System

One of the first two states to be awarded funds for the growth model pilot project was Tennessee. Although there are several value-added models in existence (e.g., McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004b), which differ in their model and statistical assumptions (Braun, 2005; Tekwe et al., 2004), the most recognized and widely implemented model is an offshoot of the Tennessee Value-Added Assessment System (TVAAS; Battelle for Kids, 2007a; Olson, 2004a). The TVAAS was originally developed by William L. Sanders. Sanders and his associates at SAS inSchool expanded the TVAAS to produce the Education Value-Added Assessment System (EVAAS), the name by which the system is now commonly known (Sanders, 1998).

EVAAS is a statistical process that allows for large-scale tracking of change in student achievement test scores over time. With complex software and hardware, data are merged regardless of the completeness of the dataset and without data imputation. Then the system, built largely on algorithms calculated by computer, permits large-scale analyses of student achievement data from which determinations may be made about growth in student achievement and the effectiveness of districts, schools, and teachers over time (Sanders & Horn, 1994; Sanders et al., 2002).

The system promises to measure gains that students make from year to year, for up to 5 years, as students move through school. Gain scores are calculated by computing the differences between students' scale scores on state tests from one grade level to the next (Sanders et al., 2002). Then educators can specify degrees of student growth and determine the contributions that teachers, schools, and districts make toward positive or negative changes in student performance. Using mixed-model equations and sophisticated controls when needed, analyses of student test data can be run almost effortlessly (Sanders, Saxton, & Horn, 1997), though not flawlessly.

Why is the EVAAS model so popular in comparison with other value-added models? The only real differences between the EVAAS model and others are that the EVAAS model is advertised as being (a) unimpaired by students' backgrounds (race and levels of poverty), which distort all other analyses of student test score data; (b) not compromised by issues of missing data; and (c) suitable for wide implementation across states because the software for processing EVAAS data permits large-scale analyses. Other value-added models do not promise such efficient and robust applications. Additional claims in support of EVAAS include the following:

- "Educational findings that were invisible in the past are now readily apparent." (Sanders & Horn, 1994, p. 310)
- "Both accelerators and impediments to sustained academic growth can be measured in a fair, objective and unbiased manner." (SAS, 2007)
- "Without this information, educational improvement efforts cannot address the real factors that have been proven to have the greatest effect on student learning." (Sanders & Horn, 1998, p. 256)
- "NCLB raises the academic standard for all kids, while the value-added approach is going beyond that and attempting to reach an even higher standard for individuals." (Sanders, 2004)
- "There may be *some important unintended consequences* of this legislation if states do not go beyond the *No Child Left Behind* AYP [Adequate Yearly Progress] requirement [and adopt a value-added model]" (Sanders, 2003, p. 1)

Yet there is no evidence that research reports conducted internally, and especially externally, have validated such claims.

Validity

One major criticism of the EVAAS is that too few analyses have been conducted to examine and evaluate the validity of the model or, more specifically, the validity of the inferences made in EVAAS value-added reports (Braun, 2004; Glass, 1995; Kupermintz, 2003; Meyer, 1997; Walberg & Paik, 1997). The EVAAS method still needs a great deal more validity research before wide implementation is justified. To my knowledge, none of the validity studies called for have commenced.

On the website of Battelle for Kids, the nonprofit promoting the use of EVAAS (http://battelleforkids.com), is the following statement: "Combining value-added analysis and improved high school assessments will lead to: Improved high school graduation rates, increased rigor in academic content, higher college going rates, less college remediation and increased teacher accountability." Support for these assertions would require a major series of validity studies, but none are cited or available.

Elsewhere, Battelle for Kids commissioned a study that the Voinovich Center at Ohio University conducted to measure the effects on student achievement of access to value-added data on the part of the students' school districts (Battelle for Kids, 2006). Results indicated that districts whose leaders took advantage of value-added information showed statistically significant gains in student achievement. But the summary of the technical report reveals that only 50% (6 in 12) of the school districts using value-added data posted greater gains than similar districts with which they were matched.

Content-Related Validity

To collect content-related evidence of validity, it is necessary to examine whether test scores measure what students learn and are able to do. Initially, this was a significant problem with the EVAAS model because the model used norm-referenced tests that were not aligned with state standards to make judgments about whether districts, schools, and teachers were effective. Now, with the increased use of criterion-referenced tests linked to state standards, this is less of a pressing issue.

However, whether states' criterion-referenced tests can be used to accurately measure student growth presents a different problem. Grade-level assessments are not sensitive measures of growth; and the further the student learning is from grade level, the less reliable those assessments are. "The progress of students who are well above or below grade level is effectively invisible," according to the Delaware Statewide Academic Growth Assessment Pilot of 2007 (Rodel Foundation of Delaware, p. 9). That study found that multigrade adaptive growth assessments

yield more valid interpretations about student learning over time, particularly in high-needs schools.

Criterion-Related Validity

To generate criterion-related evidence of validity, it is also necessary to assess whether teachers who post large gains from year to year are the teachers deemed most effective through other, independent measures of teacher quality (see, e.g., McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004a). Value-added analysts might examine whether teachers who post large gains from year to year are the same teachers whom supervisors described as the most effective in teacher reviews or evaluations. Analysts might examine whether teachers who post the smallest gains from year to year are the same teachers whom supervisors categorize as least effective. Analysts might also examine whether teachers who post high or low scores on teacher licensure tests, teachers with more or fewer years of experience, and teachers with graduate or basic degrees in content and pedagogy post respectively larger or smaller value-added gains over time. Investigations must be conducted to determine whether the results yielded by EVAAS are supported or contradicted by other predictors with high face validity. These investigations are needed to validate the EVAAS model and should have been conducted before the widespread adoption that is currently under way.

EVAAS developers recently began using their value-added data to project how far students will go in their educational futures, for example, by predicting what students' scores will be on the ACT when they seek entrance to college (Olson, 2002). In these predictions, no mention is made of whether students will be followed to confirm that the predictions based on the EVAAS model come true, but projections are being made nonetheless (Sanders, 2003). This situation has another set of very troublesome implications and consequences.

