Center for
K–12 Assessment
& Performance Management

An independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

Exploratory Seminar:

Measurement Challenges Within the Race to the Top Agenda

December 2009

Value-Added Models and the Next Generation of Assessments

Robert H. Meyer and Emin Dokumaci

Created by Educational Testing Service (ETS) to forward a larger social mission, the Center for K–12 Assessment & Performance Management has been given the directive to serve as an independent catalyst and resource for the improvement of measurement and data systems to enhance student achievement.

Value-Added Models and the Next Generation of Assessments[1]

Robert H. Meyer and Emin Dokumaci

Value-Added Research Center, University of Wisconsin–Madison

This paper was presented by Robert H. Meyer at the Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda, December 2009. Download copies of other papers presented at the seminar at http://www.k12center.org/publications.html.

This paper discusses some of the fundamental features of value-added models, with particular focus on the interaction between the design and interpretation of value-added models and the design and properties of student assessments. We present a case study using actual state and district data from Wisconsin.

A value-added model is a quasi-experimental statistical model that yields estimates of the contribution of schools, classrooms, teachers, or other educational units to student achievement (or other student outcomes), controlling for other (non-school) sources of student achievement growth, including prior student achievement and student and family characteristics. The model produces estimates of school productivity—value-added indicators—under the counterfactual assumption that all schools serve the same group of students. This facilitates apples-and-apples school comparisons rather than apples-and-oranges comparisons. The objective is to facilitate valid and fair comparisons of student outcomes across schools, given that the schools may serve very different student populations.

A useful (high-quality) value-added model produces indicators of educational productivity at multiple levels of the educational system that are valid and reliable (in the sense of accurately measuring educational productivity). The degree to which a particular value-added system produces high-quality value-added indicators depends directly on five major factors:

1. The quality and appropriateness of the student outcomes (for example, mathematics and reading achievement[2]) used to measure value-added productivity. In particular,

[1] Although we are responsible for the views and analyses presented in this paper, we particularly thank our colleagues at the Value-Added Research Center (VARC) and Deb Lindsey, Director of Research and Assessment in the Milwaukee Public Schools. Our work has benefited from extensive interactions with our district and state research partners, including the public school systems in Chicago, Madison, Milwaukee, Minneapolis, New York City, and Racine, the state of Wisconsin, and the Wisconsin Cooperative Educational Service Agencies. This paper draws on research funded by the Wisconsin Department of Public Instruction, the Milwaukee Public Schools, the Joyce Foundation, the Walton Family Foundation, and the Institute for Education Sciences, U.S, Department of Education, through Grant R305D100018. The opinions expressed are those of the authors and do not represent views of the funding organizations.

these outcomes need to be curriculum sensitive; that is, capable of measuring the contributions of teachers, programs, and policies.[3]

2. The availability and quality of longitudinal data on students, teachers, and schools, particularly the degree to which students, classrooms/courses, and teachers are correctly linked.
3. The design of the value-added model (or models) used to produce measures of value-added productivity (and associated measures of the statistical precision of productivity). Our objective is to develop models that yield productivity estimates with low mean squared error (MSE).
4. The volume of data available to estimate the model.[4]
5. The degree to which the student outcomes (and other variables included in a value-added model) are resistant to manipulation or distorted measurement.[5]

We believe that it is possible to produce valid, reliable, and useful measures of educational productivity if sufficient attention is paid to addressing the five factors listed above. Our focus in this paper is primarily on factors one and three, the design and properties of assessments and the design and assessment requirements of value-added models. Our objective is to identify issues and priorities for enhancing the quality of value-added models and indicators with respect to these two factors.

[2] Although value-added systems have generally measured educational productivity using student test scores, VARC has recently worked with the Milwaukee Public Schools to develop value-added models for non-assessment outcomes, such as student attendance.

[3] The second part of this paper addresses the issues of the quality and appropriateness of student outcomes in much greater detail.

[4] At the micro (unit) level, the precision of value-added estimates depends directly on the number of student observations available for a given educational unit. The amount of usable data can be increased by pooling data for given educational units over time. At the macro level, the interpretative utility of value-added indicators depends on which classrooms, schools, districts, and (possibly) states are included in the data used in the analysis. Estimates based on a wide "reference group" (for example, a statewide database) are typically much more useful that those based on a narrow reference group (for example, a small district). Later in this paper we illustrate the utility of estimates derived from a statewide system.

[5] Accountability systems (including value-added systems or attainment/proficiency-based systems) based on assessments that are open to manipulation could distort the incentives (and thus behavior) of educational stakeholders. For example, if students are tested using the same test form year after year, narrow teaching to the test form (and not the content domain that underlies the test) could be effective in raising test scores without actually increasing student achievement. In principle this problem can be addressed by changing test forms each year. As discussed later in the paper, this requires test developers to equate test forms horizontally if the desire is to produce test scores that can be validly compared over time.

Value-Added Analysis in Context

As a prelude to addressing the specific objectives of this paper, we briefly discuss some of the larger issues related to the appropriate use of value-added indicators in an education system.

In order for a value-added system to be a powerful engine of school improvement, it should be systemically aligned with the fundamental needs and operations of schools, districts, and states, in ways that include at least the following:[6]

- Be used to evaluate effectiveness of instructional practices, programs, and policies.
- Be embedded within a framework of data-informed decision making.
- Be aligned with school, district, and state policies, practices, and governance procedures.
    - Vertical alignment: alignment across all levels of the system, including state, district, cascade of district management levels, school (multiple grades), grade-level team, classroom, teacher, student subgroup, and student.
    - Horizontal alignment: alignment across departments and divisions at each level (e.g., teaching and learning, human resources, and accountability).
- Provide extensive professional development to support understanding and application of value-added information.

A well-developed and aligned value-added system can be used to stimulate school improvement in several different ways, including when it can:

Provide evidence that schools can generate high student achievement growth (that is; high value-added productivity) even if they predominantly serve students with low prior achievement.[7]

- Facilitate triage by identifying and providing assistance to low-performing schools or teachers.
- Contribute to district knowledge about what works (including professional development).
- Be incorporated within a performance management system.
- Hold educational stakeholders accountable for performance.
- Provide bonuses to high-performing teachers, teams of teachers, and schools.

[6] In our work at VARC we have explored the connections between value-added systems and school district needs and operations in the following reports: Carl, Cheng, and Meyer (2009); Jones, Geraghty, Nisar, Mader, and Meyer (2010); and Lander, Keltz, Pautsch, Carl, Geraghty, and Meyer (2009).

[7] Although it is beyond the scope of this paper to address the mechanisms by which value-added indicators can be used to stimulate school improvement, we later will discuss how value-added systems support the premise that all schools can be highly productive .

- Provide information to teacher preparation institutions on the value-added performance of the teachers they have trained.[8]

In most applications it is essential to use value-added information in conjunction with other sources of information, such as observational data (based on well-defined rubrics) or value-added information based on multiple student outcomes.[9] In addition it typically is sensible to use information from multiple years in order to dampen variability due to statistical noise and authentic variation in educational outcomes. Reliance on a single value-added indicator, as opposed to multiple indicators, could provide educators with an incentive to focus their efforts to improve measured student performance in too narrow a manner.

Finally, it is important to contrast the appropriateness of using value-added indicators as measures of educational productivity versus attainment/proficiency indicators of the type currently required under the No Child Left Behind (NCLB) legislation. We address this issue later in the paper after providing a fuller description of a value-added model.

Description of a Simple Value-Added Model

Since our objective is to discuss conceptual issues in the design and interpretation of value-added models and assessments, we intentionally focus on a relatively simple statewide (multi-district) value-added model of school productivity at a given grade level. We then discuss options for making the model more complex. Most if not all value-added models (including classroom and teacher value-added models) produce value-added parameters of the type included in this model.

The key features of the model are:

- Two years of (consecutive grade) longitudinal assessment data for each student (measured annually at the end or beginning of the school year).[10]

[8] VARC is currently participating in a 10-year project in Minnesota, North Dakota, and South Dakota, funded by the Archibald Bush Foundation, to produce value-added measures of teacher performance and provide them to teachers, teacher preparation institutions, and districts. For additional information on the Teacher Effectiveness Initiative see Bush Foundation (n.d.).

[9] As an example, schools participating in the TAP system (The System for Teacher and Student Advancement) use a blend of teacher/classroom-level value-added ratings, school-level value-added ratings, and observational ratings (based on a well-defined rubric) to rate teacher performance.

[10] Note that since statewide testing begins in third grade in many states, only 2 years of (up-to-date) attainment data are typically available to estimate value-added models of achievement growth from third to fourth grade. In later grades, where additional years of longitudinal data are available, it is possible to expand the two-period model to include multiple grades—for example, a model of achievement growth from third to fourth to fifth grade. In a two-period model differences across schools in student growth trajectories are captured directly by the student characteristics that are included in the model. In this model, systematic differences in student growth trajectories that are not captured by student characteristics included in the model are absorbed by the estimated value-added effects. One of the key advantages of including three or more achievement outcomes for each student (when those data are available) is that it is possible to better control for differences across schools in the

- School/district value-added productivity effects η_{klt} (for school *k* in district *l* in year *t*, at a given grade).
- Statewide value-added productivity effects π_t.
- A posttest-on-pretest link parameter λ_t (which may vary across grades and over time). This parameter allows for the possibility that achievement growth may differ for students with high and low prior achievement and in situations where the distribution of the posttest and pretest variables may be non-uniform over time (more on this below).
- Demographic variables X_{it} to capture differences across students (within-classrooms) in achievement growth.

Figure 1 provides a schematic diagram of this two-period value-added model.

The model indicates that achievement at the end of a period (posttest Y_2) is the sum of:

1. Student achievement at the beginning of the period (pretest Y_1) times a posttest-on-pretest link parameter (λ).[11]
2. Student growth that is correlated with student characteristics such as income, English language learner (ELL) status, and race/ethnicity (β).

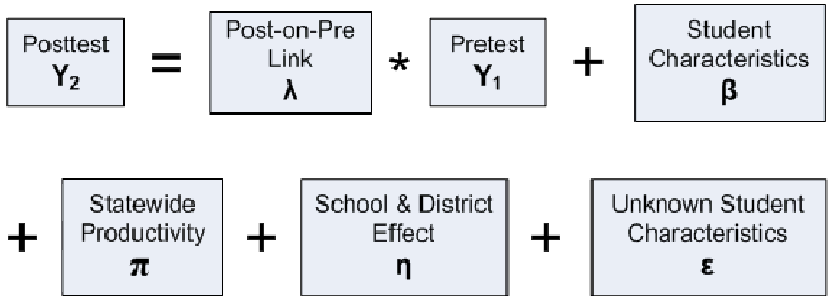

Figure 1. Diagram of two-period state value-added model.

student-level determinants of achievement growth than in a model based on two achievement outcomes. See the following for discussions of alternative value-added models: Ballou, Sanders, and Wright (2004); Boardman and Murnane (1979); Hanushek et al. (2005); McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004); Meyer (1996, 1997, 2004, 2008); Rothstein (2007); Sanders and Horn (1994); and Willms and Raudenbush (1989).

[11] As discussed in this paper, the value-added model could be extended to include measures of prior achievement in multiple subject areas, as well as measures of noncognitive outcomes. A model of mathematics achievement, for example, could include prior reading achievement as well as prior mathematics achievement. We have found that including multiple measures of prior student outcomes typically increases the predictive power of a model and improves the accuracy of estimated value-added effects.

3. Statewide productivity (π ; see further explanation below).
4. School and district productivity (η) (see further explanation below).
5. Student growth that is due to unknown student characteristics and random test measurement error (ε).

The value-added productivity parameters produced by this model are defined in greater detail in Table 1.[12]

Table 1. Value-Added Effect Parameters

Parameter	Definition
π_t	*Statewide productivity* in year *t* (for a given grade). Note that this parameter can be interpreted as a genuine statewide productivity effect only if test scores are accurately horizontally equated over time so that changes in test score growth do not reflect test form effects (the issue of horizontal equating is discussed later in this paper). This parameter is typically estimated as a contrast effect relative to a baseline year. In this case statewide productivity is equal to 0 in the baseline year, and productivity in other years is measured relative to productivity in the baseline year.
η_{klt}	*Relative school productivity* for school *k* in district *l* in year *t* (for a given grade). This parameter is referred to as a relative value-added parameter because it is centered around 0 in each year, so that the average school in the district has a value-added rating equal to 0 and school productivity is measured relative to the average school. Changes in statewide productivity are thus absorbed by the parameter π_t .
$\eta_{klt}^{ABSOLUTE} = \pi_t + \eta_{klt}$	*Absolute (total) school, district, and state productivity*. This indicator incorporates relative school productivity plus overall changes in statewide productivity, provided (as mentioned above) that test scores are accurately horizontally equated.

The school/district productivity parameter defined above (either η_{klt} or $\eta_{klt}^{ABSOLUTE}$) is referred to as the beat the average (BTA) rating in the Milwaukee value-added system and the beat the odds rating in the Minneapolis value-added system, because the value of the indicator equals the amount by which it exceeds or falls short of average district productivity in each year (in the case of η_{klt}) or in the baseline year (in the case of $\eta_{klt}^{ABSOLUTE}$).

[12] Appendix A presents the model using formal statistical notation and defines the model parameters and variables.

The simple value-added model presented above provides a framework for discussing issues related to model design and assessment design and their interaction.

