CASE STUDIES on DATA USE:
A Series of Reports Developed by the CCSSO Accountability Systems and Reporting (ASR) Collaborative

Report Submitted to the ASR Collaborative
on September 29, 2009
Boston, Massachusetts

Lead Researcher:
Carole Gallagher, Ph.D.
Assessment and Accountability Comprehensive Center at WestEd

Prologue:
Purpose of This Series of Reports on Effective Data Use[1]

This series of reports on data use is an ongoing project of the Council of Chief State School Officers, Accountability Systems and Reporting Collaborative (ASR). The purpose of these reports is to highlight ways in which state or federal accountability systems have presented opportunities for collection, reporting, and use of data at the state and local levels that have contributed to improvements in valued educational outcomes. Primarily in the form of annual test scores from standards-based assessments, such data have been shown to be useful at a number of levels for setting and reaching goals for learning for all students (Braun, 2005; Gong, Perie, & Dunn, 2006; Heritage & Yeagley, 2005; Herman, Yamashiro, Lefkowitz, & Trusela, 2008; Phelps, 2008; Ross, Sanders, Wright, Stringfield, Wang, Weiping, & Albert, 2001; Singer & Willet, 2003).

Data generated through accountability systems may be used in important ways by the higher education and research communities; by policy makers and state departments of education; by district superintendents, school principals, and teachers; and by parents and students. Researchers and state department staff regularly analyze performance data for cross-school or student subgroup trends and patterns in levels of proficiency and for indications of what is working and what it not. District and school leaders rely on these data largely to help organizational strengths and limitations and make decisions about programs, professional development for staff, and allocation of resources. Teachers may use accountability-based data to evaluate instructional strategies and identify those students most in need of academic support. Parents and students may examine test scores to consider progress toward annual learning targets and preparedness for success at the next grade or level. Policy makers use accountability data to set meaningful performance standards for all students and to monitor the degree to which student achievement in districts and schools meets annual performance goals in core content areas.

In all cases, the accessibility of timely, high-quality data and the capacity to support appropriate data use are critically important (Gong, Perie & Dunn, 2006; Marsh, Pane, & Hamilton, 2006). This series of papers intends to highlight a range of states' assessment and accountability systems and to discuss the historical and political contexts in which they were developed, provide an in-depth examination of the theoretical foundation and methodology for the accountability model, and to present specific examples of effective use of data. In all reports, comments from semi-structured interviews with representatives from the respective state departments of education will be included.

[1] This series is a follow-up to previous reports developed by the CCSSO ASR Collaborative.

Introduction to Growth Models[2]

In 2005, the USED began accepting proposals from states seeking to incorporate a growth component in their federal accountability systems. The growth models for two states (Tennessee, North Carolina) were approved in Round 1 of the pilot study (2005-2006), and another five (Arkansas, Delaware, Florida, Iowa, Ohio, Alaska, and Arizona) were approved in Round 2 (2006-2007). In the third round, the USED opened the pilot program to all other eligible states, and in 2008, the growth models for Michigan and Missouri were approved. All growth models approved for federal accountability purposes must meet the seven core principles set by the USED for growth model pilots and provide a viable alternative to states for measuring and reporting on student and school academic performance (USED, 2006).[3]

Growth models are one subset of the family of longitudinal models that use data from multiple points in time to examine changes in learning outcomes (Singer & Willet, 2003). Specifically, growth models have been found to be useful tools for assessing the degree to which students are progressing toward achievement targets or standards, monitoring changes in student achievement over time, examining the cumulative effects of teaching and learning, and evaluating program effectiveness (Choi, Goldschmidt, Yamashiro, 2005; Goldschmidt & Choi, 2006; Osgood & Smith, 1995). By modeling achievement growth over time, a growth model can account for the cumulative processes of learning (CCSSO, 2007).

Unlike cohort or status models, these models rely on longitudinal data systems to track the achievement scores of individual students over time and across schools. Increasingly, states developing comprehensive accountability systems have a need for the types of data generated by these models to guide decision-making about instructional improvement. When used in conjunction with a status model for accountability purposes, educators and policy makers have both a snapshot image of a school's annual level of achievement as well as more detailed information about the ways in which students' and schools' scores are changing over time.[4]

Growth models focus on changes in performance of individual students (and/or the aggregate of individual growth at the school or district level). For this reason, proponents believe that they are more directly linked to teaching and learning than cohort models that do not track the same students or groups of students over time

[2] See *Policymakers' Guide to Growth Models for School Accountability: How do Accountability Models Differ?* (2005), *Implementer's Guide to Growth* (2007), and *Guide to USED Growth Model Pilot Program 2005-2008* (2009) for more detailed discussions of growth models.

[3] The accountability model must ensure 100% student proficiency by 2014 & ensure a closing of the achievement gap for all student groups; establish high expectations for low-achieving students without setting annual achievement expectations based on student demographic or school characteristics; produce separate accountability decisions for reading & mathematics; include all students, schools, & districts; hold schools & districts accountable for the performance of student subgroups; include annual assessments in grades 3-8 & high school in reading & mathematics; produce comparable results from grade to grade and year to year; have been operational for at least one year & approved through the NCLB peer review process; track student progress; & include student participation rates & student achievement on an additional academic indicator.