Construct-Related Validity

To establish construct-related evidence of validity—a significant undertaking—it is important to discuss whether achievement tests can effectively measure the constructs of school and teacher quality. At the root of the problem is that these models rely on large-scale standardized tests to make valid statements about what students know and are able to do. Whether high-stakes tests can be used to make valid inferences about student knowledge, teacher quality, or school or district effectiveness is far from certain. And the question whether analyzing gains in test scores using value-added models can effectively measure "growth" in other ways than measuring "growth upward" from one year to the next (Reckase, 2004) warrants further inquiry.

Kupermintz (2003) conducted a validity investigation of the EVAAS model examining its definition of teacher effectiveness and concluded that the model's heavy reliance on test score gains oversimplified the construct of teacher effectiveness. This is not the only educational measurement tool that treats the relationship between test scores and teacher effectiveness inadequately and simplistically, however.

Consequence-Related Validity

To establish consequence-related evidence of validity, empirical studies are needed to evaluate whether the EVAAS method will help to improve student learning in schools. Very few studies of this type exist (Rubin, Stuart, & Zanutto, 2004). But for the model to serve schools well, it must work to improve schools, not just to report on them.

Using Data in Formative Ways

One major finding about the EVAAS value-added output and score reports is that, of the districts and schools that have implemented the EVAAS model, too few appear to be using output data in formative ways (Raudenbush, 2004). Morgan (2002) found that confusing data reports and a lack of training for teachers and administrators in how to understand the data reports were preventing schools and teachers from using value-added data to improve student learning and achievement. The students in districts and schools that implemented the EVAAS value-added model did not benefit strategically from their districts' or schools' involvement, as expected.

For the EVAAS model to work, it must be statistically sophisticated; however, as the model becomes more complicated, it becomes less user-friendly. Ballou clarifies this point: "When statistical methods are used to minimize error or 'noise,' the systems quickly become incomprehensible to educators, losing the 'transparency' that many argue is a hallmark of effective accountability systems" (as quoted in Olson, 2002, p. 14; see also Reckase, 2004; Tekwe et al., 2004). Educators want to use relatively simple, understandable statistical models to analyze educational phenomena, but social complexity demands that statistical models be sophisticated enough to capture reality with integrity (Andrejko, 2004; Callendar, 2004). The EVAAS value-added model is caught on the horns of this dilemma.

Sanders responds that "one [doesn't] have to understand how a car works in order to drive it" (as cited in Braun, 2005, p. 16).

> Most everyone can use a cellular telephone, but virtually no one knows, or needs to know, how to build the phone. . . . If it were necessary for each user to know how to build the device prior to appropriate use, then all of our phones would be restricted to tin cans and string. (Sanders, 2000, p. 336)

Sanders and his colleagues (1997) contend that the spokespeople for the system are the educators who use value-added reports to inform educational practices and reforms in their schools; yet no citations are provided to validate these claims. Walberg and Paik (1997) also raise concerns about how these anecdotes were gathered and whether they are representative.

Currently, representatives of Battelle for Kids acknowledge these points of confusion and, in Ohio, are offering large-scale training sessions so that sets of value-added specialists may learn more about the benefits and uses of the EVAAS value-added model and how to use and interpret value-added score reports at the regional and district levels (Battelle for Kids, 2007b).

Peer Review

Another significant issue with the EVAAS value-added method is that the developers have not made this method completely open for peer review. Specifically, they hold as proprietary information the computational algorithms needed to manage and solve large systems of linear equations. This makes peer review by

external statisticians impossible (Dorn, 1994; Glass, 1995; Kupermintz, 2003). Nor have the EVAAS developers released their value-added data to allow other educational researchers to conduct replications or confirmatory analyses of their findings. My own and others' attempts (Kupermintz, 2003) to access the EVAAS value-added data have consistently gone without response or been refused with the justification that the value-added data, if released to external researchers, might be misrepresented. But it is not unusual for researchers to allow external statisticians to conduct replications or confirmatory analyses of similar datasets to validate research findings using different methods, if all sensitive identifying data are removed.

In 1997, Sanders, Saxton, and Horn (p. 140) asserted that they had undertaken "extensive efforts" to increase understanding of the value-added method and to explain the model in the greatest detail to be found in all reports to date. Sanders (1998) stated that "detailed external reviews from both the statistical and educational evaluation communities have confirmed that the properties of the TVAAS results are as claimed" (p. 26); however, nowhere does he provide citations of or references to these external reviews. It seems that three groups of external reviewers examined the model in depth. Two groups praised the model; one group raised significant points of contention, including concerns about the model's uncritical acceptance, widespread adoption, and rapid application across states (see also Braun, 2005).

In 1995 a statistician served as an external reviewer and examined the value-added system, formerly referred to as the TVAAS. In a brief technical review he endorsed it as a "statistically sound and appropriate system" for evaluating teacher, school, and district effectiveness (Harville, 1995, p. 1). In a more in-depth technical review also conducted in 1995, a statistician analyzed a hypothetical set of data using the model's value-added software. He concluded that the model, the statistical assumptions made, and the software used were reasonable and defensible, particularly because analyses of the hypothetical data yielded the same results as did Sanders's value-added model (Stroup, 1995). However, as he used the same software, his conclusion is no surprise.

In 1996, Bock, Wolfe, and Fisher assessed the model. Although they praised it for its structure, its method, and the sheer size of its student performance database, they expressed significant reservations. They noted concerns about the terms of value-added agreements with the state of Tennessee and recommended a set of audits to monitor contractual procedures. They observed that the data were not disaggregated by subgroups (race, socioeconomic status, gender), which was problematic, especially given federal mandates that states report student achievement levels by subgroup to assess levels of Adequate Yearly Progress. And the reviewers attributed large fluctuations in scores across schools and districts to the improper horizontal and vertical equating of tests (see also Braun, 2004, 2005; Stroup, 1995).