The Design of Value-Added Models

We begin by discussing two basic features of the two-period model: (a) the provision for connecting post and prior achievement via a post-on-pre linking parameter and (b) the inclusion of student-level demographic variables in the model.

Post-on-Pre Link: The Coefficient on Prior Achievement[13]

One of the important features of the value-added model considered above is that it allows for the possibility that the coefficient on prior achievement (λ_t) could differ across grades and years and might not equal 1, a parameter restriction that is imposed in some value-added models. The model would be simpler to estimate if it were appropriate to impose the parameter restriction $\lambda = 1$, but there are at least four factors that could make this restriction invalid. First, λ could be less than 1 if the stock of knowledge, skill, and achievement captured by student assessments is not totally durable, but rather is subject to decay. Second, λ could differ from 1 if school resources are allocated differentially to students as a function of prior achievement. If resources were to be tilted relatively toward low-achieving students—a remediation strategy—then λ would be reduced. The opposite would be true if resources were tilted toward high-achieving students. Third, λ could differ from1 if posttest and pretest scores are measured on different scales, perhaps because the assessments administered in different grades are from different vendors and scored on different test scales or due to instability in the variability of test scores across grades and years. In this case, the coefficient on prior achievement partially reflects the difference in scale units between the pretest and posttest. Fourth, the different methods used to scale assessments could in effect transform posttest and pretest scores so that the relationship between post and prior achievement would be nonlinear. In this case a linear value-added model might still provide a reasonably accurate approximation of the achievement growth process, but the coefficient on prior achievement (as in the case of the third point) would be affected by the test scaling. See Meyer, Dokumaci, Morgan, and Geraghty (2009) for discussion of these issues.

[13] Although our focus in this paper is primarily on model design issues rather than statistical estimation techniques, it is important to point out that in order to properly estimate the value-added model presented in the text it is necessary to account for measurement error in test scores (using methods from structural equation modeling) and possible endogeneity due to correlation of prior achievement with the equation error term. We discuss the issue of test measurement error later in this paper. Fuller (1986) and Meyer (1992, 1999) discussed methods for correcting for measurement error. VARC has used these methods to control for test measurement error in all of its value-added systems. Methods for addressing the endogeneity of prior achievement have been developed by numerous researchers, including Anderson and Hsaio (1982). Arellano and Honore (2001) provided a recent survey of this literature.

In summary, there are four factors that could make it problematic to impose the parameter restriction that the coefficient on prior achievement (λ_t) be identical in all grades and years and equal to a particular value (such as 1): (a) durability/decay in achievement, (b) differential resource allocation, (c) differences in the pretest and posttest test scales, and (d) nonlinearity in the test scaling algorithm.

In the case study presented later in this paper we explore whether the standard deviations of prior and post achievement test scores exhibit instabilities across grades and over time (as discussed above).

Does Achievement Growth Differ for Students With Different Student Characteristics?

An important feature of the value-added model presented above is that it includes explicit measures of student characteristics.[14] Since most district or state value-added systems are based on administrative databases (as opposed to special-purpose data collections), value-added models generally include a limited number of student measures—for example, poverty status (participation in free or reduced-price lunch), participation in special education, participation in an ELL program, gender, or race/ethnicity.

Including measures of student characteristics in a value-added model serves two purposes. First, including these measures makes it possible to measure district or statewide differences in achievement growth by student subgroups (e.g., low vs. high poverty). We refer to these differences as *value-added growth gaps*. They are analogous to attainment gaps. Growth gaps are in some ways more fundamental than attainment gaps, because attainment gaps arise via year-to-year accumulation of growth gaps. These statistics are important for policy purposes; over time a district can monitor changes in growth gaps to evaluate the success of policies and programs designed to reduce inequality in student attainment and student growth.[15]

The second purpose of including characteristics in a value-added model is to control for differences in student composition across schools so that estimates of educational performance reflect differences in school productivity, rather than differences in school composition. In other words, control variables (including prior achievement) are included in the model to achieve, to the extent possible, apples-and-apples school comparisons rather than apples–and-oranges comparisons.[16] Models that fail to include student-level variables will yield results systematically biased against schools and educators that

[14] It is common practice to include student demographic variables in statistical models of student test scores. For an early reference, see the well-known study by Coleman (1966).

[15] The value-added model presented above allows for state- or district-level growth gaps (by student subgroups) and changes over time in these growth gaps. We have developed a generalized value-added model (which we refer to as a *differential effects* value-added model) that captures differences in growth gaps (by student subgroups) across schools, classrooms, and teachers (and over time). Working with our district partners, we have applied this model in Chicago, Milwaukee, and New York. For additional information, see Dokumaci and Meyer (2010).

[16] In the model considered in this paper we have included student-level measures of student characteristics but have not included school-level (or classroom-level) measures of these variables or other school-level variables—for example, the proportion of students in poverty (by school). We discuss the option of including school-level variables in a value-added model in Appendix C. See Meyer (1996, 1997) and Willms and Raudenbush (1989) for further discussion of this issue.

disproportionately serve students who, on average, exhibit relatively low within-classroom and within-school achievement growth (for example, low-income students). In an era where public policy is focused on providing high-quality teachers for all students, it seems particularly unwise to build educational productivity indicators that are biased against exactly the types of students society is most eager to help.

Some analysts contend that despite the arguments in favor of including student-level demographic variables in a value-added model, doing so (or even including prior student achievement) could lead to reduced achievement expectations for subgroups with relatively low achievement or achievement growth.[17] We strongly reject this contention (although we believe that it needs to be fully and appropriately addressed). We suggest that it arises from a failure to recognize that measuring the productivity of schools, classrooms, and teachers is different from setting student achievement expectations (or standards) and measuring whether students have met those expectations. There are two dimensions to this issue, not one. Attainment information can appropriately be used to identify students who do not satisfy standards and thus are in need of additional resources. Value-added information can appropriately be used to measure the productivity of schools attended by both low- and high-achieving students. There is nothing conceptually or practically difficult about addressing both dimensions simultaneously. Later in the paper we illustrate this point using data from the Milwaukee Public Schools.

Value-Added Productivity, Student Achievement Growth, and Student Attainment

In this section we highlight the differences and connections between value-added productivity, average student achievement growth (gain), and average student achievement (measured prior to and at the end of the school year). Note that average achievement is a measure quite similar to percent proficient, in that both indicators measure some feature of the level and distribution of student attainment. The connection between these indicators, given the value-added model presented above, is shown in Figure 2.

(1)		(2)		(3)		(4)
Gain: Average Growth	=	$(\lambda - 1)\left(\begin{array}{c}\text{Average Prior}\\ \text{Achievement}\end{array}\right)$	+	Average Growth Effect of Student Characteristics	+	Value-Added Productivity
(1)		(2)		(3)		(4)
Attainment: Average Posttest	=	$\lambda\left(\begin{array}{c}\text{Average Prior}\\ \text{Achievement}\end{array}\right)$	+	Average Growth Effect of Student Characteristics	+	Value-Added Productivity

Figure 2. The connections between value-added, gain, and attainment indicators.

[17] See Appendix C for a discussion of the merits of including classroom- and school-level variables in value-added models.

The gain (or growth) indicator differs from value-added productivity in two ways. One, it absorbs growth differences across schools due to differences in student characteristics, if any (column 3). Two, differences in average prior achievement across schools leak into average gain, if the coefficient on prior achievement (λ) does not equal 1 (column 2). These are conditions that can be checked empirically (as illustrated later in the paper). In our experience, estimates of the post-on-pre link parameter (λ) are generally less than 1.[18] In this case, average gain absorbs a negative fraction of average prior achievement, since the multiplier $(\lambda - 1)$ in the above equation is negative if $(\lambda < 1)$. The bottom line is that if there are empirically large differences between average gain and value-added indicators, then it is problematic to rely on the gain indicator as a valid measure of school productivity.

Similarly, average post achievement differs from value-added productivity in two ways. First, as in the case of the gain indicator, average post achievement absorbs growth differences across schools due to differences in student characteristics, if any. Second, average post achievement, as expected, absorbs differences in average prior achievement across schools as long as student achievement is a cumulative growth process, in which case ($\lambda > 0$). Unless average prior achievement and average student characteristics are identical across schools or are perfectly correlated with value-added productivity (all very unlikely circumstances), average post achievement, and other attainment indicators are highly inaccurate measures of school performance. Meyer (1996, 1997) presented additional evidence on why attainment indicators generally fail as measures of school performance.

Value-Added Information and Value-Added Reports

In this section we discuss approaches for reporting value-added information. We begin by considering reports for a single school district and then consider reports that feature comparisons of multiple districts.

District Value-Added Reports

Figure 3 is an example of a school report card from the Milwaukee Public Schools (MPS) that provides information on a school's value-added rating in reading and mathematics defined using two different metrics. The BTA metric is equivalent to *relative school productivity* as defined in Table 1.[19] This indicator is centered about at 0, so that the average school in the district has a relative value-added rating equal to 0. The indicator is expressed in units that are identical to the units of student achievement measured at the end of the school year (typically scale score units). The performance tier metric is a standardized measure of value-added that generally ranges from 0 to 6, with a district mean equal to 3. The tier rating

[18] As discussed in a previous footnote, we assume that the parameters of the value-added model have been estimated using an estimation strategy that controls for measurement error in prior achievement. Failure to control for test measurement error yields estimates of the post-on-pre link parameter that are biased downward, unless the level of test measurement error is small (say, less than 5%).

[19] The information reported in this figure applies to all students in Grades 3 to 5, the elementary school grades covered by the Wisconsin state assessment. The value-added ratings are thus the average of grade-level value-added indicators for Grades 3 to 5.

is equal to $3 + \text{BTA} / SD$, where *SD* is equal to the standard deviation (corrected for estimation error) of BTA value-added. In other words, the tier metric is equal to a *z*-statistic, with 3 added so that typical values are positive.[20] As an example, a school with a tier rating equal to 4 has a relative value-added rating that is one standard deviation greater than the average school. The tier metric is relatively easy to use and interpret because it does not require knowledge of the units in which student achievement is measured, as in the case of the BTA metric.

MILWAUKEE PUBLIC SCHOOLS

MPS School Report Card

School Elementary

Value-Added Growth Analysis - Elementary School Grades

Mathematics

Year	Beat the Average Performance (Mathematics Scale)	Performance Tier
2006-07 to 2007-08	7.2	4.6
2005-06 to 2006-07	5.4	3.6
2004-05 to 2005-06	2.1	3.1

Year / Tier	0	1	2	3	4	5	6
2006-07 to 2007-08							
2005-06 to 2006-07							
2004-05 to 2005-06							

Reading

Year	Beat the Average Performance (Reading Scale)	Performance Tier
2006-07 to 2007-08	-1.5	2.7
2005-06 to 2006-07	-8.2	1.7
2004-05 to 2005-06	-5.7	2.1

Year / Tier	0	1	2	3	4	5	6
2006-07 to 2007-08							
2005-06 to 2006-07							
2004-05 to 2005-06							

Data is based on WKCE and Terra Nova Scale Scores
Tier 0 = Well below district average growth
Tier 3 = Average growth
Tier 6 = Well above district average growth

Value-Added and Attainment Data Status Over Seven Years

		Value-Added Performance Tier							Achievement WKCE Prof./Adv. Across Grades							1 = High Value-Added High Attainment 2 = High Value-Added Low Attainment 3 = Low Value-Added High Attainment 4 = Low Value-Added Low Attainment						
Subject	Elem. Level	01-02	02-03	03-04	04-05	05-06	06-07	07-08	01-02	02-03	03-04	04-05	05-06	06-07	07-08	01-02	02-03	03-04	04-05	05-06	06-07	07-08
Mathematics	School	2.4	2.7	2.5	2.6	3.1	3.6	4.6	35%	37%	38%	38%	41%	49%	53%	4	4	4	4	2	1	1
	MPS	3.0	3.0	3.0	3.0	3.0	3.0	3.0	41%	46%	53%	46%	42%	48%	49%							
Reading	School	2.5	2.6	2.2	2.3	2.1	1.7	2.7	43%	43%	49%	52%	52%	53%	54%	4	4	4	4	4	4	4
	MPS	3.0	3.0	3.0	3.0	3.0	3.0	3.0	54%	62%	67%	62%	61%	63%	61%							

Figure 3. Example of school report card with value-added and attainment information.

[20] If value-added indicators for a given sample of schools (for example, a district or state) were approximately normally distributed, then the distribution of schools would follow a bell-shaped curve: Roughly 68% of schools would have a value-added tier rating between 2 and 4; another 27% would have tier ratings between 1 to 2 and 4 to 5; finally, 5% of the schools would have tier ratings less than 1 or greater than 5.