[4] The reader is referred to a number of other sources (e.g., CCSSO, 2005; 2007; Choi, Goldschmidt & Yamashiro, 2005; Gong, Perie, & Dunn, 2006; Ladd & Lauen, 2009) for additional information about the strengths and limitations of different accountability models.

but instead report on the performance of successive groups of students at a particular grade level (Betebenner, 2008; Rumberger & Palardy, 2004; Seltzer, Choi, & Thum, 2003; Willms & Raudenbush, 1989).[5] Their use has been shown to support the types of data-driven decision-making associated with improved student performance (Black & Wiliam, 1998; Chrispeels, Brown, & Castillo, 2000) and has contributed to the development of innovative reporting strategies that more fully inform stakeholders about valued learning outcomes (Stringfield, Wayman & Yakimowski, 2005; Wayman, 2005; Thum, 2003). While once used primarily for educational research and program evaluation, a number of states now include a growth model as one component of their comprehensive state or federal accountability system.

Yet growth model use is associated with specific challenges. They are demanding in terms of technical resources, in that states must have the capacity to track individual students over time and across schools via unique student identifiers, to match new test data with archived data for each student, and to maintain and store large data sets in a secure environment (McCaffrey, Lockwood, Koretz & Hamilton, 2003; Sanders, Wright, & Rivers, 2006). They require assessments that meet high standards of technical adequacy, with evidence that inferences drawn from results are valid for this purpose, that tests are reliable and fair, that the items are aligned in meaningful and substantive ways to content standards; that the content assessed is representative of the domain's range of breadth, depth, and scope at that grade level and is linked in developmentally appropriate ways across grades; and that annual growth targets are defensible (Braun, 2005; Goldschmidt & Yamashiro, 2005; Rabinowitz, 2004). Prior to implementation, a number of tradeoffs must be weighed and key decisions made about what type of growth will be measured, how much growth will be considered sufficient, and if students with different starting points should be expected to grow at the same rate (Gong, Perie, & Dunn, 2006).

Types of Growth Models Used for Accountability Purposes

When evaluating Adequate Yearly Progress (AYP) for federal accountability purposes under NCLB, states report the proportion of schools meeting a set standard for performance (percent proficient). This status snapshot (*How are students in grade 6 doing this year?*) does not take into consideration that students and schools enter with different levels of achievement, so it is more challenging for some schools than others to make enough progress to meet annual AYP targets (Braun, 2005; Goldschmidt, 2004). NCLB-based status models do not reward those schools whose test scores needed to show the greatest improvement—or growth—in order to meet target proficiency levels. Growth models, alternatively, provide additional incentive to those schools who may lag behind in overall percent proficient but who are contributing significantly to the movement of students toward proficiency each year (Goldschmidt, Roschewski, Choi, Auty, Hebbler, Blank, Williams, 2005).

For state and federal accountability purposes, three types of growth models have emerged. These include (1) growth to proficiency, growth to a standard, or trajectory models; (2) value table or transition models; and (3) projection models. Each is used for a specific purpose and each has unique strengths and limitations.

[5] E.g., this year's sixth graders compared to last year's sixth graders.

The *growth to proficiency* family of models are intended to show if a student is on track to reach a proficiency target as some specified point in future. The models work by setting a proficiency target for some future grade (generally 3-4 years beyond the current grade), determining the gain required to reach the proficiency target from the current score, then dividing the required gain into annual increments. Students are considered on track to reaching proficiency if, assuming their performance trend continues, their score gains match or exceed the annual increment required to reach proficiency in the future grade. For AYP calculations, schools may count as proficient those students whose gains match or exceed the annual increment. One challenge associated with this model is that achievement targets may need to be reset each year. States currently using this type of growth model for accountability purposes include Alaska, Arizona, Arkansas, Florida, Missouri, and North Carolina. Colorado recently joined the list of approved federal Growth Model Pilot states with a growth-to-proficiency model.

Value table or *transition* models are intended to evaluate student transitions across performance levels, with the goal of moving students from lower performance levels (or sub-levels[6]) to higher performance levels. For AYP purposes, schools may count as proficient those students who moved into higher performance levels or sublevels during the school year. Because growth is measured by change in performance level, a vertical scale is not necessary (CCSSO, 2007). However, transition from one level to another (e.g., below basic to basic, basic to proficient, or proficient to advanced) at one grade level (e.g., early grades vs. later grades) may be assigned different values or weights, and it may be challenging for all stakeholder groups to reach consensus on the relative value of change from one level to another (Lissitz, 2005). States currently using this type of growth model for accountability purposes include Delaware, Iowa, and Michigan.

The *projection model* is intended to predict or project student performance into the future. Like the growth to proficiency models, for AYP calculations, schools may count as proficient those students who have not yet met proficiency but are predicted or projected to reach proficiency in the future (generally within 3-4 yrs). However, in projection models, student performance is predicted based on past performance *and* the performance of a normative sample of peers (prior cohorts) in the target grades, and then compared to the proficiency standard for the target grade. These models are among the more statistically complex growth models, generally estimated via linear or multi-level regression equations. Three states currently use a particular type of projection model for accountability purposes; Tennessee, Ohio, and Pennsylvania all have a Value-Added Model (VAM) as a component of their comprehensive accountability systems (see details below).

Value-added models represent one special class of projection models. Via longitudinal analyses, each student's past performance is used to estimate a projected score for that student. As with other projection models, in the value-added models (VAM), attained student scores are compared to projected scores.