Bock and his colleagues (1996) mention concerns with the inability of the model to disentangle one teacher's effects on an adjacent teacher's student test scores over time to produce statistically unbiased estimates of teacher effectiveness (Braun, 2005; Kupermintz, 2003). Developers respond that the model tempers these effects through a strategy of stacked blocking, which enables the partitioning of these effects (Sanders, 1994; Stroup, 1995). But whether this method of blocking works satisfactorily is unknown, especially given the powerful additive and cumulative effects that Sanders and Rivers (1996) found that teachers have on student achievement over time. This is very problematic because we know well that the gains a student produces in one year include contributions (small and large) made by that student's prior teachers (Meyer, 1997). The precise size of these additive and cumulative effects and how long they persist are unknown (as discussed in Olson, 2004a).

Reviewers also express concerns with the ways that data are reported to the public and how educators might misuse the data when they rely solely on value-added data as a single measure to evaluate educational effectiveness. Ross, Stringfield, Sanders, and Wright (2003) and others (Braun, 2005; Carter, 2004; Dorn, 1994; Kim & Sunderman, 2005; Koretz, 2002; McCaffrey et al., 2004a; Rubin et al., 2004) argue that summative uses of these data in isolation from other indicators of effectiveness would be negligent, especially if high stakes are attached to results. Yet this is likely to occur, despite warnings against inappropriate uses of the results derived from the EVAAS model.

Rubin et al. (2004) suggest that value-added results be used only as "face-value" indicators of school improvement and teacher effectiveness; they advise that the results not be used in isolation from other indicators of school and teacher quality. Others researchers (McCaffrey et al., 2004a) consider making high-stakes decisions using value-added data to be better than the traditional choice, given that both are substandard practices.

Missing Data

Bock and colleagues (1996) point to issues caused by incomplete records. Value-added models require complete and high-quality longitudinal data that many states currently do not have. EVAAS developers claim that the model can operate regardless of the amount of missing or fractured data always found in large student achievement databases (Sanders, 2000). They argue that data can be merged at a rate of about 90% (Sanders et al., 2002), but their argument seems improbable, especially if the percentage is expected to stand up over a period of 5 years, which EVAAS developers argue the model can tolerate (see, e.g., Ross, Wang, et al., 2001).

Usually, student test score data are not linked to teacher names or identification numbers; and, often, teachers are missing some or even most of their students' data. Students are also often misreported by class and grade level. These simple miscodes affect thousands of student records. Conducting longitudinal analyses also complicates this issue enormously. If a teacher has a nearly full set of data in one year but is missing student records from the year before, it is functionally impossible to measure learning gains or to evaluate learning trajectories over time. The missing data problem biases estimates of teacher effectiveness (see also Braun, 2004, 2005; Lockwood, Doran, & McCaffrey, 2003; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Walberg & Paik, 1997). As noted by McCaffrey et al. (2004b),

> Given the large proportion of missing data in many achievement databases and known differences between students with complete and incomplete test data, it is possible that estimates may be highly sensitive to this (or other) assumptions about missing data. (p. 97)

In short, built into value-added models is an assumption that missing data are irrelevant or randomly distributed. This assumption is extremely problematic because it is well known that disproportionate numbers of students who do not participate in large-scale tests are low performing (and some are even encouraged to miss school the day of the test). These missing data cannot be ignored or assumed to be irrelevant (Raudenbush, 2004; Rubin et al., 2004).

Regression to the Mean

Adding to the problem is that the model measures teacher effectiveness by error or deviation from the mean (Boyd et al., 2006; Medina, 2008; Rivkin, 2007). And teachers who teach smaller classes are pulled toward the mean (Kupermintz, 2003; McCaffrey et al., 2003; Sanders et al., 1997). These teachers are more likely assumed to be average regardless of whether they are in fact excellent or inadequate. An expert teacher, because he or she teaches more students, may be labeled above average, whereas an equally expert teacher with significantly fewer students might not be acknowledged at all, being misclassified as average because of having fewer student records. Inversely, an ineffective teacher who teaches a large class might be penalized for being below average, whereas an equally ineffective teacher who teaches a smaller class may go undetected.

This is likely why special education students are not included in the EVAAS value-added model unless data are aggregated at the school or district level (Topping & Sanders, 2000). This exclusion is problematic if schools are to disaggregate and report value-added data by subgroups as mandated by NCLB. Ultimately, the issue of small class size alters the practicality of the EVAAS value-added approach. EVAAS becomes a model that effectively distinguishes only between the best and the worst teachers whose class sizes are above an arbitrary number.

Extraneous Variables

Controversially, within the EVAAS model, student risk factors (race and poverty) are not controlled for. This makes the EVAAS the only sophisticated analytical model for measuring student achievement that does not account for student background factors—factors that have been shown by decades of research to bias achievement data (Braun, 2005; Kupermintz, 2003; McCaffrey et al., 2003, 2004b; Tekwe et al., 2004). Instead, the EVAAS system measures teacher effectiveness using student gain scores that *implicitly* control for students' backgrounds (Ballou, Sanders, & Wright, 2004). The model controls for these extraneous variables by allowing students to serve as their own controls.

Developers of the EVAAS model state that the effects attributable to race and socioeconomic status on student growth are negligible. Because the students' gains are analyzed from one year to the next and across subject areas, the influence that their backgrounds might have on their learning is controlled for or buried within their previous years' test scores, across years and subjects. Accounting for students' prior levels of knowledge in this way, the EVAAS developers argue, cancels out the influence that students' backgrounds would otherwise have had on test score gains. This approach, the developers say, allows for a "truer" assessment of what students learn from one year to the next (Ballou et al., 2004; see also Ross, Sanders, et al., 2001; Ross et al., 2003; Ross, Wang, et al., 2001; Sanders, 1998, 2004; Sanders & Horn, 1994, 1998; Wright, Horn, & Sanders, 1997).