Note that the BTA and tier ratings can be used to report relative or absolute value-added productivity. As discussed previously, the difference between the two types of indicators is that relative value-added is centered around the same mean year after year (0 in the case of BTA, 3 in the case of tier ratings). Absolute value-added, in contrast, is centered around mean productivity for a given baseline year. Absolute value-added ratings are not restricted to a pre-specified range, but could shift to the right (in the case of an overall increase in educational productivity) or to the left (in the case of an overall decrease in educational productivity). The MPS school report card reports relative value-added ratings for the most recent three school years in both table and chart formats. (We will discuss the decision to report relative, as opposed to absolute, value-added ratings later in this paper.) The precision of the ratings (as captured by conventional confidence intervals) is also reported in the charts. Attainment information—the percentage of students who are proficient or advanced—is reported in the bottom of the table.[21]

The value-added information in Figure 3 describes a school that exhibited substantial growth in mathematics productivity over the past 3 years, going from tier ratings initially close to the average level (or below) up to a high of 4.6. In contrast, value-added ratings in reading were consistently below average. The value-added ratings were estimated with sufficient precision so that it is possible to reject the null hypothesis that mathematics productivity in the most recent year was equal to the district average. Despite the fact that mathematics value-added was quite high in the most recent year, the percent of students who met the proficiency standard at the end of the school year was only 53%, somewhat greater than the district average of 49% for that year. These two pieces of information tell an important story: This is a school that serves students with low incoming achievement (relative to established proficiency standards) but has managed to dramatically improved its productivity in mathematics over the past 3 years. With respect to mathematics achievement, the attainment information signals that it would be appropriate to provide students in this school with additional resources to spur growth in mathematics achievement (for example, summer school or after-school instruction or tutoring). On the other hand, the value-added information indicates that low productivity in mathematics is not the source of the achievement deficit. Indeed, this is a school in which policies should be directed at sustaining excellence with respect to mathematics instruction and pushing the quality of mathematics instruction to the next level (a tier level greater than 5). With respect to reading achievement, the evidence suggests a need to improve productivity substantially. The priorities for the school and the district are to replicate in the area of reading the evident turnaround in mathematics instruction.

[21] Note that value-added/post-attainment quadrant information is reported in the lower-right part of the report card. Schools are assigned to one of four quadrants depending on whether they have value-added ratings and post-attainment ratings that are above or below the average for the district. For example, a quadrant 1 school has value-added and post-attainment ratings that are both above average. Figure 5 provides information on quadrant status for all MPS elementary schools.

One of the key characteristics of the school report card depicted in Figure 3 is that it presents data focused on a single school (although the information is reported in such a way that it is clear where a school stands relative to the district average). Below, we consider a second method for reporting value-added and attainment information that provides a more holistic view of productivity across all schools in the district. In Figures 4 and 5 we present two-dimensional value-added/attainment graphs based on information from the MPS for mathematics and reading for a single year. Each data point represents value-added and attainment data for a single school.[22] Figure 4 presents information on value-added productivity and incoming (prior) achievement (measured in terms of a proficiency rate). Figure 5 presents information on value-added productivity and end-of-year achievement (again, measured in terms of a proficiency rate).

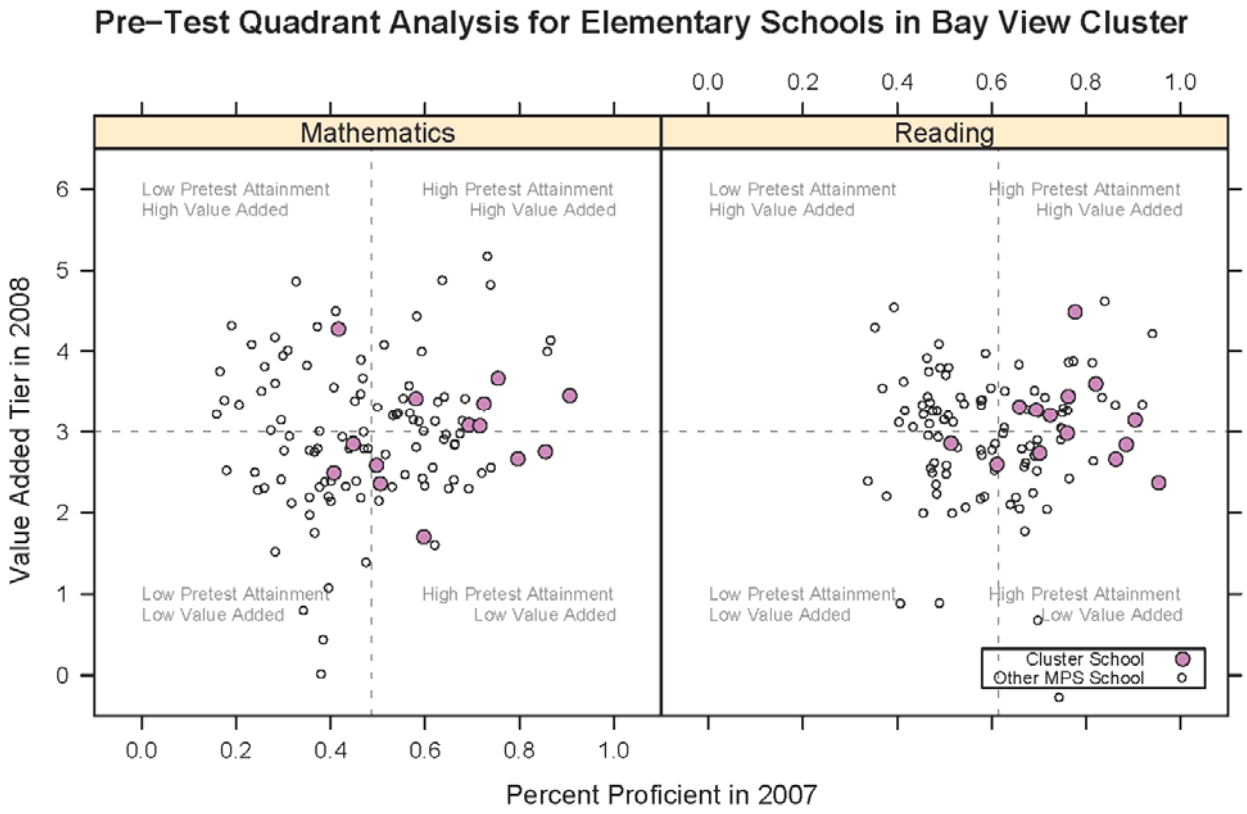

Figure 4. Two-dimensional prior attainment and value-added graph. Larger dots denote schools located in the Bay View cluster of schools in Milwaukee; smaller dots denote schools not in this cluster.

[22] In Figures 4 and 5 the two types of schools are distinguished in part to permit district staff to utilize this data more efficiently. Charts of this kind can be included in district and state performance management systems. Additional data features can be incorporated into the graphs, such as the capacity to drill down to obtain additional information on the students enrolled in each school.

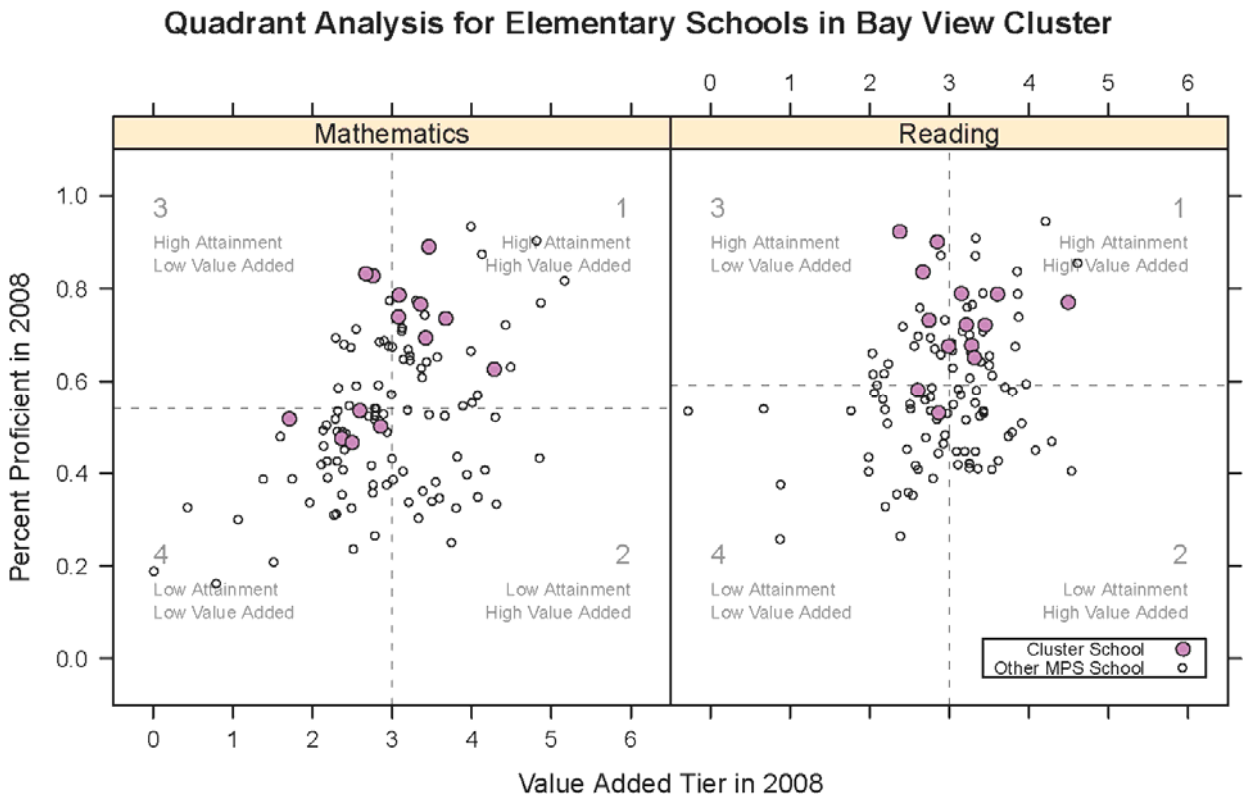

Figure 5. Two-dimensional value-added and post-attainment graph. Larger dots denote schools located in the Bay View cluster of schools in Milwaukee; smaller dots denote schools not in this cluster.

The data in Figure 4 convey a powerful message: Schools can generate high student achievement growth (that is; high value-added productivity) even if they predominantly serve students with low prior achievement. The data on the left side of the figure should serve as a wake-up call to schools that have given up on the premise that all students can learn. The proof of the proposition is clearly indicated on the graph—schools in the Milwaukee system can generate very high value-added growth in student achievement despite serving students with very low incoming achievement. These high value-added/low prior attainment schools are located in the upper-left part of the graph. The graph also reveals that some schools serving high-achieving students (the schools on the right side of the graph) have failed to generate high value-added growth. These schools are effectively coasting and may not even be aware of it. Graphs of this type can be used by school, district, and state staff to (a) identify low- and high-performing schools (that is, schools that need to turn around their performance or sustain excellent performance, respectively) and (b) provide concrete evidence that high performance is an attainable, realistic option for all schools. NCLB, with its exclusive focus on student attainment, is poorly equipped to transmit this message and the strong school performance expectations that go with it.

Figure 5 provides a policy-relevant complement to the preceding graph. Whereas Figure 4 displays information on value-added productivity and incoming achievement, Figure 5 displays information on

value-added productivity and end-of-year achievement. This graph is useful for guiding resource allocation decisions designed to ensure that all students meet high achievement standards, as required by NCLB. Student achievement in high–value-added schools may be unacceptably low if students enter these schools with very low achievement. Student growth could be accelerated for these students by providing additional resources, such as summer school and after-school instruction and tutoring. On the other hand, transferring a low-achieving student from a high-performing school to another school in the district (or reconstituting that school) would probably not improve the learning environment for that student, but rather would worsen it.

In summary, two-dimensional value-added/attainment graphs can provide school and district staff with information that can be credibly used to set high school performance expectations (or standards) and guide efficient allocation of resources to at-risk students.

Value-Added Reports for Multiple Districts: The Power of Statewide Comparisons

All of the reports discussed in the previous section can of course be produced using value-added and attainment information derived from a statewide system. There are several advantages to a statewide system. First, a system based on all of the districts in a state includes many more schools than a single-district system and thus many more opportunities to establish concrete examples of high value-added productivity. In general, the observable frontier (i.e., the maximum observed level) of value-added productivity tends to increase as the size of a *reference group* increases. This is important from a practical policy perspective: It undoubtedly is much easier to establish ambitious productivity standards if policy makers can show that the standard has actually been realized by a school, classroom, or teacher in a given reference group.

Second, in a statewide system it is possible to compare districts with respect to two dimensions: average value-added productivity (across all schools and classrooms in the district) and the consistency of value-added productivity across these entities. Consistency could be measured as the standard deviation of value-added productivity or as the percentage of entities having value-added productivity greater than a specified standard. Districts with high average value-added productivity and high consistency could be said to have high quality control. This kind of information can easily (and fruitfully) be incorporated into district and state performance management systems.

Third, the overall productivity of a large reference group (such as a state system) is likely to change much more slowly over time than the overall productivity of a small reference group. This suggests that comparisons of performance indicators that are not horizontally equated[23] (more on this below) may reasonably be interpreted as providing (approximate) evidence of absolute changes over time.[24] Alternatively, it may be very policy relevant to measure relative performance within a large reference group, even when it is much less informative to do so within a small reference group.

[23] Horizontally equated test scores are measured on the same scale over time.

[24] This follows from the fact that relative changes in productivity are identical to absolute changes in productivity if there is no change in average absolute productivity.

We should note that the arguments supporting statewide, as opposed to single district, reference groups could equally be used to support the utility of multistate or national reference groups. Comparisons at this scale presumably would be facilitated by the development of state assessment consortia, if a common scale were used to measure achievement for all members of the consortium.

Below we present the estimates that illustrate the importance of cross-district comparisons of district mean productivity and the consistency of productivity (as captured by the district standard deviation). To simplify the analysis we focus on two of the largest districts in Wisconsin, Milwaukee, and Madison. We report estimates of relative value-added productivity (in the BTA metric), rather than absolute value-added productivity (as defined previously in this paper). (In a later section we consider whether it is feasible to report valid absolute value-added measures, given the assessment data available in Wisconsin.)