[6] Typically, performance levels are subdivided such that students reach proficiency in a set number of years (generally 3-4).

A student whose actual score exceeds the projected score has demonstrated growth. For federal accountability purposes, students whose scores meet or exceed the score needed to reach proficiency by the target date may be included in AYP calculations as proficient. Depending on the structure of the data and purpose of analyses, VAMs range from simple gain score or fixed effects models to more complex multivariate or cross-classified models that track students across teachers and schools over time. In these models, the estimates describe the residual, or the part of the score left unexplained by other factors following analysis, which are assumed to be related to the combined effects of school and classroom context (Raudenbush & Bryk, 2002). A major challenge associated with using VAMs for accountability purposes is the controversy about the attribution of causal effects associated with these estimates.

Emerging Context for Growth Models

States seek to provide educators with the types of data that can inform decision-making about the types of instruction that effectively support student achievement. Recent developments in federal funding support this need by providing new opportunities for states to consider implementing a growth model as part of their comprehensive accountability systems. In the *Federal Register* announcement for Race to the Top funds posted in July 2009, growth models are generally described as one component of a potentially fundable proposal. In that announcement, USED defines "student growth" quite specifically:

> *Student growth* means the change in achievement data for an individual student between two points in time. Growth may be measured by a variety of approaches, but any approach used must be statistically rigorous and based on student achievement (as defined in this notice[7]) data, and may also include other measures of student learning in order to increase the construct validity and generalizability of the information. (p. 37811-37812)

Growth data have the potential to provide a clear and valued complement to status measures that focus on the percentage of students reaching proficiency annually (Braun, 2005; Goldschmidt, 2004).

[7] Academically challenging, technically sound annual standards-based assessments.

CASE STUDY on DATA USE:
THE TENNESSEE VALUE-ADDED ASSESSMENT SYSTEM (TVAAS)

Report to the CCSSO
Accountability Systems and Reporting (ASR) Collaborative
Report 1 in Series

Carole Gallagher, Ph.D.
Assessment and Accountability Comprehensive Center at WestEd

In collaboration with:

Tennessee Department of Education
Dan Long, Executive Director
Assessment, Evaluation, Research, and e-Learning
Marcy Tidwell, Associate Director
Assessment, Evaluation, and Research

Report 1: The Tennessee Value-Added Assessment System

This first report in the series focuses on the **Tennessee Value-Added Assessment System** (TVAAS). Following an introduction to the genre of accountability models based on growth, the background and methodology of TVAAS is examined. Three types of data use are highlighted in this report:

- Section I: Using TVAAS Data to Monitor Changes in Student Achievement
- Section II: Using TVAAS Data to Evaluate Teacher Effect
- Section III: Using TVAAS Data to Examine School Effectiveness

In the final section, strengths and challenges associated with using the TVAAS system for accountability and other purposes are discussed. A number of examples of the types of reporting documents associated with data collected at the student, teacher, and school levels are included as appendices.

Introduction to Value-Added Models

Value-added models represent one special class of projection models. Via longitudinal analyses, each student's past performance is used to estimate a projected score for that student. As with other projection models, in the value-added models (VAM), attained student scores are compared to projected scores. A student whose actual score exceeds the projected score has demonstrated growth. For federal accountability purposes, students whose scores meet or exceed the score needed to reach proficiency by the target date may be included in AYP calculations as proficient.

Depending on the structure of the data and purpose of analyses, VAMs range from simple gain score or fixed effects models to more complex multivariate or cross-classified models that track students across teachers and schools over time. These VAMs use mixed model and multilevel or hierarchical modeling approaches to estimate both initial status and score gains over time while accounting for measurement error (Bryk, Thum, Easton, & Luppescu, 1998).[1] More statistically complex mixed models, such as the Education Value-Added Assessment System (EVAAS), offer the added advantage of making use of all data available for each student as "…the entire observational vector of each student's test data is fitted simultaneously" (Sanders & Wright, 2009; p. 1).

In estimating growth, VAMs attempt to account for students' prior achievement (e.g., previous years' test scores) in various ways as all students do not start at the same academic level (Goldschmidt, 2004). In some models, the interaction between growth and initial status is explicitly modeled and used (1) to estimate the average expected school growth and (2) to describe the distribution of student growth within a school (Choi, Seltzer, Herman, & Yamashiro, 2004). The EVAAS models—including the value-added models developed in Tennessee and Pennsylvania—account for student prior achievement and other student-level factors *implicitly* by using students as their own controls and creating growth patterns through statistical procedures such as stacked blocking (Sanders, 1994).[2]

[1]The mixed-model methodology indicates that the statistical model contains both fixed and random effects. Multilevel models are one subclass of mixed models (Betebenner, 2004).

[2] See Chapter 2, Section II for a more detailed description of this process and rationale for its use.