In practical terms, EVAAS developers claim that looking at change or growth eliminates the background factors that normally contaminate analyses of test score data. "It appears that the biggest factor affecting academic progress of children is classroom instruction; not race, ethnicity or ability of the student" (Sanders, 2004). In other words, bright students do not learn any faster than their less able classmates. Obviously, this defies common knowledge and common sense, especially given what multiple studies have evidenced throughout the history of educational research on student achievement (Coleman et al., 1966; Marchant, Paulson, & Shunk, 2006; White, 1982). Family income, ethnicity, ability, and other background variables unquestionably affect levels of student achievement and the progress that students make from year to year.

Gene V. Glass once posed the following question to Sandra P. Horn, an advocate of the model: "Given two classrooms, one with average IQ 80 and one with average IQ 120 and identical in every other respect and taught exactly identically by two identical teachers, do you believe your analysis would produce equal measured gains?" She replied, "Yes." Glass commented that Horn, of course, had to say yes, or the method would be exposed as invalid (G. V. Glass, personal communication, April 5, 2007). Students with higher levels of intelligence undoubtedly learn more than students with lower levels of intelligence, and because intelligence is correlated with background factors, Horn's reasoning is simply wrong. If two classes are equated on past achievement but differ greatly on IQ, one will make more progress during the year, and it can have nothing to do with the teacher.

In addition, if student background variables do not affect measures of growth in student achievement, why is it that the achievement gap persists between White, Asian American, and wealthier students, on the one hand, and students from traditionally marginalized backgrounds, on the other, even since the passage of NCLB (Northwest Evaluation Association, 2006)? How can the achievement gap continue to widen if all students, regardless of race, ethnicity, or ability, learn at the same rate?

And what does this mean for evaluating teacher effectiveness, when students are not randomly placed in certain teachers' classes? If one teacher is given a stellar set of students and another equally effective teacher is given an average set of students, the former set of students will undoubtedly learn and gain more over time than will the average set of students. Yet the teacher with the average set of students will be penalized as relatively less effective. Without randomized assignment of students to teachers, never can statements be made that one teacher caused students to learn more than another, unless one accepts a set of "heroic assumptions" (Rubin et al., 2004).

EVAAS developers must recheck their model and adjust it to properly account for students' backgrounds. Instead of claiming that their model is impervious to background and other demographic variables, they must use logic and reason to investigate why these effects have nearly vanished in their model. Relying on the precision of their statistical model instead of on common sense is remiss, particularly when this model is being adopted at a great rate across the country.

National Board–Certified Teachers and Student Achievement

In an unpublished technical report, Sanders, Ashton, and Wright (2005) examined whether National Board–certified teachers (NBCTs) are more effective than their regularly certified peers. NBCTs are regularly certified school teachers with at least 3 years of classroom experience who go through a rigorous evaluation process set forth by the National Board for Professional Teaching Standards. This process assesses and validates candidates' expertise as teachers. In the three leading independent studies on the topic (Cavaluzzo, 2004; Goldhaber & Anthony, 2004; Vandevoort, Amrein-Beardsley, & Berliner, 2004), results converged; the researchers found that students of NBCTs make about one month's greater gains per year across subjects than do students taught by regularly certified teachers.

Sanders et al. (2005) put these findings to their value-added test. They compared 4 years of elementary mathematics and reading test scores of students of NBCTs and non-NBCTs in two large North Carolina school districts ($n > 260{,}000$), using four different analytical models. They applied two statistical models that were similar to those used by Cavaluzzo (2004), Goldhaber and Anthony (2004), and Vandevoort et al. (2004). But they discounted the findings from these studies on the grounds that the researchers did not use "models to properly account for the nested structure of the data" (Sanders et al., 2005, p. 4). Sanders et al. also applied two value-added models to the data, which they deemed more appropriate.

Across analyses, Sanders et al. (2005) concluded that students of NBCTs *did not* increase their achievement at significantly greater rates than students of non-NBCTs. Until then, it had been widely accepted that NBCTs were in fact expert teachers and had proven themselves as such in the classroom. This claim, counterintuitive as it was, had a notable impact.

If the technical report by Sanders et al. (2005) had been sent out for peer review before its release, the peer reviewers would have found a different picture painted by a very simple reanalysis of the data provided in the report. But the report authors did not provide their data for external researchers. Only reanalyses of the data provided in their reports and appendixes could be used to verify or challenge the findings derived from their value-added model.

In their report, Sanders et al. (2005) provide figures in four tables titled 3A, 3B, 3C, and 3D (pp. 14–17). Data derived from the two models similar to those used in the three aforementioned studies were provided in Table 3A for Model 1 and in Table 3C for Model 3 (Sanders et al., 2005, pp. 14, 16). The data are presented here in Table 1.

A simple reanalysis of these data illustrates that of the 60 comparisons made across mathematics and reading that were based on Sanders et al.'s (2005) replication of the traditional models, students of NBCTs outperformed students of non-NBCTs 78.3% of the time (47 of 60 math and reading estimates). Of the statistically significant comparisons, students of NBCTs outperformed students of non-NBCTs 91% of the time (21 of 23 observations).

Effect sizes were also similar to those reported in the three earlier studies, supporting these researchers' initial conclusions. Of the statistically significant findings, students of NBCTs made about one month's greater gains in math (mean $ES = 0.10$) and about one third of a month's greater gains in reading (mean $ES = 0.03$) than did students of non-NBCTs. Combined, students of NBCTs made about three fourths of a month's greater gains in achievement (mean $ES = 0.08$) than did students in classrooms with non-NBCTs.

Some educational statisticians believe that only statistically significant effect sizes should be included in calculations of average effect sizes (Robinson & Levin, 1997); others criticize this position on the basis that it can lead to misinterpretations of overall results. Members of the second camp believe that all effect sizes should be reported and averaged regardless of statistical significance (Thompson, 2006).

Including all effect sizes, students of NBCTs made about three fourths of a month's greater gains in math (mean $ES = 0.08$) and about one third of a month's greater gains in reading (mean $ES = 0.03$) than students of non-NBCTs. Combined, students of NBCTs made just over one-half month's greater gains in achievement (mean $ES = 0.05$) per year than did students in classrooms with non-NBCTs.