Estimates from the mathematics and reading value-added models are presented in Table 2. The table reports district mean value-added productivity, the standard error of that estimate, and the standard deviation of school productivity within each district. Figures 6 and 7 report the district mean and standard deviation, respectively, of value-added productivity for mathematics (the results for reading are quite similar). As indicated in Table 2 and Figure 6, district average productivity is higher in Madison than in Milwaukee at all grades and in both growth years, except for Grade 3–4 in Growth Year 1 for mathematics and Grade 7–8 in Growth Year 1 for reading. For example, an average school in Madison contributes 2.77 more points to a student's mathematics scale score than an average school in the state in Growth Year 1 from Grade 4–5. On the other hand, an average school in Milwaukee contributes 4.79 fewer points to a student's mathematics scale score than an average school in the state in Growth Year 1 from Grade 4–5. In contrast, the standard deviation of school productivity is generally much lower in Madison than in Milwaukee. In short, Madison is a more consistent provider of school productivity than Milwaukee.

In order to understand the interplay of differences between the two districts in the mean and consistency of productivity, it is useful to directly examine the distributions of estimated school productivity for each district. Figure 8 reports these distributions for school productivity in mathematics for Madison and Milwaukee for the 2006–2007 school year. A separate figure is presented for each grade.

The results are quite striking: The range of value-added productivity is much wider for Milwaukee than Madison (consistent with the result reported in Table 2 that the standard deviation of productivity is higher in Milwaukee than in Madison). Given the wider range of productivity in Milwaukee, it is apparent that the district has schools that are very low performing and schools that are very high performing. In fact, despite the fact that average productivity is higher in Madison than in Milwaukee, the highest performing Milwaukee schools tend to have somewhat higher productivity than the highest performing Madison schools. These data suggest that the MPS system needs to work aggressively to improve the performance of its lowest performing schools. This district has a significant quality control problem.

Table 2. District Value-Added Effects: Madison and Milwaukee

Grade	District	Growth Year 1: 2005–2006			Growth Year 2: 2006–2007		
		District average	Standard error	District standard deviation	District average	Standard error	District standard deviation
		Mathematics					
3	Madison	-3.48	0.59	3.56	0.78	0.61	8.16
4	Madison	2.77	0.59	5.60	-1.03	0.64	3.06
5	Madison	-0.95	0.60	5.85	3.84	0.59	5.75
6	Madison	0.62	0.50	5.37	2.06	0.51	2.93
7	Madison	2.53	0.58	3.05	0.66	0.43	2.59
3	Milwaukee	-0.66	0.40	11.39	-0.76	0.41	11.34
4	Milwaukee	-4.79	0.37	8.53	-4.83	0.43	9.54
5	Milwaukee	-6.22	0.39	10.22	-5.06	0.39	8.53
6	Milwaukee	-2.64	0.29	6.38	-2.32	0.36	6.57
7	Milwaukee	-0.14	0.39	6.53	-1.69	0.37	9.40
		Reading					
3	Madison	0.52	0.61	4.71	-0.49	0.63	4.92
4	Madison	3.36	0.61	3.45	2.59	0.60	4.91
5	Madison	0.90	0.65	5.33	0.82	0.63	2.83
6	Madison	1.01	0.63	5.75	0.91	0.64	2.88
7	Madison	1.35	0.64	5.63	1.32	0.54	4.24
3	Milwaukee	-1.89	0.41	7.69	-5.37	0.41	8.28
4	Milwaukee	-1.99	0.39	8.23	-4.33	0.41	8.16
5	Milwaukee	-4.17	0.42	8.63	-5.13	0.42	8.21
6	Milwaukee	-2.34	0.39	6.09	-2.77	0.43	6.08
7	Milwaukee	1.92	0.42	6.43	0.24	0.34	5.58

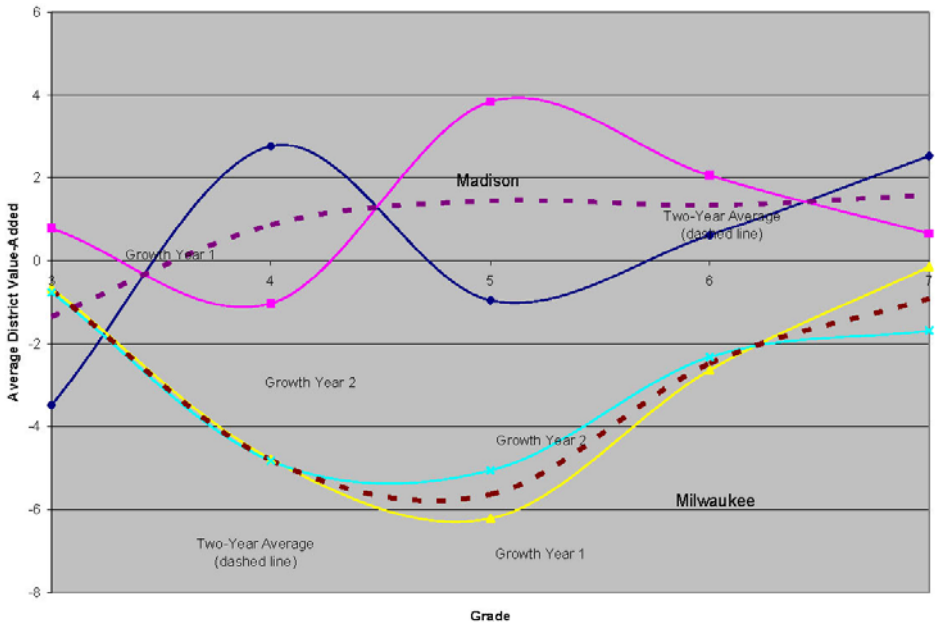

Figure 6. Average district value-added productivity in mathematics for Madison and Milwaukee by grade and year.

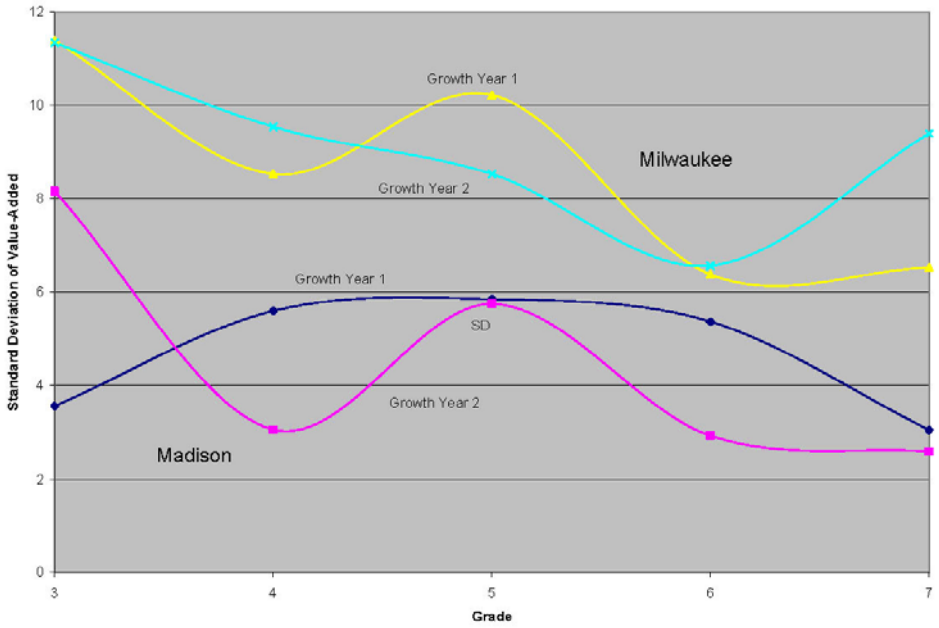

Figure 7. District standard deviation of value-added productivity in mathematics for Madison and Milwaukee by grade and year.

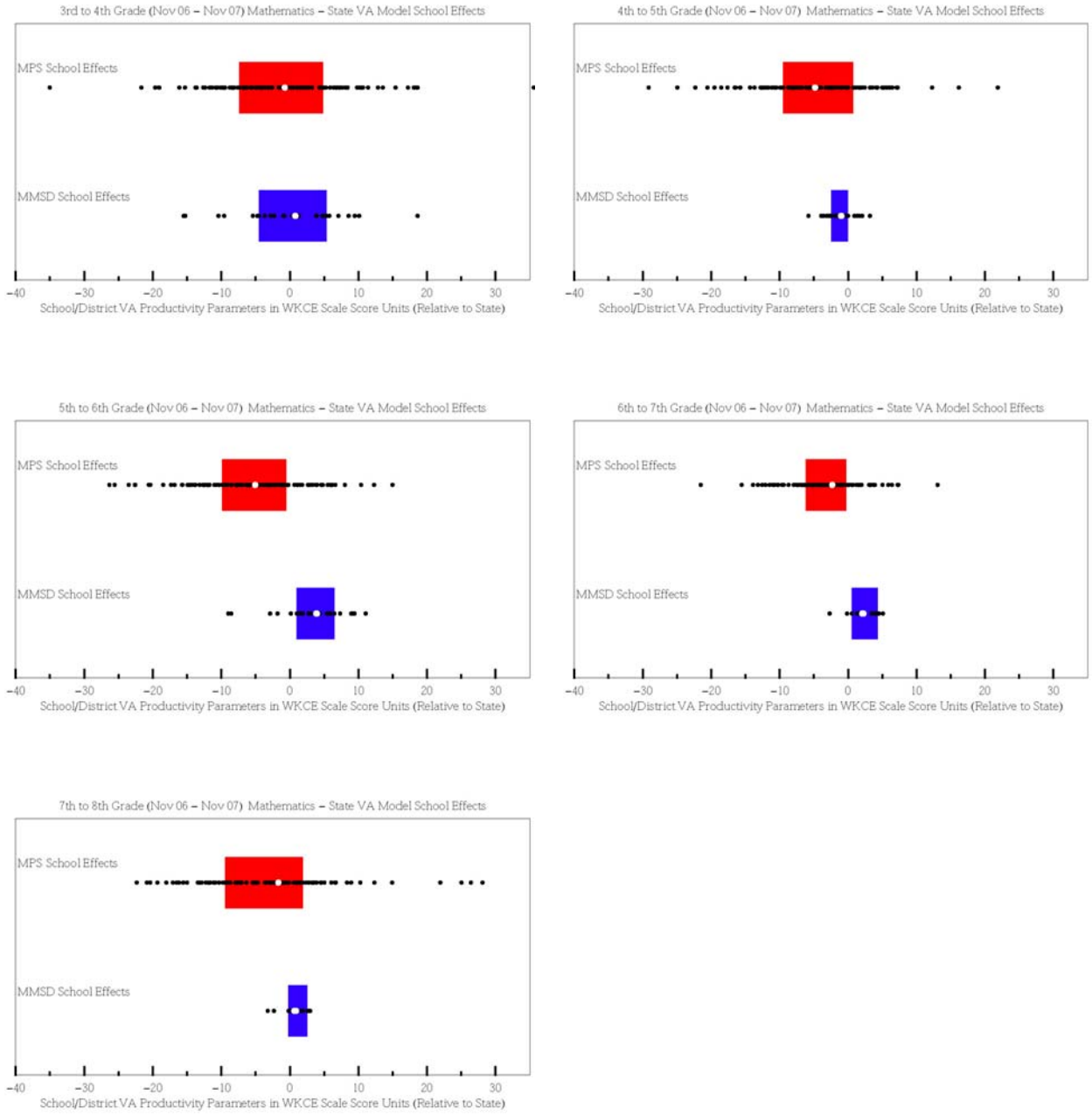

Figure 8. Distribution of value-added school productivity in mathematics in Milwaukee (upper plots in each chart) and Madison (lower plots in each chart), 2006–2007 school year. The left and right edges of the boxes in the graphs represent the 25th and 75th percentiles of the school productivity distribution in each district. The white dots represent the district averages, and the black dots represent individual school productivity effects. VA = Value-Added; WKCE = Wisconsin Knowledge and Concepts Examination.

We conclude this section with several observations. First, the current practice of using only status and/or simple growth models to analyze student growth limits a school or district's ability to attribute changes in student performance to specific programs or to the instruction provided by teachers. From a public policy perspective, value-added models provide the necessary statistical grounding to determine the causal reasons behind low and high performance more accurately, and thus can provide the information needed to identify the most effective school improvement strategies. Second, two-dimensional value-added/attainment graphs can provide school and district staff with information that can be credibly used to set high performance expectations (or standards) and guide efficient allocation of resources to at-risk students. Third, value-added information based on statewide or multistate data facilitates setting performance expectations, derived from actual data, at high levels. Finally, consistency is an important attribute of school district productivity and should be monitored along with average productivity.

In the remainder of this paper we discuss issues related to the design of value-added models and the design and properties of student assessments.

How Complex Should a Value-Added Model Be

In our work at the Value-Added Research Center (VARC), we have followed the following rule in designing and implementing value-added models: Simpler is better, unless it is wrong. This rule implies that designers of value-added models need to be vigilant in protecting against possible threats to validity and reliability. In practice this means considering possible model extensions and generalizations and engaging in rigorous diagnostic evaluation and testing of these model enhancements using actual (and possibly simulated) data. Some model enhancements (in some district and state contexts, but perhaps not others) may improve model validity and/or reliability. Others may turn out to have a limited impact on results; thus they can be either dropped from the model or retained to demonstrate that the model is robust to inclusion/exclusion of these enhancements.[25] In our work with Chicago, Milwaukee, and New York City, educational stakeholders have been very helpful in identifying possible threats to model validity. For example, a principal in the Chicago Public Schools stimulated a line of inquiry that ultimately led to collecting data on whether a student was homeless and including this student characteristic in the Chicago value-added model.