Two key assumptions are associated with most value-added models. First, because growth is determined by comparing scores from the same student at different points in time, background factors and prior achievement are assumed to be "controlled" so that other possible explanations for change can be examined (Hanushek, Rivkin, & Taylor, 1996; Raudenbush & Bryk, 2002; Sanders, 1998).[3] According to Sanders, Saxton, and Horn (1998),

> Each child can be thought of as a "blocking factor" that enables the estimation of school system, school, and teacher effects free of the socioeconomic confoundings that historically gave rendered unfair any attempt to compare districts and schools based on the inappropriate comparisons of group means. (p. 138)

Second, in most VAMs, score gains are interpreted as the outcome of highly effective instruction; scores that fall below the projected are interpreted as the outcome of ineffective instruction (Sanders, 2006; Wright, Sanders, & Rivers, 2006). This interpretation is based on the assumption that changes in scores over time can be attributed to a specific time, agent, or experience. When the tests from which the scores are derived are aligned with standards for instruction, the specific effects of a particular instructional program, teacher, or school on test performance can be separated from non-school related factors, such as family background (Goldschmidt, Roschewski, Choi, Auty, Hebbler, Blank, & Williams, 2005; Ladd & Walsh, 2002; Meyer, 1996; Sanders, 2000).

Some value-added models are designed to take into account the nature of nested data (i.e., test scores within students [scores for each test event], students within classrooms, and classrooms within schools) inherent to educational testing and the reality that students are *not* randomly assigned to classrooms or schools (Raudenbush & Bryk, 2002; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Weisberg, 1979). These models have the statistical capacity to isolate student-, teacher-, and/or school-level effects and to specify teacher or school effects as constant or varying randomly. Such models may be more robust to missing data as they have the statistical capacity to capitalize on existing data and to maintain all cases with at least two data points. These are significant strengths when conducting longitudinal studies in real-world contexts (Bryk, Thum, Easton, & Luppescu, 1998; Darlington, 1997).

Data from value-added models may be used for purposes other than state or federal accountability and have been adapted to meet a variety of needs in over 300 school districts across the nation (TDOE, 2008). They may be used to estimate teacher and school effects as well. Value-added estimates describe a residual, or the part of the score left unexplained by other factors following analysis; these factors are assumed to be related to the combined effects of school and classroom context (Raudenbush & Bryk, 2002). A teacher whose actual classroom mean score exceeds the expected in comparison to other teachers at his/her school is viewed as having contributed "value" to students'

[3] The theoretical premise is that person-specific factors (e.g., gender, race/ethnicity, family background, innate ability) are controlled because they remain constant across all testing events in which that student participates. However, school-specific factors (e.g., classroom or teacher assignment, instructional program) do change across testing events, and therefore can be linked causally to the outcome.

performance (Sanders, 2000). Similarly, schools whose mean growth is greater than the expected have added “value” to students’ performance.

Proponents of VAM find that, when used responsibly and interpreted in conjunction with other types of information, data generated from these models enable educators and stakeholders to see the progress students make each year and how much specific teachers, schools, or districts—relative to other teachers, schools, or districts—contributed to that progress (Braun, 2005; Meyer, 1996; Lockwood, McCaffrey, Mariano, & Setodji, 2007; Sanders, 2006). Data collected through VAM systems may be used by administrators for formal program evaluation and to identify those teachers who might benefit most from professional development support and resources. These data also may be used formatively by teachers to improve instructional strategies. Most models can be adjusted to accommodate changing assessments over time as well as incorporation of multiple measures and/or multiple cohorts.

Tennessee's Value-Added Assessment System[4]

> Our work indicates that the biggest impediment to ever higher achievement is the years in which individual students are not making realistic growth... Without yearly feedback from responsible measurement, often teachers and principals do not recognize that these hurtful patterns exist. However, we certainly know of cases in which teachers, after being presented with the results from the data, have engineered for themselves strategies within their classrooms that have made instruction more equitable—addressing the needs of all students, rather than just a few (Sanders, 2000, p. 337).

The Tennessee Value-Added Assessment System (TVAAS) was introduced as part of the Tennessee Education Improvement Act of 1992. Developed primarily in response to complaints about inequities in school funding from rural schools, this model was recommended by the Tennessee Department of Education (TDOE) and approved by the legislature because of its potential to provide concrete evidence of the nature and degree of impact of classroom instruction in five content areas. Under the TVAAS, students are tested annually using standardized assessments (Tennessee Comprehensive Assessment Program, or TCAP) in grades 3 through 8 that measure student learning and changes in achievement from one school year to the next in mathematics, reading/language arts, writing, science, and social studies. Analyses of these data are mandated to take into account differences in prior achievement (or starting points) when estimating the impact of a teacher, school, and/or district on individual student gains.

In May 2006, the TVAAS was one of only two systems approved by USED as a growth pilot for federal accountability purposes.[5] This projection model uses all available past TCAP scores to project if a student will score proficient or advanced on the statewide assessment in three years. Projection in the TVAAS accountability model is estimated via linear regression and is based on two assumptions: (1) students will receive the average Tennessee schooling experience and (2) students will receive future instruction of average effectiveness (Sanders, Saxton, & Horn, 1997). While prior to 2006 TVAAS data were used primarily to measure the effect over the course of one-year's instruction of a teacher, school, or district on a specific student group, the approved growth model for accountability purposes predicts each individual student's future achievement as it relates to state academic curriculum standards (TDOE, 2005). Per federal guidelines for approved growth models, no student background factors (e.g., socioeconomic status) are included as covariates in TVAAS growth model analyses.