Sanders et al. (2005) also analyzed their data using two value-added models that they deemed more sophisticated than the models used in the previous three studies. These models (a) controlled for teacher effects, even though two of the three aforementioned studies controlled for teacher effects as well; (b) used students' test scores from their previous grades as covariates, as did researchers in the prior three studies; and (c) used value-added gain scores to assess the academic value that NBCTs added to their students' achievement in comparison with the NBCT's regularly certified peers. Again, a simple reanalysis of the data contradicts the conclusions of Sanders et al.

Data derived from the two more sophisticated models were provided in Table 3B for Model 2 and Table 3D for Model 4 (Sanders et al., 2005, pp. 15, 17). They are presented here in Table 2.

A simple reanalysis of these data illustrates that of the 60 comparisons made across mathematics and reading using Sanders's value-added method of analysis, students of NBCTs outperformed students of non-NBCTs 78.3% of the time (47 of 60 math and reading estimates). Of the statistically significant comparisons, students of NBCTs outperformed students of non-NBCTs 83% of the time (5 of 6 observations).

Effect sizes were also similar but smaller than the previously reported effect sizes. Of the statistically significant findings, students of NBCTs made about three fourths of a month's greater gains in math (mean $ES = 0.08$) and about one week's greater gains in reading than students of non-NBCTs (mean $ES = 0.02$). Combined, students of NBCTs made almost one-half month's greater gains in achievement (mean $ES = 0.04$) than students in classrooms with non-NBCTs.

Including all effect sizes, students of NBCTs made just over one-half month's greater gains in math (mean $ES = 0.06$) and about one week's greater gains in reading than students of non-NBCTs (mean $ES = 0.02$). Combined, students of NBCTs made almost one-half month's greater gains in achievement (mean $ES = 0.04$) than did students in classrooms with non-NBCTs.

Yet Sanders and his colleagues (2005) chose to highlight and overemphasize the negative observations in the data. They overemphasized the three significant negative findings posted in their analysis, overlooking the fact that only 3 in 120 (2.5%)

Table 1
Math and Reading Figures: Traditional (Replicated) Models 1 and 3

No.	Model	Grade	Comparison	Math Estimate	Math Effect Size	Reading Estimate	Reading Effect Size
1	1	4	NBCT vs. Failed	0.34*	0.08	0.27	0.05
2	1	4	NBCT vs. Future	0.46**	0.10	0.17	0.03
3	1	4	NBCT vs. Never	0.19	0.04	0.02	0.00
4	1	5	NBCT vs. Failed	0.44*	0.09	0.26	0.05
5	1	5	NBCT vs. Future	0.74**	0.15	0.06	0.01
6	1	5	NBCT vs. Never	0.36**	0.07	0.28**	0.06
7	1	6	NBCT vs. Failed	0.99**	0.22	0.50*	0.10
8	1	6	NBCT vs. Future	0.54**	0.12	0.35	0.07
9	1	6	NBCT vs. Never	0.56**	0.12	0.60**	0.12
10	1	7	NBCT vs. Failed	0.00	0.00	–0.42*	–0.09
11	1	7	NBCT vs. Future	0.09	0.02	0.21	0.04
12	1	7	NBCT vs. Never	0.34	0.07	–0.11	–0.02
13	1	8	NBCT vs. Failed	1.37**	0.28	0.40	0.08
14	1	8	NBCT vs. Future	–0.07	–0.01	0.28	0.06
15	1	8	NBCT vs. Never	0.24	0.05	0.11	0.02
16	3	4	NBCT vs. Failed	0.26	0.05	0.06	0.01
17	3	4	NBCT vs. Future	0.33*	0.07	–0.13	–0.02
18	3	4	NBCT vs. Never	0.11	0.02	–0.24	–0.04
19	3	5	NBCT vs. Failed	0.46*	0.09	0.29	0.05
20	3	5	NBCT vs. Future	0.74**	0.15	0.07	0.01
21	3	5	NBCT vs. Never	0.34**	0.07	0.13	0.02
22	3	6	NBCT vs. Failed	0.49*	0.10	0.51	0.09
23	3	6	NBCT vs. Future	0.04	0.01	–0.20	–0.03
24	3	6	NBCT vs. Never	–0.01	0.00	0.22	0.04
25	3	7	NBCT vs. Failed	–0.13	–0.03	–0.28	–0.05
26	3	7	NBCT vs. Future	0.22	0.04	0.22	0.04
27	3	7	NBCT vs. Never	0.46*	0.09	–0.46**	–0.07
28	3	8	NBCT vs. Failed	1.09**	0.20	1.39**	0.23
29	3	8	NBCT vs. Future	–0.01	0.00	0.13	0.02
30	3	8	NBCT vs. Never	0.30*	0.05	–0.31	–0.05

Note. NBCT = National Board–certified teachers. Sanders, Ashton, and Wright (2005) analyzed various subgroups of regularly certified teachers in other parts of their analysis: NBCTs versus regularly certified teachers who had (a) failed the National Board exams ("Failed," in this table), (b) stated that in the future they might seek National Board certification ("Future"), and (c) never been involved with the National Board certification process ("Never"). Data on these subgroups are aggregated here to compare NBCTs with all regularly certified teachers regardless of their involvement with the National Board, as was done in the previous three studies.
*$p < 0.05$. **$p < 0.01$.

analyses conducted across models in their study yielded a negative significant effect. Seemingly to downplay the positive effects of NBCTs, the researchers focused on these negative observations, made no mention of the considerably larger number of comparisons favoring NBCTs, and used these negative observations to speculate that NBCTs may in fact be *hindering* levels of student achievement or *disadvantaging* student learning.

But this simple reanalysis of their data in fact confirms, again, that students of NBCTs learn significantly more than students of regularly certified teachers. Effect sizes vary depending on model, but across all models the findings still stand that students of NBCTs make about one month's greater gains in student achievement than do students of regularly certified teachers. Sanders's more sophisticated value-added analyses just (a) reduced the numbers of statistically significant findings and (b) weakened effect sizes.