Although it is beyond the scope of this paper to provide an extensive discussion of important value-added model features, beyond those features included in the two-period model presented above, we describe three features below that we have found to be important in our work with districts and states.

Multiple-Year Longitudinal Test Data

In a two-period value-added model, differences in student growth trajectories across schools are captured directly by the student characteristics that are included in the model. Systematic differences in

[25] A model enhancement that does not significantly improve the validity of model estimates may actually reduce statistical precision. In this case it generally is optimal to drop the enhancement.

student growth trajectories that are not captured by student characteristics included in the model are absorbed by the estimated value-added effects (thus resulting in bias). One of the key advantages of including three or more achievement outcomes for each student (when those data are available) is that it is possible to better control for differences in the student-level determinants of achievement growth across schools than with a model based on two achievement outcomes. The advantage of having multiple years of longitudinal test data is much greater in classroom and teacher value-added models than in grade-level/school value-added models; this is because the benefits of having multiple years of data depend heavily on the degree of student mobility from unit to unit (classroom-to-classroom mobility tends to be much higher than school-to-school mobility). See the following for discussions of alternative value-added models: Ballou et al. (2004); Boardman and Murnane (1979); Hanushek, Kain, O'Brien, and Rivkin (2005); McCaffrey et al. (2004); Meyer (1996, 1997, 2004); Rothstein (2007); Sanders and Horn (1994); and Willms and Raudenbush (1989).

Student Mobility

Thus far we have not discussed how to measure value-added productivity in real-world situations where some (perhaps many) students change schools during the school year. It is customary when calculating NCLB proficiency rates to exclude students who have not been enrolled in a school for a full academic year. Gain (growth) indicators typically either exclude mobile students or pretend that growth can be fully attributed to the school that a student attended at a given point in time (for example, the date of the posttest). We strongly believe that it is problematic from a policy perspective to systematically exclude students from a measurement system that serves an evaluation and accountability function. Systematic exclusion of mobile students (or any other student group) from an accountability system creates an incentive for agents to allocate fewer resources to this group. Creating an incentive of this type is a bad idea even if we believe (as we do) that few educators would respond to this incentive.

It turns out that student mobility can be introduced into a value-added model with only a slight tweak in the traditional definition of the school variables included in a standard multilevel (hierarchical) model. In such a model a school variable is set to 1 if a student (*i*) attended a given school (*k*) during the school year and 0 otherwise. In order to accommodate students who changed schools during the school year, the school variable is redefined so that it measures the fraction of time that a student attended a given school during the school year. We refer to this variant of the value-added model as the dose model. Although this model requires more extensive student attendance data than a model that ignores student mobility, we have found that most districts have data warehouses that can support implementation of this model.

Differential Effect Value-Added

The conventional value-added model (including the model discussed thus far) imposes the restriction that a high-performing classroom or school (at a given grade level at a given point in time) is identically high performing for all types of students, including, for example, students with low and high prior achievement and low and high income status. If this assumption is approximately true, schools can validly be compared on the basis of a single performance indicator. However, this assumption might be

incorrect: A given school could be very effective for students with low prior achievement, for example, but less so with talented and gifted (TAG) students. These differences in effectiveness could stem from differences in the effectiveness of the multiple programs and courses offered by schools. For example, schools that provide tutoring, after school, and summer school programs for low-performing students, but no additional programs for TAG students, might be relatively more effective with low-achieving students than high-achieving students.

We have developed a generalized value-added model (which we refer to as a *differential effects* value-added model) that captures differences in value-added productivity (by student subgroups) across schools, classrooms, and teachers (and over time). Working with our district partners, we have applied this model in Chicago, Milwaukee, and New York. For additional information see Dokumaci and Meyer (2010).

The Intersection of Assessment Design and the Design and Interpretation of Value-Added Models

In this section we consider several aspects of the design and properties of student assessments, with particular focus on the interaction between the characteristics of assessments and the design and interpretation of value-added models. The analysis presented below draws on student assessment data from the state of Wisconsin.

Assessment Scales

Student achievement can be expressed using many different test scales, including development scale scores, percentile scores, and normal curve equivalents.[26] Many test developers use a developmental scale as the fundamental scale for building student test scores and then construct other scales and statistics from it. In this paper we report results using development scale scores.

Conventional value-added indicators, as well as average gain and average achievement indicators, exploit the full range of information contained within student scale scores. As a result it is important that scale scores are measured well along the entire distribution of scores—from the lowest to the highest scores. As is well known, assessments that fail to include a sufficient number of relatively easy or relatively difficult test items generally fail to measure low and high levels of achievement with high precision. In fact, excessively easy or difficult assessments may exhibit floors and ceilings—that is, minimum and maximum test scores that are well within the true distribution of scores. More generally, in order to obtain precise measurements along all parts of the scale it is best to construct tests so that all parts of the scale are covered by test items of a given difficulty.[27]

[26] For an extensive review of assessment design and test scaling procedures, see Brennan (2006), Kolen and Brennan (2010), and Wilson (2005).

[27] It is challenging to measure achievement precisely over a wide range of scores using only a single test form. Such a test would need to include a relatively large number of items. Alternatively, some vendors, such as the Northwest Evaluation Association (NWEA), developers of the MAP assessment, use computer adaptive testing

In sharp contrast to value-added indicators, proficiency rates (the indicators required by NCLB) require accurate measurement only at the cut scores that defined the boundaries between proficiency categories (for example, basic, minimal, proficient, and advanced). We speculate that many state assessments have been designed to support accurate measurement at these cut points rather than accurate measurement along the entire achievement spectrum. It is important that the next round of assessments be developed to fully support value-added and growth analyses of the type discussed in this paper.

Many but not all states have developed test scales that are intended to be comparable from one grade to the next. Assessments of this type are said to be *vertically scaled*. Scale scores based on different forms of an assessment (designed for a given grade or achievement level) are said to be *horizontally scaled* if they are scored on the same developmental scale. All state tests are required to be horizontally equated so that it is possible to determine whether school and state proficiency rates have changed over time.

Since it is costly to build assessment systems in which test scores are horizontally and possibly vertically scaled, it is important to be clear about the benefits of building assessments that satisfy these properties. In order to measure student gain (growth) it is, of course, essential that prior and post achievement (more generally, achievement at multiple points in time) be measured on the same (vertical) scale. However, as discussed in the next section, it is not clear that data users should unquestioningly accept test developer claims that test scores from different grades are successfully vertically equated. An important strength of value-added models is that they do not require assessments to be vertically scaled if they include a post-on-pre link parameter, as discussed above.[28,29]

(CAT) to select an optimal set of items for every student, thereby increasing the precision of measured achievement and dramatically reducing the number of test items needed to assess a given individual. See Wainer (2000) for information on computer adaptive testing.

[28] Another alternative, sometimes used in practice, is essentially to discard the developmental scale produced by a test vendor and follow one of two options: (a) Linearly transform all scale scores, separately by grade and possibly test administration date, so that the transformed test scores have a prespecified mean and standard deviation (often 0 and 1, respectively, as in a *z*-statistic) or (b) nonlinearly transform all scale scores, separately by grade and possibly test administration date, so that the transformed scores conform (approximately) to a normal distribution with prespecified mean and variance. The normal curve equivalent (NCE) scale implements the second strategy. Unfortunately, neither approach produces test scales that are vertically equated. In some subject/skill areas, an accurately vertically equated assessment could naturally exhibit expanding or declining test score variability across grade levels. Forcing these tests to have a uniform variance would clearly not produce vertically scaled test scores. On the other hand, either of these two approaches could be helpful in providing some standardized meaning to test scores that are clearly not measured on the same scale (perhaps having been produced by different vendors). Note that transforming test scores at a given level *separately by years* (or test administration dates) should generally be done only if there is reason to suspect that test scores from different years have not been successfully equated. The bottom line is that test scores produced using these methods should not be treated as vertically equated scores and thus should probably not be used to measure gain (growth). They can be used in a value-added model that includes a post-on-pre link parameter.

[29] There is an important exception to this rule in the case of short-cycle assessments—that is, assessments that are given multiple times during the school year (for example in September, December, March, and May). If different students and different schools take/administer short-cycle assessments at substantially different times during the

The value of horizontal equating is potentially much higher and is, in fact, a requirement of NCLB-required assessments. In short, if assessments are properly horizontally equated (and built to accurately measure achievement along the entire continuum of scores), then it is feasible to measure changes in student attainment and value-added productivity over time. In the context of the value-added model presented earlier, the overall year-specific productivity parameter (π) and absolute value-added ($\eta^{ABSOLUTE}$) can be estimated only if assessments are properly horizontally equated. If assessments are not horizontally equated, then the best that can be accomplished is to compare value-added productivity relative to average productivity (given by the parameter η) over all of the teachers, classrooms, and schools included in the system (the reference group). Since the overall productivity of a large reference group (such as a state system) is likely to change much more slowly over time than the overall productivity of a small reference group, relative productivity indicators may reasonably be interpreted as providing (approximate) evidence of absolute changes over time. Hence relative indicators based on large reference groups may be quite useful for schools and districts (since they represent only a small part of an entire reference group). They obviously are useless for tracking the overall performance of the reference group (since the overall productivity average for the reference group is always equal to 0). In our experience, local stakeholders and policy makers universally prefer absolute, rather than relative, comparisons of productivity. They prefer the idea of a fixed performance standard, rather than one that moves up or down, depending on the performance of other entities.[30] Moreover, it is important from a national perspective to be able to track the absolute productivity of states and the nation as a whole.

Despite the obvious value of ensuring that assessments used for accountability purposes are horizontally equated, in our experience many assessments do not appear to be properly horizontally equated, particularly when put to the demanding use of supporting measurement of growth or value-added. We present evidence on this later in this paper. One of the problems is that state assessment systems may currently be constructed so that horizontal equating errors for a *single* test are relatively small, since a single test is all that is required to compute a proficiency rate. In contrast, *four* different tests are required to determine whether a state's value-added productivity increased or decreased (a pretest and posttest for two different cohorts). As a result, value-added indicators could in some cases be subject to unacceptably large horizontal equating errors. We strongly recommend that states (and state assessment consortia) require that test developers ensure that assessments satisfy clearly specified tolerances for horizontal equating error.

As an example we present a case study describing the investigation of assessment data from the state of Wisconsin to determine whether the state tests have been properly horizontally equated. We also consider a variety of statistics to determine whether the distributions of tests scores are comparable over time. These are statistics that could routinely be used to evaluate the degree to which assessments

---

school year, then it is important that assessments administered on different days be equated correctly and reported on the same scale. Short-cycle assessments such as the NWEA MAP meet this criterion.

[30] An accountability/reward system entirely based on relative performance is akin to a tournament, where the outcomes depend on the performance of all participants. See Lazear (1995) and Lazear and Rosen (1981).

have been successfully horizontally equated and designed to accurately measure achievement over a wide spectrum.

Stability in the Distribution of Test Scores[31]

The analysis presented below draws on student assessment data from the state of Wisconsin for the schools years 2005–2006, 2006–2007, and 2007–2008. The state assessment, the Wisconsin Knowledge and Concepts Examination (WKCE), is administered to all students in Grades 3 to 8 and 10 in November of each year. We used this data to estimate value-added models of growth in reading and mathematics achievement for Grades 3 to 4, 4 to 5, 5 to 6, 6 to 7, and 7 to 8. There were approximately 55,000 to 60,000 students in each grade and a total of 425 school districts.

We begin our analysis by examining the means and standard deviations of Wisconsin test scores by grade for four points in time—November 2005, November 2006, November 2007, and November 2008. Students were included in the analysis if they had both a pretest score and posttest score in two consecutive years. We refer to these samples as matched samples. Similar results were obtained for unmatched samples (samples that included all students.) Table 3 includes the number of students (N), the state means of pretest and posttest scale scores (MEAN), and the standard deviations of pretest and posttest scale scores (STD) for each year, grade, and subject.

Table 3. Summary Statistics for Wisconsin State Assessment Data, Matched Samples

			Nov 05 - Nov 06					Nov 06 - Nov 07					Nov 07 - Nov 08				
Subject	Grade		N	Mean		Std		N	Mean		Std		N	Mean		Std	
	Pre	Post		Pre	Post	Pre	Post		Pre	Post	Pre	Post		Pre	Post	Pre	Post
Math	3	4	54079	432.47	468.25	44.35	42.11	55717	435.73	467.80	45.91	44.32	56256	432.98	473.21	43.63	43.44
	4	5	55879	463.90	490.85	44.91	43.28	55388	467.61	494.69	42.34	47.64	56461	467.58	496.78	44.48	47.06
	5	6	56099	485.09	514.67	41.61	44.76	56401	490.72	515.17	43.11	45.58	56026	494.44	515.46	47.66	46.24
	6	7	58971	508.63	536.27	42.60	42.69	57476	514.54	534.51	44.66	43.88	57554	515.13	536.92	45.55	45.26
	7	8	60779	529.05	544.15	44.11	47.31	59915	536.57	543.09	42.39	48.65	58537	534.71	547.65	43.79	47.60
Reading	3	4	54079	458.83	479.05	36.42	43.79	55717	459.32	477.34	38.15	46.15	56265	458.28	478.09	39.55	45.44
	4	5	55879	477.66	486.22	45.35	45.36	55388	478.07	485.79	44.63	45.20	56461	476.96	482.46	46.56	46.61
	5	6	56099	485.80	504.62	46.17	48.00	56401	485.75	504.23	46.06	47.81	56025	485.47	504.28	45.59	51.03
	6	7	58971	501.99	514.20	47.65	46.44	57476	504.29	514.91	48.44	47.13	57530	504.27	517.80	47.97	49.07
	7	8	60779	511.95	528.24	45.80	50.95	59915	514.24	529.04	46.69	50.57	58505	514.99	528.83	47.36	49.99

[31] Although this case study draws on data from a particular state to illuminate issues on building and interpreting value-added models, we should emphasize that the results presented in this paper are very similar to results we have obtained working with other district and state databases.