In Tennessee, schools can meet AYP in three ways: via the status model (AYP), the safe harbor (cohort improvement) provision of AYP, or the growth model, by including in the calculation of status those students who are projected to be proficient in the future year. Growth is projected for students in grades 4-8 who have at least one previous test score; for those students in grade 3, in high school, or who took the alternate assessment, growth is not projected and instead they are included

[4] The authors wish to acknowledge the contributions of Dan Long, Marcy Tidwell, and Vicky Smith from the Tennessee Department of Education in writing this chapter.
[5] North Carolina's model also was approved at that time.

in AYP calculations based on status. In terms of AYP, the growth component currently is an option only for schools and has not been extended to districts.

Scores for *Achievement Grades* are reported on the State normal grade equivalent (NCE) scale. Tennessee administered both NRT and CRT tests in the spring of 2004.[6] Since each student took both tests, the two tests could be equated and previous NRT test data was mapped onto the CRT scale. After mapping onto the CRT scale, the data were converted into state NCEs using 1998 data as the baseline. For Achievement Grades, then, a school with an NCE of 50 has a mean achievement score equal to the state average in 1998 (TDOE, 2009b).

Scores for *TVAAS/Value Added Grades* are based on state NCEs with a 0.0 growth standard. This conversion provides a way for TVAAS analyses to measure achievement and academic gain for each district and each school against a consistent metric, expressed in state NCEs, as students move from grade to grade (TDOE, 2009b). By measuring student progress within a grade and subject, TVAAS score reports are intended to highlight the influence of in-school factors on student achievement.

TVAAS uses up to five years of the most recent data for each student when calculating growth. Scores from state tests and subtests in five content areas (reading/language arts, writing, mathematics, science, social studies) over all years available are stored in the database. This model boasts an estimation algorithm that capitalizes on existing data and is robust to missing student-level data.[7] District means include all student-level data, even if student data cannot be tracked to a specific teacher (Braun, 2005).

Tennessee as a Pioneer for Longitudinal Data Use

For the past sixteen years, comprehensive reports that include value-added data have been sent by the TDOE to stakeholders that include parents, teachers, administrators, policy makers, and the research community.[8] The value-added reports complement TCAP Achievement, Writing, and End-of-Course reports by providing descriptive information about growth patterns over time (TDOE, 2009a). The elementary and middle school reports compare NCEs to the growth standard, while the high school reports include a school effect score based on the mean difference between the mean observed and mean predicted scale scores.

These efforts were bolstered in 2005 when the TDOE received a longitudinal data system (LDS) grant that allowed them to focus on data management goals and develop the infrastructure for a comprehensive data warehouse designed for performance reporting. Building the LDS required cross-program buy-in, with opportunities for TDOE staff to identify gaps in current data reporting capacity, help refine draft reports, and provide feedback on planned quality assurance

[6] Tennessee has transitioned from using a norm-referenced test to using a criterion-referenced test and scores have been back-mapped for correlation purposes to 1998 to assure consistency in scales.

[7] By including data over time and across students, the types of systematic variability introduced by missing data are assumed to be minimized in TVAAS estimates (Sanders, Saxton, & Horn, 1997). See Ballou, Sanders, & Wright (2004) for detailed description of TVAAS estimation procedures or McCaffrey, Lockwood, Koretz et al.(2004) for general model specifications.

[8] Examples of different reports are included in this report's appendices.

measures. Data now stored for each student include exit status, test results, attendance, discipline history, teachers, and course enrollment (TDOE, 2008). Truancy and graduation rates are reported at the school-, district-, and state-levels annually and course-taking patterns are tracked at the student- and school-levels. The LDS also is used to identify those districts most in need of resources in order to reach annual learning targets. These functions signal an intentional shift towards using data more proactively for school improvement (TDOE, 2008).

In addition, state educators can access an online, password-protected site that includes archived and current TVAAS data.[9] From this database, teachers can create customized reports with finely grained student-, subgroup-, or classroom-level data related to growth. The TDOE has plans to develop online "Learning Paths" at the TVAAS site that guide teachers to other sources of data (e.g., TCAP) and provide tutorials for appropriate data use at each site. These types of ad hoc information collecting and reporting capabilities are emerging as user-friendly strategies for supporting educators with data-supported decision-making (Long, 2009).

Data from TVAAS are used in a number of ways other than for federal accountability purposes. They are used to support student achievement, improve teaching methods and teacher quality, and narrow achievement gaps (Sanders, 2004). Information is available to educational decision-makers that can be used formatively to improve instructional practice and diagnostically to pinpoint individual student's strengths and limitations (Sanders & Wright, 2009). Growth data on school reports help educators identify at-risk students and target resources toward those most in need. The rich database is used extensively for research purposes, to address questions such as the degree to which student growth is sensitive to SES, racial differences, school location, or prior status (Ballou, Sanders, Wright, 2004; Bratton, Horn, & Wright, 1996). According to TDOE (2009b),

> In Tennessee, we are moving on from numbers and talking more about what these data really mean. Using feedback from practitioners—we try to provide teachers with what they need—we are building district-level 'data teams' who can go out and have conversations with teachers about what these data say about their students and their classroom instruction. We believe this information is guiding instructional practice, not just by 'tweaks' showing up in test scores, but by what we hear teachers in the field talking about in terms of system improvement. There are large centers of best practice across the state…these schools are making a difference. In one district, every building principal has a weekly session with staff on how to use TVAAS data in conjunction with TCAP data. Another district is piloting a project where TVAAS data are updated for each teacher, reconfigured to include all students in the classroom. The next step is for these teams to begin having conversations with parents.