Conclusions

So what does the reality of this one study suggest about the credibility of findings from other studies using the EVAAS model? Do teachers really have the claimed residual or carryover effects year after year (Sanders & Rivers, 1996)? Do classroom heterogeneity and class size have little or no relationship to student achievement (Wright et al., 1997)? Does the effect of the teacher in the classroom completely outweigh the effects of students' backgrounds (Wright et al., 1997; Sanders & Horn, 1998)? Do students who read more books, particularly if they are difficult books, learn more about reading (Topping & Sanders, 2000)? Do students in schools that undergo significant restructuring realize significant learning gains after reform (Ross, Wang, et al., 2001)? When students change buildings in school, do they lose significant amounts of knowledge after the transfer (Sanders et al., 2002)? How do we know? How can we be sure? Is it unfair to use the flawed interpretation of the results from this one study of NBCTs to question the assertions derived from other such value-added analyses?

In this article I argue that although the EVAAS model is probably the best and most sophisticated one we have of this type (Gormley & Weimer, 1999), or "the least bad" (Walberg & Paik,

Table 2
Math and Reading Figures: Sanders's Models 2 and 4

No.	Model	Grade	Comparison	Math Estimate	Math Effect Size	Reading Estimate	Reading Effect Size
1	2	4	NBCT vs. Failed	0.22	0.05	0.17	0.03
2	2	4	NBCT vs. Future	0.22	0.05	0.06	0.01
3	2	4	NBCT vs. Never	0.05	0.01	–0.02	0.00
4	2	5	NBCT vs. Failed	0.46	0.10	0.27	0.06
5	2	5	NBCT vs. Future	0.64*	0.14	0.06	0.01
6	2	5	NBCT vs. Never	0.32	0.07	0.32*	0.07
7	2	6	NBCT vs. Failed	0.59	0.14	0.46	0.09
8	2	6	NBCT vs. Future	0.07	0.02	0.12	0.02
9	2	6	NBCT vs. Never	0.27	0.06	0.58**	0.11
10	2	7	NBCT vs. Failed	0.23	0.05	–0.49	–0.10
11	2	7	NBCT vs. Future	0.44	0.09	0.15	0.03
12	2	7	NBCT vs. Never	0.40	0.08	0.05	0.01
13	2	8	NBCT vs. Failed	0.82	0.17	0.36	0.08
14	2	8	NBCT vs. Future	–0.20	–0.04	0.37	0.08
15	2	8	NBCT vs. Never	0.24	0.05	0.49**	0.10
16	4	4	NBCT vs. Failed	0.16	0.04	0.03	0.01
17	4	4	NBCT vs. Future	0.14	0.03	–0.15	–0.03
18	4	4	NBCT vs. Never	–0.01	0.00	–0.22	–0.04
19	4	5	NBCT vs. Failed	0.53	0.11	0.31	0.06
20	4	5	NBCT vs. Future	0.70*	0.15	0.03	0.01
21	4	5	NBCT vs. Never	0.29	0.06	0.09	0.02
22	4	6	NBCT vs. Failed	0.28	0.06	0.44	0.08
23	4	6	NBCT vs. Future	–0.30	–0.06	–0.44	–0.08
24	4	6	NBCT vs. Never	–0.47	–0.10	–0.04	–0.01
25	4	7	NBCT vs. Failed	0.13	0.03	–0.49	–0.08
26	4	7	NBCT vs. Future	0.56	0.12	0.10	0.02
27	4	7	NBCT vs. Never	0.56	0.12	–0.75*	–0.12
28	4	8	NBCT vs. Failed	0.54	0.10	1.42	0.24
29	4	8	NBCT vs. Future	0.16	0.03	0.34	0.06
30	4	8	NBCT vs. Never	0.05	0.01	–0.24	–0.04

Note. NBCT = National Board–certified teachers. Sanders, Ashton, and Wright (2005) analyzed various subgroups of regularly certified teachers in other parts of their analysis: NBCTs versus regularly certified teachers who (a) failed the National Board exams ("Failed," in this table), (b) stated that in the future they might seek National Board certification ("Future"), and (c) never had been involved with the National Board certification process ("Never"). Data on these subgroups are aggregated here to compare NBCTs v. all regularly certified teachers regardless of their involvement with the National Board, as was done in the previous three studies.
*$p < 0.05$. **$p < 0.01$.

1997, p. 171), and is not necessarily wrongheaded, there are significant issues that must be addressed before wide acceptance. The insufficiency of validity studies, the difficulties with the user-friendliness of the model, the lack of external reviews, and the methodological issues with missing data, regression to the mean, and student background variables were explored in depth. A paradigm case in which the model was used to advance unfounded assertions was also examined.

The mission of the U.S. Food and Drug Administration is to ensure that no harm is done to consumers of foods and drugs. The agency protects and advances public health and provides consumers with scientifically based information needed to improve or preserve their well-being. It ensures that foods are safe and wholesome and that drugs are safe and effective. And it ensures that foods and drugs are honestly, accurately, and informatively represented to the public.

Do not students and teachers in America's schools deserve similar protection? Who protects them from assessment models that could do as much harm as good? Who protects their well-being and ensures that assessment models are safe, wholesome, and effective? Who guarantees that assessment models honestly and accurately inform the public about student progress and teacher effectiveness? Who regulates the assessment industry?

The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) endorse a common set of assessment standards, set forth in *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2000). The standards represent the professional consensus on the appropriate uses of tests.

AERA (2000) has issued 12 recommendations for high-stakes testing based on these standards, all of which pertain to the EVAAS value-added model. Six of them (paraphrased below) are most relevant here, as they pertain to the practical problems with this model that were addressed earlier:

- High-stakes decisions should not be made on the basis of a single test score.
- High-stakes tests must be validated for each intended use.
- The negative side-effects of a high-stakes assessment program must be fully disclosed to policy makers.
- The accuracy of achievement levels (based on gains in this case) must be established.
- Students with disabilities must be appropriately attended to.
- The intended and unintended effects of the testing program must be continuously evaluated and disclosed.