The information in Table 3 is presented graphically in Figures 9 to 12. In each figure a line represents a different student cohort. For example, in Figure 9 the line for Grades 3-4-5-6 tracks the average mathematics scale score of students in Grade 3 in 2005, Grade 4 in 2006, Grade 5 in 2007, and Grade 6 in 2008 for the same student cohort. As shown in Figure 9, the average mathematics score increased with grade, although at a declining rate—the increase in the average mathematics score is less for higher grades. In addition, there is substantial variation in mathematics scores at a given grade level. In contrast, average reading scores exhibit relatively rapid growth in the even-numbered grades and relatively slow growth in odd-numbered grades, and there is very little variation in reading scores at a given grade level.

Figures 11 and 12 display the standard deviations of test scores across grades and years. As in the previous two graphs, each line represents a different cohort. Note that the standard deviations of mathematics scores vary substantially across grades (about 0 to 6 scale score units) and cohorts (about 2 to 6 scale score units). In contrast the standard deviations of reading scores vary substantially across grades (about 1 to 8 scale score units), but not much across cohorts (about 1 to 2 scale score units).

The bottom line is that there is substantial inconsistency across grades, years, and cohorts in the means and standard deviations of Wisconsin test scores. Taking the data at face value, it would appear that students in Wisconsin learn substantially less in reading in the fourth and sixth grades than they do in other grades. The range of mathematics scores, as measured by the standard deviation, floats up and down over time for the same cohort of students. This calls into question whether it is appropriate to view Wisconsin test scores, measured at different grades and different years, as scale scores that are properly scaled vertically and horizontally.

Stability in Achievement Growth

In the first part of this paper we concluded that instability in the variability of test scores, as documented in the previous section, could substantially affect the relationship between prior achievement and achievement at the end of the school year, as captured by the post-on-pre link parameter (λ). This section reports on estimates of the post-on-pre link parameter for two sets of models, models based on the original scale scores and models based on scale scores that have been transformed to have constant variance at all grade levels and in all years. Figures 13 and 14 report estimates of this parameter for both models in mathematics and reading for two growth years.

As indicated in Figure 13, the post-on-pre link parameter estimates for mathematics based on the original test scales vary widely across grades (from a low of 0.81 to a high of 1.08) and vary widely across the two growth years for the fourth-grade model. The estimates for mathematics based on the transformed (standardized) test scores lie within a narrow band ranging from 0.85 to 0.91. The estimates tend to be smaller in the early grades (approximately 0.86) and larger in the later grades (approximately 0.90). The estimates for reading based on the original test scales also vary widely across grades, with exceptional high values (close to 1 or greater) in Grades 3 and 7 and consistently lower in Grades 4 to 6. As is the case for mathematics, the estimates for reading based on the transformed (standardized) test scores lie within a narrow band and are very similar at all grade levels.

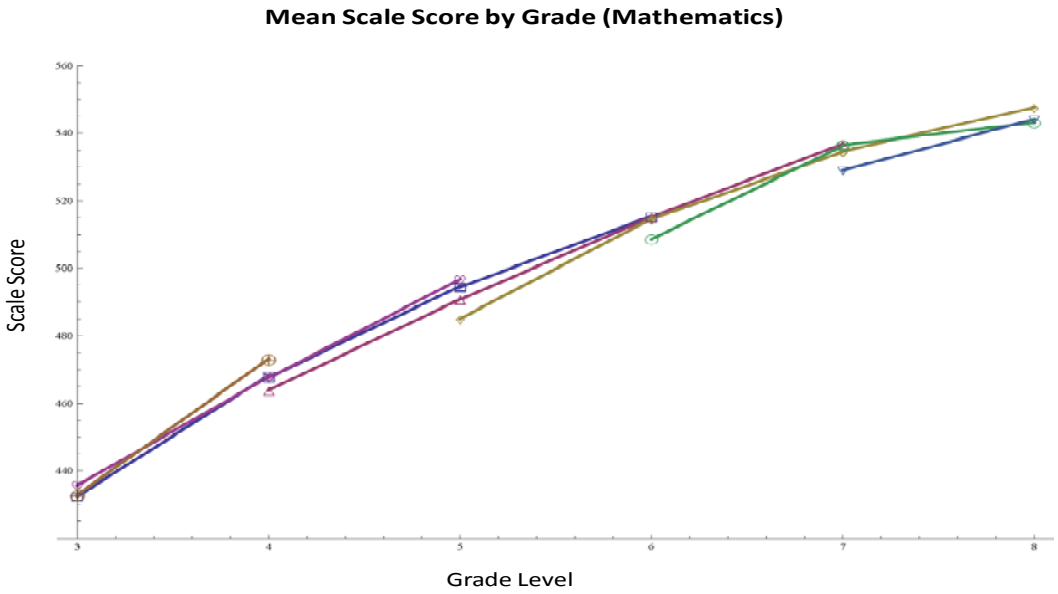

Figure 9. Growth in average mathematics scale scores, Wisconsin. Each line represents a different student cohort.

Figure 10. Growth in average reading scale scores, Wisconsin. Each line represents a different student cohort.

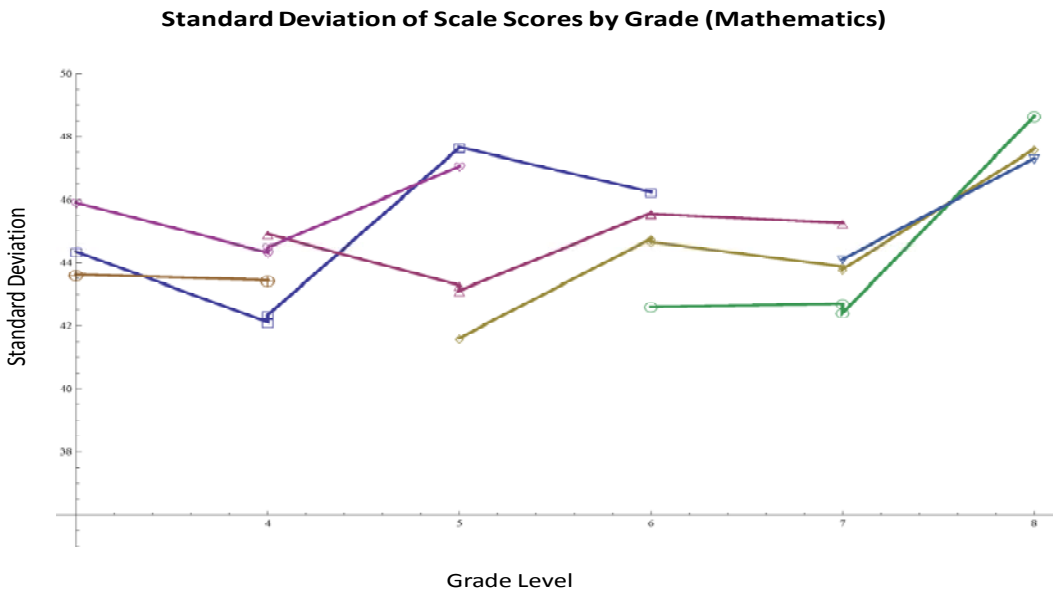

Figure 11. Standard deviations of mathematics scores across grades, Wisconsin. Each line represents a different student cohort.

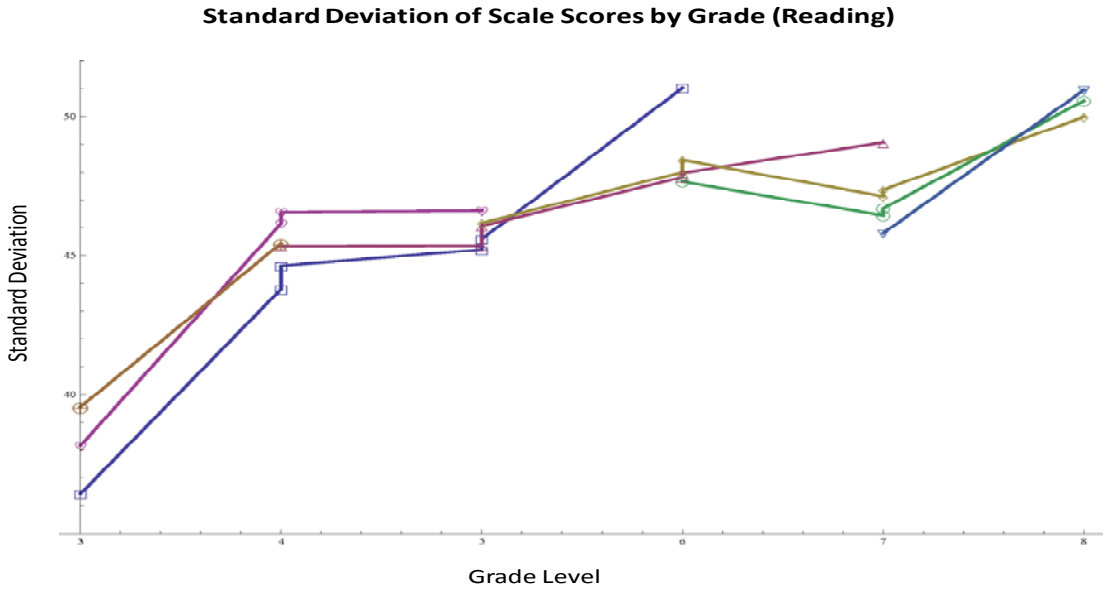

Figure 12. Standard deviations of reading scores across grades, Wisconsin. Each line represents a different student cohort.

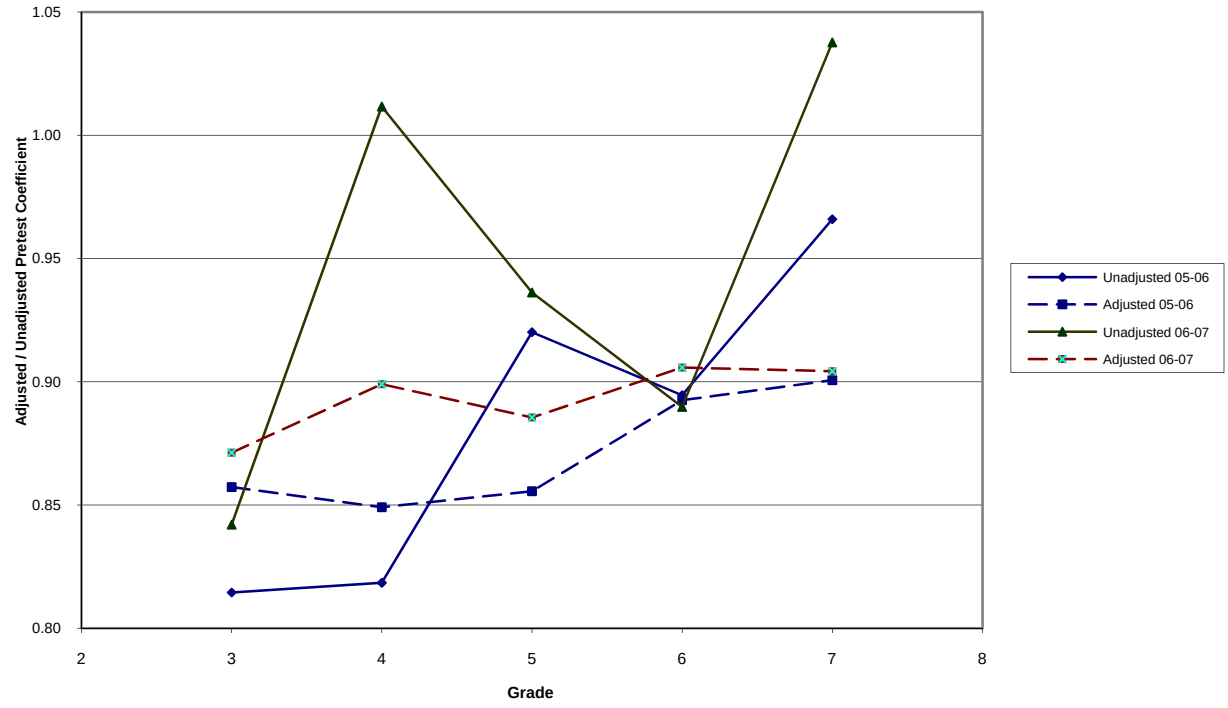

Figure 13. Post-on-pre link coefficients on mathematics.