Recent Developments Related to TVAAS

In a recent communication with state educators, the TDOE announced that in 2010, the state would be implementing a new curriculum and set of assessment

[9] Currently, the level of access to this database is determined by the district superintendent. In some districts, all teachers have ready access to the data; in others, district administrators review the data and develop periodic reports for teachers.

standards. As a result, new expectations for student growth will need to be set. Starting with 2009 results, TDOE will reset the growth standard to reflect the state's average student performance in 2009; going forward, the state's performance in 2009 will replace the 1998 performance as the baseline year. According to TDOE (2009a), implementation of the new baseline year "offers an opportunity to delineate among schools that have, on average, made more than expected progress with its students. This will allow the possibility for all schools to meet these new and higher standards in future years."

In the following three sections, the ways in which TVAAS data are used to monitor student achievement and evaluate teacher and school effects are explored.

Section I: Using TVAAS Data to Monitor Changes in Student Achievement

> The Tennessee experience suggests that where local leadership has provided the opportunities for teachers and principals to learn to use the reports provided, then cynicism has been replaced by teachers asking why more information cannot be supplied more quickly (Sanders, 2000, p. 336).

Data from TVAAS are intended to be used diagnostically to improve educational opportunities for students at all achievement levels.[10] To do so, TVAAS analyses model effects at a number of levels. Concurrently, TVAAS models district or "system" effect (mean score for that grade and year), current teacher effect (current classroom membership), past teacher effect, and error variance (systematic and nonsystematic).

Students' scores across years and subjects are statistically linked (Braun, 2005). For TVAAS analyses, a student's first-year score is subtracted from his/her second-year score; this becomes the student's raw gain score during growth analyses. Past and current teacher effects and variance terms then are added to the raw gain score in the equation, in addition to the error terms. This outcome then is compared to the mean gain for that grade in that district. A student demonstrates growth by achieving a score that exceeds the expected in comparison to the classroom mean, and an effective school or teacher is one whose performance profile (mean of aggregated student scores) shows higher than expected average growth, relative to district norms.[11]

As with other projection models, TVAAS also predicts how a student is likely to perform in the future. By using historical data, i.e., all of each student's prior test scores, projections of performance on tests up to three years in the future can be modeled. These projections have been found to be more reliable than a single score from the adjacent year (Sanders & Wright, 2009; Wright, Sanders, & Rivers, 2006).

The Tennessee model differs from other value-added models in two key ways. First, while the TVAAS model *implicitly* accounts for students' initial status (e.g., prior achievement or previous test scores) by allowing students to act as their own controls, it does not assume that growth is linear nor take into account the interaction between initial status and growth during modeling (Goldschmidt et al., 2005; Sanders & Rivers, 1996). Instead, it uses multiple cohorts and panel data to adjust for prior achievement while examining "layered" gains,[12] assuming that student scores are uniformly affected by the teacher each year and that this effect persists and is cumulative (Sanders, Saxton, & Horn, 1997). This allows each

[10] Per state law, students with disabilities and students with low attendance records are excluded from value-added analyses.

[11] Use of separate growth norms is intended to address regression to the mean due to measurement error and other factors, or the tendency of score gains for students with the lowest levels of prior achievement to exceed those for students with the highest levels of prior achievement (Campbell & Kenny, 1999).

[12] I.e., teacher effects build annually from the effects of previous years.

student's most recent scores to be compared over time only to his or her own previous test scores (Sanders & Horn, 1998).

Second, TVAAS analyses do *not* include student background characteristics during modeling (Ballou, Sanders, & Wright, 2004). That is, unlike other value-added models, a student's socioeconomic status, race/ethnicity, gender, language background, etc. are not explicitly entered into the TVAAS equations as covariates. According to Sanders and Wright (2009), "At the student level, if the entire multivariate longitudinal vector is fitted, then the inclusion of SES variables is not needed to ensure fairness" (p. 4). In their 2004 study of the impact of including covariates on the precision of estimates derived from TVAAS analyses, Ballou, Sanders, and Wright explain the rationale for this methodology as follows:

> Measuring student progress requires controlling in some fashion for initial level of achievement…Introducing a prior test score as a regressor controls for initial achievement, so that the contribution of schools and teachers to student progress is based on residual differences in the post-test scores. Because the value-added method measures gain from a student's own starting point, it implicitly controls for socioeconomic status and other background factors to the extent that their influence on the post-test score is already reflected in the pre-test score. (p. 38)

Further discussion of this model feature is presented in greater detail in Sections II and III.

Recent Example: Using TVAAS Data to Examine Student-Level Achievement
In Spring 2009, TVAAS data were used to predict the future ACT scores of sixth grade students in Tennessee (TDOE, 2009a). Using longitudinal data sets, the TDOE conducted both prospective (how might these students perform in grade 11 or 12, assuming an average school experience in Tennessee?) and retrospective (how did these students perform in the past?) analyses of student-level data. The prospective analyses were intended to predict (1) the number and percentage of students in grade 6 in 2008 who have at least a 50% chance of scoring at a particular level on the ACT and (2) the probability that grade 6 students scoring at each proficiency level (low, middle, high) on the state test will reach particular performance levels on the ACT.