These are the time-tested principles and commitments that should be applied to all large-scale assessment systems, especially when high stakes are attached to the test results. These questions have yet to be addressed satisfactorily by EVAAS developers.

NOTE

I would like to thank Chris Clark, David Berliner, Gene Glass, Thomas Haladyna, Ray Buss, and my writers' group at Arizona State University for their thoughts, inspiration, and editorial assistance in the writing of this article.

REFERENCES

American Educational Research Association. (2000, July). *AERA position statement on high-stakes testing in pre-K–12 education*. Washington, DC: Author. Available at https://www.aera.net/policyandprograms/?id=378

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2000). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Andrejko, L. (2004). Value-added assessment: A view from a practitioner. *Journal of Educational and Behavioral Statistics, 29*(1), 7–9.

Ballou, D. (2002, Summer). Sizing up test scores. *Education Next.* Retrieved February 17, 2007, from http://www.hoover.org/publications/ednext/3365706.html

Ballou, D., Sanders, W. L., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37–66.

Battelle for Kids. (2006, May 2). *Achievement gains study of SOAR pilot and matched districts: Does SOAR make a difference? A summary of the final report to Battelle for Kids.* Columbus, OH: Author.

Battelle for Kids. (2007a). *Bringing clarity to school improvement.* Retrieved February 11, 2008, from http://battelleforkids.com

Battelle for Kids. (2007b). *Ohio's value-added rollout.* Retrieved February 11, 2008, from http://battelleforkids.com/home/value_added/v_a_0hio

Bock, R. D., Wolfe, R., & Fisher, T. H. (1996). *A review and analysis of the Tennessee value-added assessment system: Summary and recommendations.* Nashville: Tennessee Office of Education Accountability.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., Michelli, N., & Wyckoff, J. (2006). Complex by design: Investigating pathways into teaching in New York City schools. *Journal of Teacher Education, 57*(2), 102–119.

Braun, H. I. (2004, December). *Value-added modeling: What does due diligence require?* Princeton, NJ: Educational Testing Service.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models.* Princeton, NJ: Educational Testing Service. Retrieved March 13, 2007, from http://www.ets.org/Media/Research/pdf/PICVAM.pdf

Callendar, J. (2004). Value-added student assessment. *Journal of Educational and Behavioral Statistics, 29*(1), 5.

Carter, R. L. (2004). Rejoinder. *Journal of Educational and Behavioral Statistics, 29*(1), 135–137.

Cavaluzzo, L. (2004). *Is National Board certification an effective signal of teacher quality?* Alexandria, VA: Center for Naval Analysis. Retrieved February 20, 2006, from http://www.cna.org/documents/CavaluzzoStudy.pdf

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F. D., et al. (1966). *Equality of educational opportunity.* Washington, DC: Department of Health, Education and Welfare.

Dorn, S. (1994, September 9). Message posted to Discussion of Tennessee Value-Added Assessment System, archived at http://epaa.asu.edu/tvaas.html

Glass, G. V. (1995). Message posted to Discussion of Tennessee Value-Added Assessment System, archived at http://epaa.asu.edu/tvaas.html

Goldhaber, D., & Anthony, E. (2004). *Can teacher quality be effectively assessed? National Board certification as a signal of effective teach*ing. Seattle, WA: Center on Reinventing Public Education. Retrieved January 20, 2006, from http://www.crpe.org/workingpapers/pdf/NBPTSquality_report.pdf

Gormley, W. T., & Weimer, D. L. (1999). *Organizational report cards.* Cambridge, MA: Harvard University Press.

Harville, D. A. (1995). *A review of the Tennessee Value-Added Assessment System (TVAAS).* Ames: Iowa State University. Retrieved March 13, 2007, from http://www.cgp.upenn.edu/pdf/Harville-A%20Review%200f%20TVAAS.pdf

Hoff, D. J. (2007, February 13). NCLB panel calls for federal role in setting national standards. *Education Week.* Retrieved February 13, 2007, from http://www.edweek.org/ew/articles/2007/02/13/23aspen_web.h26.html

Kim, J. S., & Sunderman, G. L. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher, 43*(8), 3–13.

Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources, 37*(4), 752–777.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value-Added Assessment System. *Educational Evaluation and Policy Analysis, 25,* 287–298.

Lockwood, J. R., Doran, H., & McCaffrey, D. F. (2003). Using R for estimating longitudinal student achievement models. *R News, 3*(3), 17–23.

Marchant, G. J., Paulson, S. E., & Shunk, A. (2006). Relationships between high-stakes testing policies and student achievement after controlling for demographic factors in aggregated data. *Education Policy Analysis Archives, 14*(30). Retrieved March 22, 2007, from http://epaa.asu.edu/epaa/v14n30/

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: Rand.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004a). Let's see more empirical studies on value-added modeling of teacher effects: A reply to Raudenbush, Rubin, Stuart and Zanutto, and Reckase. *Journal of Educational and Behavioral Statistics, 29*(1), 139–143.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004b). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67–101.

Medina, J. (2008, January 21). New York measuring teachers by test scores. *New York Times.* Retrieved January 22, 2008, from http://www.nytimes.com/2008/01/21/nyregion/21teachers.html

Meyer, R. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review, 16,* 183–301.

Morgan, J. P. (2002). *Multiple choices: Testing students in Tennessee.* Nashville, TN: Comptroller of the Treasury, Office of Education Accountability.

No Child Left Behind Act of 2001, Pub. L. No. 107–110 (2001).

Northwest Evaluation Association. (2006). *Achievement gaps: An examination of differences in student achievement and growth.* Lake Oswego, OR: Author. Retrieved January 29, 2007, from http://www.nwea.org/research/achievementgap.asp

Olson, L. (2002, November 20). Education scholars finding new "value" in student test data. *Education Week.* Retrieved March 13, 2007, from http://www.edweek.org/ew/articles/2002/11/20/12value.h22.html

Olson, L. (2004a, November 16). "Value added" models gain in popularity. *Education Week.* Retrieved January 16, 2006, from http://www.edweek.org/ew/articles/2004/11/17/12value.h24.html

Olson, L. (2004b, November 30). NCLB law bestows bounty on test industry. *Education Week.* Retrieved January 16, 2006, from http://www.edweek.org/ew/articles/2004/12/01/14tests.h24.html

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*(1), 121–129.

Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics, 29*(1), 117–120.

Rivkin, S. G. (2007, November). *Value-added analysis and education policy.* Washington, DC: Urban Institute, National Center for Analysis of Longitudinal Data in Education Research. Retrieved December 27, 2007, from http://www.urban.org/UploadedPDF/411577_value-added_analysis.pdf

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21–26.

Rodel Foundation of Delaware. (2007, October 25). *A more accurate growth model: Using multigrade adaptive assessments to measure student growth* (Delaware Statewide Academic Growth Assessment Pilot Steering Committee). Wilmington, DE: Author. Retrieved December 29, 2007, from http://www.rodelfoundationde.org/pdfs/GrowthReportCongress102507.pdf

Rose, L. C., & Gallup, A. M. (2007). The 39th Annual Phi Delta Kappa/Gallup Poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 89*(1), 33–48. Retrieved October 14, 2007, from http://www.pdkintl.org/kappan/kpollpdf.htm

Ross, S. M., Sanders, W. L., Wright, S. P., Stringfield, S., Wang, W. L., & Alberg, M. (2001). Two- and three-year results from the Memphis restructuring initiative. *School Effectiveness and School Improvement, 12*(3), 323–346.

Ross, S. M., Stringfield, S., Sanders, W. L., & Wright, S. P. (2003). Inside systemic elementary school reform: Teacher effects and teacher mobility. *School Effectiveness and School Improvement, 14*(1), 73–110.

Ross, S. M., Wang, L. W., Alberg, M., Sanders, W. L., Wright, S. P., & Stringfield, S. (2001, April). *Fourth-year achievement results on the Tennessee Value-Added Assessment System for restructuring schools in Memphis.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103–116.

Sanders, W. L. (1994, October 27). Message posted to Discussion of Tennessee Value Added Assessment System. Retrieved February 11, 2008, from http://epaa.asu.edu/tvaas.html

Sanders, W. L. (1998). Value-added assessment. *The School Administrator, 55*(11), 24–27.

Sanders W. L. (2000). Annual CREATE Jason Millman Memorial Lecture: Value-added assessment from student achievement data: Opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14*(4), 329–339.

Sanders, W. L. (2003, April). *Beyond "No Child Left Behind."* Paper presented at the annual meeting of the American Educational Research Association, Chicago. Retrieved February 10, 2007, from http://www.sas.com/govedu/edu/no-child.pdf

Sanders, W. L. (2004). *How can value-added assessment lead to greater accountability?* [Biographical sketch with article]. Report presented at First Annual Policy Conference of the New York State Educational Conference Board (Investment and Accountability for Student Success), Albany, NY. Retrieved February 10, 2007, from http://www.nysecb.org/2004conference/04sanders.html

Sanders, W. L., Ashton, J. J., & Wright, S. P. (2005, March 7). *Comparison of the effects of NBPTS* [National Board for Professional Teaching Standards] *certified teachers with other teachers on the rate of student academic progress.* Arlington, VA: National Board for Professional Teaching Standards. Available at http://www.nbpts.org/UserFiles/File/SAS_final_NBPTS_report_D_-_Sanders.pdf

Sanders, W. L., & Horn, S. (1994). The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*(3), 299–311.

Sanders, W. L., & Horn, S. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247–256.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee Value-Added Accountability System: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137–162). Thousand Oaks, CA: Corwin Press.

Sanders, W. L., Saxton, A., Schneider, J., Dearden, B., Wright, S. P., & Horn, S. (2002). *Effects of building change on indicators of student achievement growth: Tennessee Value-Added Assessment System.* Knoxville: University of Tennessee Value-Added Research and Assessment Center.

SAS. (2007). *Dr. William L. Sanders* [Biographical sketch]. Retrieved March 16, 2007, from http://www.sas.com/govedu/edu/bio_sanders.html

Stroup, W. (1995). *Assessment of the statistical methodology used in the Tennessee Value-Added Assessment System.* Knoxville: Tennessee Value-Added Research and Assessment Center.

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., et al. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11–36.

Thompson, B. (2006). Research synthesis: Effect sizes. In J. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 583–603). Washington, DC: American Educational Research Association.

Topping, K. J., & Sanders, W. L. (2000). Teacher effectiveness and computer assessment of reading: Relating value-added and learning information system data. *School Effectiveness and School Improvement, 11*(13), 305–337.

U.S. Department of Education. (2006a, May 17). *Secretary Spellings approves Tennessee and North Carolina growth model pilots for 2005–2006.* Retrieved March 14, 2007, from http://www.ed.gov/news/pressreleases/2006/05/05172006a.html

U.S. Department of Education. (2006b, November 9). *Secretary Spellings approves additional growth model pilots for 2006–2007.* Retrieved March 14, 2007, from http://www.ed.gov/news/pressreleases/2006/11/11092006a.html

Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004, September 8). National Board certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46). Retrieved January 20, 2005, from http://epaa.asu.edu/epaa/v12n46/

Walberg, H. J., & Paik, S. J. (1997). Assessment requires incentives to add value: A review of the Tennessee Value-Added Assessment System. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 169–178). Thousand Oaks, CA: Corwin Press.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*(3), 461–481.

Wright, P., Horn, S., & Sanders, W. L. (1997). Teachers and classroom heterogeneity: Their effects on educational outcomes. *Journal of Personnel Evaluation in Education, 11*(1), 57–67.

AUTHOR

AUDREY AMREIN-BEARDSLEY is an assistant professor at Arizona State University, College of Teacher Education and Leadership, P.O. Box 37100, Phoenix, AZ 85069; *audrey.beardsley@asu.edu.* Her areas of interest are educational policy, research methods, and tests and assessments.

Manuscript received October 20, 2007
Revision received January 28, 2008
Accepted February 1, 2008