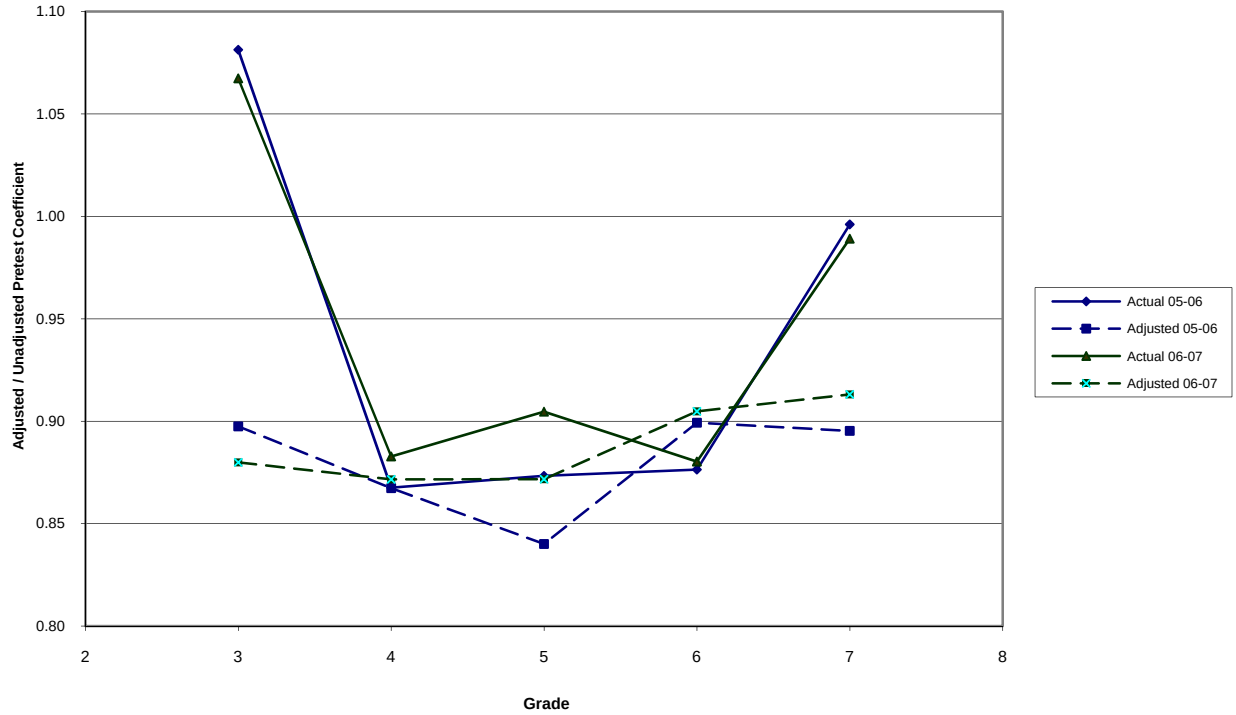

Figure 14. Post-on-pre link coefficients on reading.

The analysis in this section reinforces the concerns raised above that the units of the Wisconsin scale scores are unstable from grade to grade and year to year. We note that a value-added model with a post-on-pre link parameter provides built-in protection against instability in the variability of prior and end-of-year test scores. Alternatively, the vendor-developed test scales can be essentially discarded and replaced by scale scores that have been transformed to have constant variance at all grade levels and in all years. We find consistent evidence that the post-on-pre link parameters in the models with variance-standardized test scores are substantially less than 1, a parameter restriction that is imposed in some value-added models.

Comparability of Value-Added Estimates Over Time: A Test of Horizontal Equating

In this section we consider whether it is appropriate to compare value-added estimates from different cohorts, given the pattern of results presented in the previous section. It is important to assess whether these data support valid comparisons of growth over time. Unreasonably large increases or decreases in average statewide productivity might indicate problems in maintaining the comparability of test scores over time (that is, the accuracy of horizontal equating across test forms). Errors in horizontal equating of test scores over time would make it illegitimate to interpret a change in an estimated state productivity effect as a genuine change in productivity. In such a case, school value-added indicators could only be used to compare the productivity of schools *relative* to other schools in the same year. Relative value-added indicators may reasonably be compared over time if we believe that the true change in state average value-added is relatively small. In that case a reported change in the relative value- added would be approximately equal to the change in absolute value-added.

Since test vendors generally do not provide direct evidence for the degree to which test scores are successfully equated, we believe that analysts should exercise caution in accepting unreasonably large increases or decreases in productivity as evidence of true changes in productivity. We propose the following subjective rules for detecting possibly erroneous changes in productivity for given schools and for the state as a whole:

- School-level rule: Consider a change in productivity greater than 1 standard deviation unit (also referred to as a tier unit) in value-added productivity (corrected for estimation error) as possible evidence of faulty horizontal equating.
- State-level rule: Consider a change in statewide productivity greater than 0.5 standard deviation units (also referred to as tier units) as possible evidence of faulty horizontal equating.

Based on our experiences we suspect that horizontal equating errors are common for statewide tests not originally designed to support growth and value-added analysis.

Now we examine changes in statewide test scores over time to detect whether the changes are consistent with the suggested statewide rule. As indicated in Appendix B, an approximate estimate of

the difference in productivity between two growth years is given by the change in average gain between the two years at a given grade level. These numbers can be computed quite readily.

The average gain in Wisconsin scale scores for each growth year is reported in the first three columns of Table 4. These numbers are taken directly from the attainment statistics presented earlier in Table 3. The change in growth, reported in columns 4 to 6, is computed from the data in columns 1 to 3. Column 7 reports the value-added tier unit (the standard deviation in value-added school effects) for the baseline year 2005–2006. (The value-added standard deviations computed in other years are nearly identical to the reported baseline standard deviations.) The change over time in average gain, reported in tier units, is presented in the bottom panel of the table for each grade. These numbers are equal to the change in growth divided by the value-added tier unit.

As indicated in the bottom panel of Table 4, the change in average gain in mathematics achievement exceeded the specified cutoff of 0.5 tier units in almost all grades. Indeed, most of the changes in mathematics were substantially greater than a full tier unit. The change in average mathematics gain was especially large in grades 6 and 7, substantially in excess of a full tier unit. In fact, in seventh grade the average gain in achievement declined from approximately 15.1 scale score points in Growth Year 1, to 6.5 points in Growth Year 2, and then back up to 12.9 points in Growth Year 3. In short this data implies that it would be unwise to view the absolute value-added productivity estimates in mathematics as comparable based on the current Wisconsin state assessment. As a result we conclude that the Wisconsin mathematics assessment data only supports comparisons based on relative value-added indicators.

In sharp contrast to the results for mathematics, most (but not all) of the changes in gain in reading are less than the cutoff of 0.5 tier units. As a result it may be reasonable to view these numbers as being comparable between the three growth years.

We should caution the reader that the rule used in this section to identify failed horizontal equating is designed to pick up only the most egregious violations of successful horizontal equating. Horizontal equating errors that are less than 0.5 tier points will not, of course, be detected, but could still be much larger than is acceptable. This is an important area for further analysis.

Test Measurement Error

Substantial test measurement error exists in virtually all assessments that are not adaptive. In our experience the average reliabilities of NCLB assessments tend to be around 85%, although the magnitude of error (as typically measured by the standard error of measurement) tends to vary widely across individuals. Errors are lower for individuals with achievement levels that are closely matched to the difficulty of the items included on the test. Adaptive assessments tend to have lower levels of measurement error due to the fact that test-takers are directed toward test questions that closely match the achievement levels of the individuals. For more information on computer adaptive testing see Wainer (2000).

Table 4. Comparison of Average Gain in Achievement Over Time

Average Gain, Change in Average Gain

Subject	Grade		Avarage Gain in			Change in Avg Gain from			VA Tier Unit		
	Pre	Post	Growth Year 1	Growth Year 2	Growth Year 3	G1 to G2	G2 to G3	G1 to G3	in G. Year 1	in G. Year 2	in G. Year 3
Math	3	4	35.79	32.07	40.23	-3.72	8.17	4.45	6.75	7.10	7.00
	4	5	26.95	27.08	29.20	0.13	2.12	2.25	6.93	7.66	6.93
	5	6	29.58	24.45	21.02	-5.12	-3.43	-8.56	7.20	7.06	6.60
	6	7	27.64	19.97	21.80	-7.66	1.82	-5.84	5.42	4.83	5.37
	7	8	15.10	6.53	12.94	-8.57	6.42	-2.16	5.85	5.06	5.25
Reading	3	4	20.22	18.02	19.81	-2.20	1.78	-0.41	4.71	5.19	5.36
	4	5	8.56	7.72	5.50	-0.84	-2.21	-3.06	4.72	4.72	4.76
	5	6	18.81	18.48	18.82	-0.34	0.34	0.00	5.25	5.34	5.64
	6	7	12.21	10.62	13.53	-1.59	2.91	1.32	4.32	3.98	3.75
	7	8	16.29	14.81	13.84	-1.48	-0.97	-2.45	4.35	3.97	4.55

Average Gain Changes in Tier Units

Subject	Grade		G1toG2 Avg Gain		G2toG3 Avg Gain		G1toG3 Avg Gain	
	Pre	Post	1 Tier Unit	2 Tier Unit	2 Tier Unit	3 Tier Unit	1 Tier Unit	3 Tier Unit
Math	3	4	-0.55	-0.52	1.15	1.17	0.66	0.63
	4	5	0.02	0.02	0.28	0.31	0.33	0.33
	5	6	-0.71	-0.73	-0.49	-0.52	-1.19	-1.30
	6	7	-1.41	-1.59	0.38	0.34	-1.08	-1.09
	7	8	-1.46	-1.69	1.27	1.22	-0.37	-0.41
Reading	3	4	-0.47	-0.42	0.34	0.33	-0.09	-0.08
	4	5	-0.18	-0.18	-0.47	-0.46	-0.65	-0.64
	5	6	-0.06	-0.06	0.06	0.06	0.00	0.00
	6	7	-0.37	-0.40	0.73	0.78	0.30	0.35
	7	8	-0.34	-0.37	-0.24	-0.21	-0.56	-0.54

Note. As a rough rule of thumb, year-to-year changes in average gain that exceed 0.5 tier (value-added standard deviation) units indicate possible test form effects and thus may not represent genuine changes in average state productivity. Grades in which average gain exceeds this threshold are shaded in the tables.

Test measurement error is problematic in our context for a number of reasons. First, in order to properly estimate value-added models of the type presented in this paper it is necessary to account for measurement error in prior achievement (using methods from structural equation modeling). Fuller (1986) and Meyer (1992, 1999) discussed techniques for correcting for measurement error in linear

models. These techniques are straightforward to use and have been used to control for test measurement error in all of the value-added systems developed at VARC.[32]

Second, although measurement error correction techniques are a standard statistical tool, using this tool in the value-added context makes the model more complex and difficult to understand. Indeed it is baffling to most people that a simple graph of student gain on prior student achievement yields substantially biased results. The existence of substantial test measurement error typically produces the following: Students with low prior achievement have large positive test score gains, and students with high prior achievement have large negative test score gains. Simple graphical evidence of this type appears to support the conclusion that a school, classroom, or teacher has been very successful at remediation—raising the achievement of low achieving students—but quite unsuccessful at the raising the achievement of gifted and talented students. In general, results of this type are entirely due to test measurement error.

Finally, test measurement error is a primary culprit for producing low-precision value-added estimates when sample sizes are small—for example, when using data for a teacher that has taught for only 1 year. Given the increasing interest in producing value-added estimates at the most disaggregate level (where sample sizes are small), it would be advantageous to favor assessment designs that produce tests with the least amount of error.

Since test measurement error is problematic from the standpoint of conducting formal and informal value-added and growth analyses, we recommend that the next generation of assessments consider using adaptive testing methods to the extent feasible and appropriate.

Conclusions

The design of value-added models and the design of assessments to support these models is a new area of research that is technically challenging and very important from a policy perspective. In this paper we have presented a simple value-added framework that both illustrates what can be learned from a carefully design value-added model of educational productivity and provides a framework for discussing the interaction between value-added model design and the design and properties of student assessments. Our overall message is that value-added models place new demands on the quality and robustness of student assessments. The next generation of student assessments needs to be designed so as to fully support the development of valid and reliable value-added models of student achievement and educational productivity.

One important new area for research and development is the development of new value-added models in subjects and grades other than those covered by NCLB. To make progress in this area it may be necessary to develop new assessments that are tightly connected to the curriculum being taught in these courses.

[32] Measurement error correction techniques are much more complicated to implement in nonlinear models. VARC is currently working with several districts to develop and implement these techniques.

References

Anderson, T., & Cheng H. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, *18,* 47-82.

Arellano, M., & Honore, B. (2001). Panel data models: Some recent developments. In J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (vol. 5; pp. 3229-3296). Amsterdam, The Netherlands: North Holland.

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

Boardman, A. E., & Murnane, R. J. (1979). Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education, 52*(2), 113–121.

Brennan, R. (2006). *Educational measurement (4th ed.).* Westport, CT: Praeger Publishers.

Bush Foundation. (n.d.). *Bush Foundation teacher effectiveness initiative.* Retrieved from http://www.bushfoundation.org/education/teinitiative.asp

Carl, B., Cheng, H., Keltz, J., & Meyer, R. (2010). *A comparison of selected outcome measures across high school types in the Milwaukee public schools.* Madison, WI: University of Wisconsin-Madison, Value-Added Research Center.

Coleman, J. S. (1966). *Equality of educational opportunity (COLEMAN) study (EEOS)*. [Computer file]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2007-04-27. doi:10.3886/ICPSR06389

Dokumaci, E., & Meyer, R. H. (2010). *Multivariate and univariate value-added differential effects.* Manuscript in preparation.