Results were of interest to a wide range of stakeholders, including those pondering the degree of postsecondary preparedness of Tennessee students:

- 42% of grade 6 students were found to have at least a 50% chance of achieving a composite score of 21 on the ACT, which qualifies them for the Tennessee Hope Scholarship.
- 55% of grade 6 students had at least a 50% chance of achieving a score of 19 on the ACT mathematics portion, which is the cut-off for remedial instruction in freshman-level college mathematics.
- 27% of grade 6 students had at least a 50% chance of achieving an ACT mathematics score of 22, the cut-off score for first year college algebra.
- 4% of grade 6 students had at least a 50% chance of achieving an ACT mathematics score of 26, which is the average score for Math, Science, Engineering, and Technical field college graduates in Tennessee.

- 39% of grade 6 students had at least a 50% chance of achieving an ACT Science Reasoning score of 21, the average score for all graduates of Tennessee colleges.
- 12% of grade 6 students had at least a 50% chance of achieving an ACT Science Reasoning score of 24, the benchmark score for projected success in first-year college biology.
- A student who consistently scored in the high range on the state test had a 69% chance of achieving a score of 19 on the mathematics portion of the ACT and a 50% chance of achieving a score of 21 on the science portion.
- A student who consistently scored in the middle range on the state test had a 26% chance of achieving a score of 19 on the mathematics portion of the ACT and an 18% chance of achieving a score of 21 on the science portion.
- A student who consistently scored in the low range on the state test had a 5% chance of achieving a score of 19 on the mathematics portion of the ACT and a 4% chance of achieving a score of 21 on the science portion.

In the retrospective analyses, 2008 ACT scores were compared for each district, with a mean district score reported at four different achievement levels (TDOE, 2009a). Using ACT test takers' grade 6 mathematics and science state test scores, students were divided into quartiles. These data were used to rank each district by their means and to show the relative location of each district by quartile. Graphs were generated that displayed which districts' students outperformed those in other districts on the ACT, despite comparable grade 6 achievement.

These analyses highlighted disparities in ACT performance across districts in this state. A set of recommendations emerged that ranged from encouraging educators to access the diagnostic tools and customized student reports available on the TVAAS web site to formulating long-term strategies for building on district strengths and leveraging highly effective teaching.

Section II: Using TVAAS Data to Evaluate Teacher Effect

> If a curriculum is viewed as a ramp—not as stair steps—…differences in schooling effectiveness is the dominant factor affecting the speed that students move up the "ramp." …Teachers have primary control of the speed that students move up the "ramp" (Sanders, 2000, p. 331).

As discussed in Section I, TVAAS uses test scores from multiple test events and a series of complex analyses at the system (mean district performance), school (mean school performance), and teacher (mean student performance for each teacher[13]) levels to model student progress over time (Sanders, Saxston, & Horn, 1997). Data are analyzed and effects estimated individually for each school district (system). For those students showing improvement (i.e., score gains between test events) relative to district norms, a proportion of that gain is attributed to the professional efforts of teachers (TDOE, 2007).

Research suggests that instructionally effective classroom teachers have a strong impact on student learning (Betebenner, 2004; Wright, Horn, & Sanders, 1997) and that effects of quality teaching tend to persist and accumulate (Rivkin, Hanushek, & Kain, 1998; Sanders & Rivers, 1996). Effective teachers can enable student progress toward valued learning goals (Darling-Hammond, 2000) and foster optimal performance on standardized tests of achievement (Heneman, Kimball & Milanowski, 2006; Hershberg & Simon, 2004; Holtzapple, 2003). TVAAS is based on the premise that "…teacher effectiveness is the single largest factor affecting academic growth of populations of students (Sanders, 2000, p. 334). In Tennessee, this means that the most effective teachers in each district are those whose students experience the largest average performance gains on statewide tests of achievement, relative to the district average.

Since 1998, the degree to which each Tennessee teacher has added value, as measured by average score gains, is one component of the teacher's thorough annual evaluation process (Sanders, 2000). With endorsement from the Tennessee Education Association (TEA), growth scores based on a three-year-average estimated mean gain score from teacher reports[14] can be used as up to 8% of a teacher's evaluation. The TDOE/TEA partnership is ongoing as they work collaboratively to provide professional development opportunities for both high- and low-performing teachers related to appropriate use of TVAAS data for effecting change in student achievement.

The practice of evaluating teacher effect based on student performance on standardized tests is methodologically challenging and remains controversial (Amrein-Beardsley, 2008; Fisher, 1996; Koretz, 2002; Kupermintz, 2003; Meyer, 1996; Millman & Schalock, 1997; Shrinkfield & Stuffelbeam, 1995; Valli, Croninger

[13] The TVAAS model constrains teacher effects to average to zero within each school system and represents teacher effects as independent, additive, and linear. A combined estimate of teacher gains is computed by adding the teacher effect to the system average gain. For this reason, teachers with fewer student scores are more likely to have means close to the district mean (Kupermintz, 2003; Sanders, Saxston, & Horn, 1997).

[14] The reports cite an average of data from three successive student cohorts. Each year, the oldest cohort is dropped from analyses and the more recent one included.

& Walters, 2007).[15] Even with improvements in capacity to collect longitudinal data, advances in testing practices and statistical modeling, and increased efforts to collect evidence to support the validity of test results, questions persist about the appropriateness of the measures used and the accuracy of estimates derived from such evaluations of teacher effect (Aaronson, Barrow, & Sander, 2007; Braun, 2005; Kupermintz, 2003).