Fuller, W. (1986). *Measurement error models*. New York, NY: John Wiley & Sons, Inc.

Hanushek, E., Kain, J., O'Brien, D., & Rivkin, S. (2005). *The market for teacher quality* (NBER Working Paper No. 11154). Cambridge, MA: National Bureau of Economic Research.

Jones, C., Geraghty, L., Nisar, H., Mader, N., & Meyer, R. (2010). *Evaluation of charter schools in the Milwaukee public schools.* Madison, WI: University of Wisconsin-Madison, Value-Added Research Center.Kolen, M., & Brennan, R. (2010). Test *equating, scaling, and linking: Methods and practices*. New York, NY: Springer.

Lander, R., Keltz, J., Pautsch, C., Carl, B., Geraghty, E., & Meyer, R. (2009). *Evaluation of Milwaukee public schools' READ 180 intervention.* Madison, WI: University of Wisconsin-Madison, Value-Added Research Center..

Lazear, E. (1995). *Personnel economics*. Cambridge, MA: MIT Press.

Lazear, E., & Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy*, *89*(5), 841–864.

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

Meyer, R. (1992). *Applied versus traditional mathematics: New econometric models of the contribution of high school courses to mathematics proficiency* (Discussion Paper No. 966-92). Madison: University of Wisconsin-Madison, Institute for Research on Poverty.

Meyer, R. (1996). Value-added indicators of school performance. In E. Hanushek & W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 197–223). Washington, DC: National Academy Press.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review, 16*(3), 283–301.

Meyer, R. (1999). The production of mathematics skills in high school: What works? In S. Mayer & P. Peterson (Eds.), *Earning and learning: How schools matter* (pp. 169–204). Washington, DC: The Brookings Institution.

Meyer, R. H. (2004). *A generalized additive value-added model of school performance*. Unpublished manuscript, University of Wisconsin–Madison.

Meyer, R. H. (2008, April). *A generalized value-added model with conditional random effects and multivariate shrinkage.* Paper presented at the National Conference on Value-Added Modeling, Madison, WI. .

Meyer, R. H., & Christian, M. S. (2008, February). *Value-added and other methods for measuring school performance: An analysis of performance measurement strategies in teacher incentive fund proposals*. Paper presented at National Center for Performance Incentives Conference, Performance Incentives: Their Growing Impact on American K-12 Education, Nashville, TN.

Meyer, R. H., Dokumaci, E., Morgan, E., & Geraghty, E. (2009, February). *Demonstration of a state value-added system for Wisconsin*. *Report to the Wisconsin Department of Public Instruction.* Madison: University of Wisconsin–Madison, Center for Education Research, Value-Added Research Center.

Rothstein, J. (2007, November). *Do value-added models add value? Tracking, fixed effects, and causal inference* (CEPS Working Paper No. 159). Princeton, NJ: Princeton University, Center for Economic Policy Studies.

Sanders, W., & 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.

Wainer, H. (2000). *Computerized adaptive testing* (2nd ed.). Hillsdale, NJ: Erlbaum.

Willms, D., & Raudenbush, S. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26*(3), 209–232.

Wilson, M. (2005). *Constructing measures*. New York, NY: Psychology Press.

Appendix A

Technical Description of Two-Period Value-Added Model

In this appendix we provide a technical description of the simple statewide (multi-district) value-added model of school productivity discussed in the text. Most, if not all, value-added models (including classroom and teacher value-added models) produce value-added parameters of the type included in this model. In the text we discuss options for generalizing the model to allow for multiple longitudinal observations of student test scores, measurement error in test scores, and other factors.

The key features of the model are:

- Two years of (consecutive grade) longitudinal assessment data for each student (measured annually at the end or beginning of the school year).[33]
- School/district value-added productivity effects η_{klt} (for school *k* in district *l* in year *t*, at a given grade).
- Statewide value-added productivity effects π_t.
- A posttest-on-pretest link parameter λ_t (which may vary across grades and over time). This parameter allows for the possibility that achievement growth may differ for students with high and low prior achievement and for situations where the distribution of the posttest and pretest variables may be non-uniform over time (more on this below).
- Demographic variables X_{it} to capture differences across students (within classrooms) in achievement growth.

The two-period value-added model is defined by the following equation:

$$Y_{2it} = \xi + \lambda_t Y_{1it-1} + \pi_t + \beta_t' X_{it} + \sum_k \sum_l \eta_{klt} S_{iklt} + \varepsilon_{it} \qquad \text{(A1)}$$

[33] Note that since statewide testing begins in third grade in many states, only 2 years of (up-to-date) attainment data are typically available to estimate value-added models of achievement growth from third to fourth grade. In later grades, where additional years of longitudinal data are available (except for students with missing test data), it is possible to expand the two-period model to include multiple grades—for example, in a model of achievement growth from third to fourth to fifth grade. In a two-period model, differences across schools in student growth trajectories are captured directly by the student characteristics that are included in the model. In this model, systematic differences in student growth trajectories not captured by student characteristics included in the model are absorbed by the estimated value-added effects. One of the key advantages of including three or more achievement outcomes for each student (when those data are available) is that it is possible to better control for differences across schools in the student-level determinants of achievement growth than in a model based on two achievement outcomes. See the following for discussions of alternative value-added models: Ballou et al. (2004); Boardman and Murnane (1979); Hanushek et al. (2005); McCaffrey et al. (2004); Meyer (1996, 1997, 2004, 2008); Sanders and Horn (1994); and Willms and Raudenbush (1989).

where the variables, parameters, and indices in the model are defined in Appendix Table A1 and the grade descriptors are omitted for simplicity.

Table A1. Variables and Parameters in Value-Added Model

Variable	Definition
i	Student identifier
k	Within-district school identifier
l	District identifier
t	Year of posttest score
g	Grade (not explicitly included in above model)
Y_{2it}	Posttest score in year *t*
Y_{1it-1}	Pretest score in year *(t-1)* (prior year)
X_{it}	Student demographic characteristics (vector)
S_{iklt}	Student indicator, or fractional measure of enrollment, in school *k,* in district *l,* in year *t*
λ_t	Coefficient on pretest score: posttest-on-pretest link
β_t	Coefficient (vector) for demographic characteristics
ξ	Intercept
ε_{it}	Student level error component

Table A1. Value-Added Effect Parameters

(Parameter	Definition
π_t	Statewide productivity in year *t* (for a given grade). Note that this parameter can only be interpreted as a genuine statewide productivity effect if test scores are accurately horizontally equated over time so that changes in test score growth do not reflect test form effects.
η_{klt}	Relative school and district productivity (hereafter called relative school productivity) for school *k* in district *l* in year *t* (for a given grade). This parameter is referred to as a relative value-added parameter because it is centered at about 0 in each year so that the average school in the district has a value-added rating equal to 0 and school productivity is measured relative to the average school. Changes in statewide productivity are thus absorbed by the parameter π_t.
$\eta_{klt}^{ABSOLUTE} = \pi_t + \eta_{klt}$	Absolute (total) school, district, and state productivity. This indicator incorporates relative school productivity plus overall changes in statewide productivity, provided (as mentioned above) that test scores are accurately horizontally equated.

Note. All parameters are allowed to vary by year, including the slope parameters λ_t and β_t.

Appendix B

The Average Change in Statewide Gain is Approximately Equal to the Change in Statewide Value-Added Productivity

Using the value-added model defined in Appendix A, the average statewide gain from year *(t-1)* to *(t)* at a given grade level is given by:

$$G_t = \overline{Y}_{2.t} - \overline{Y}_{1.t-1} = \xi + \pi_t + (\lambda_t - 1)\overline{Y}_{1.t-1} + \beta_t' \overline{X}_{.t} \quad \text{(B1)}$$

where the bar over each variable (and the dot replacing the *i* index) signifies that the variable is a state mean. The change in statewide gain from posttest year *s* to *t* is similarly given by:

$$\Delta_{st} = G_t - G_s = (\pi_t - \pi_s) + (C_t - C_s) \quad \text{(B2)}$$

where C_t, a cohort variable, is defined as:

$$C_t = (\lambda_t - 1)\overline{Y}_{1.t-1} + \beta_t' \overline{X}_{.t} \quad \text{(B3)}$$

The cohort variables will typically not change much from year to year, so that the change in statewide gain approximately equals the change in statewide productivity, asserted as:

$$\Delta_{st} = G_t - G_s \approx (\pi_t - \pi_s) \quad \text{(B4)}$$

The change in statewide gain in tier units is obtained by dividing gain Δ_{st} by the standard deviation of school productivity in the baseline year ω.

Appendix C

Student- and School-Level Variables in Value-Added Models

Meyer (1996, 1997) and Willms and Raudenbush (1989) discussed some of the conceptual and empirical issues involved in including student and school-level control variables in value-added models—for example, average poverty status. The primary concern with including school-level control variables is that the estimated coefficients on these variables could be substantially biased if school resources and intrinsic school productivity are not assigned to schools such that school-level control variables are uncorrelated with unobserved school productivity. This condition would be violated, for example, if high-performance teachers and administrators preferred to work in schools with low poverty. In this case the coefficient on average poverty status would absorb this negative correlation, yielding an estimated coefficient biased in the negative direction.

Estimated value-added indicators would of course be biased if the coefficients on school-level control variables were biased. For example, if the coefficient on average poverty status is biased in the negative direction, the estimated value-added performance of schools that disproportionately serve high-poverty students would be biased upward. This is problematic if one of the important purposes of a value-added system is to identify schools that are in need of improvement. Because of this, model designers need to be cautious about including school-level control variables in value-added models. This is an important area for further research.

Note that the statistical concerns discussed above do not apply to student-level control variables. These coefficients are estimated off of variation within schools or classrooms—for example, the contrast in achievement growth between low- and high-poverty students within schools or classrooms. If resources are allocated within schools or classrooms in a way that is systematically related to student characteristics, then the coefficients on student-level control variables will capture these systematic patterns. The other role for student-level control variables is to proxy for the differences in resources provided to students by their families.

Appendix D

Additional Resources on Value-Added Models and Related Literature

Betts, J., Zau, A., & Rice, L. (2003, August). *Determinants of student achievement: New evidence from San Diego*. San Francisco, CA: Public Policy Institute of California.

Chamberlain, G. (1984). Panel data. In *Z.* Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics, 2* (pp. 1247–1318). Amsterdam, Netherlands: North-Holland.

Clotfelter, C., Ladd, H., & Vigdor, J. (2007). *Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects* (NBER Working Paper No. 13617). Cambridge, MA: National Bureau of Economic Research.

Clotfelter, C., & Ladd, H. (1996). Recognizing and rewarding success in public schools. In H. F. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 23–63). Washington, DC: The Brookings Institution.

Clotfelter, C., Ladd, H., & Vigdor, J. (2007). *Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects* (NBER Working Paper No. 13617). Cambridge, MA: National Bureau of Economic Research.

Frees, E. (2004). *Longitudinal and panel data: Analysis and applications in the social sciences*. Boston, MA: Cambridge University Press.

Ghosh, M. (1992). Constrained Bayes estimation with applications. *Journal of the American Statistical Association*, *87*(418), 533–540.

Ghosh, M., & Rao, J. N. K. (1994). Small area estimation: An appraisal. *Statistical Science, 9*(1), 55–76.

Goldberger, A. (1991). *A course in econometrics*. Cambridge, MA: Harvard University Press.

Harris, D., & Sass, T. (2006). *Value-added models and the measurement of teacher quality*. Unpublished manuscript, Florida State University, Tallahassee.

Kane, T. J., & Staiger, D. O. (2002). *Improving School Accountability Systems*. Unpublished manuscript, National Bureau of Economic Research, Cambridge, MA.

Lockwood, J., McCaffrey, D., Mariano, L., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics, 32*(2), 125–150.

Longford, N. T. (1997). Shrinkage estimation of linear combinations of true scores. *Psychometrika, 62*, 237-244.

Longford, N. T. (1999). Multivariate shrinkage estimation of small area means and proportions. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 162*(2), 227–245.

Longford, N. T. (2004). Missing data and small area estimation in the UK Labour Force Survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 167*(2), 341-373.

McCaffrey, D., & Hamilton, L. (2007). *Value-added assessment in practice: Lessons from the Pennsylvania Value-Added Assessment System Pilot Project* (RAND Technical Rep. No. TR-506-CC). Santa Monica, CA: RAND Corporation.

McClellan, M., & Staiger, D. (1999). *The quality of health care providers.* Unpublished manuscript, National Bureau of Economic Research, Cambridge, MA.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, *78*(381), 47–55.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica, 46*(1), 69–85.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, *20*(4), 307–335.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Robinson, G. K. (1991). That BLUP is a Good Thing: The estimation of random effects. Rothstein, J. (2008). *Teacher quality in educational production: Tracking, decay, and student achievement* (NBER Working Paper No. 14442). Cambridge, MA: National Bureau of Economic Research.

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

Searle, S., Casella, G., & McCulloch, C. (1992). *Variance components*. New York: Wiley-Interscience.

Stroud, T., Platek, R., Rao, J., Sarndal, C., & Singh, M. (1987). *Small area statistics*, *Vol. 26*, New York, NY: Wiley. Tate, R. L. (2004). A cautionary note on shrinkage estimates of school and teacher effects. *Florida Journal of Educational Research*, *42*, 1–21.

Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M. E., Roth, J., … Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 11–35.