As students are *not* randomly assigned to schools or classrooms *nor* are teachers randomly distributed across schools, one such challenge is accounting for possible alternate explanations for score gains (Rivkin & Ishi, 2008; Rothstein, 2008; Rubin, Stuart, & Zanutto, 2004). Student, subgroup, or school characteristics may systematically impact the context for learning, learning outcomes, or test performance (Ballou, 2002; Ballou, Sanders, Wright, 2004; Betebenner, 2004; Raudenbush & Bryk, 2002; Talbert & McLaughlin, 1993). Within-classroom variation associated with student prior achievement is another key consideration, as are between-classroom factors such as classroom size and content-specific resources and teaching practices (Darling-Hammond & Post, 2000). For example, teachers with seniority may have inflated effect scores due to assignment to classrooms comprised of highly engaged students rather than to instructional effectiveness (Braun, 2005). Finally, particularly in middle and high schools or as a result of mid-year mobility at any grade, students can be linked to more than one teacher for instruction, thereby creating the need for sophisticated statistical adjustments to account for these conditions (McCaffrey, Lockwood, Koretz, & Hamilton, 2003).

These factors can lead to biased estimates of teacher effect (Kupermintz, 2003), confound interpretation of test results (McCaffrey et al., 2004), and result in possible erroneous ranking or misclassification of teachers (NASBE, 2005). In their 2008 synthesis of research on teacher effectiveness, Goe, Bell, & Little expressed concern that by focusing primarily on standardized test results in judgments about teacher effect, value-added models may oversimplify the range of indicators associated with instructional effectiveness and provide little guidance to stakeholders about *why* effects vary within and across schools (i.e., what do effective teachers do differently in their classrooms?). As Sanders (2000) acknowledged, "…analyses at the teacher level require the utmost care and caution and present even more burden on the statistical methodology, the computing software, and the data archiving process itself" (p. 334).

For this reason, TVAAS methodology incorporates a number of statistical safeguards to support the validity of the process (Sanders, 2000). These include the following:

(1) statistical adjustments allow student prior achievement (starting place) to be taken into consideration but not treated as a covariate for analyses.
(2) each teacher's effect is estimated against the local norm (i.e., his/her own school and district/ system means) as well as against the state mean.
(3) each year's data are linked to current and previous teachers ("layering").
(4) individual teacher reports are based on multiple years of data, i.e., a three-year average of estimated gains.

[15] It is important to note that, while terms may appear to be used interchangeably, TVAAS literature generally refers to these analyses as evaluation of teacher *effect*, not teacher *effectiveness*.

In addition, TVAAS analyses rely on best linear unbiased prediction, or shrinkage estimates (Bock & Wolfe, 1996).[16] According to Sanders and Wright (2009), these estimation procedures are used to "…provide maximum correlation between the estimate and 'true effect,' "…protection against spurious estimates due to too little data," and "…greater repeatability between estimates in adjacent years" (p. 3). That is, shrinkage estimates have been found to reduce statistical noise or interference when seeking to isolate teacher (classroom) effects on student achievement (McCaffrey et al, 2004).

Finally, because of the complex, dynamic, and cumulative effects of the interactions among the contextual and individual factors associated with student learning, different types of information from multiple sources (e.g., classroom observations, work portfolios) are considered *in addition* to test scores during decision-making about teacher effectiveness in Tennessee (TDOE, 2007). Nevertheless, additional research on the criterion-related validity of annual state test scores for purposes of modeling teacher effect would bolster claims that TVAAS data are appropriate for this purpose and that results may be interpreted as trustworthy indicators of teachers' instructional skills (Amrein-Beardsley, 2008; Bock & Wolfe, 1996; Fisher, 1996). According to TDOE (2009b), such research would be welcomed.

Recent Examples: Using TVAAS Data to Examine Teacher Effect

In November 2008, TVAAS data were used a part of a state-mandated annual evaluation of the 39 teacher training programs in Tennessee institutions of higher education. The goal of the evaluation was to identify those programs that tended to produce new teachers who were evaluated as highly effective or ineffective (Tennessee State Board of Education, 2008). For purposes of this study, each elementary and middle school teacher with 1–5 years of experience was assigned a t-value effect score,[17] based on the average gain in learning for their students.[18] "Highly effective" teachers were defined as those whose t-value effect scores were in the highest quintile of the state distribution for their content area(s) and grade. This group was compared to those in the lowest quintile ("least effective" teachers). Each teacher training program received data that described the number and percentages

Teacher effects were estimated in mathematics, reading/language arts, science, and social studies. As shown in the table below, on average, teachers in the highly effective group had estimated teacher gains that were 4.9, 6.5, and 5.6 standard errors greater than their districts' means in mathematics, science and social studies, respectively, while the least effective teachers had estimated gains that were 3.8, 3.4, and 3.1 standard errors less than their districts' means in those content areas (Tennessee State Board of Education, 2008).

[16] Also called empirical Bayes estimation (Raudenbush & Bryk, 2002) or shrinkage estimation (Sanders & Wright, 2009). According to Fisher (1996), the teacher effect is specified as random while the school effect is fixed. See Ballou, Sanders, & Wright (2004) for detailed description of TVAAS estimation procedures or McCaffrey, Lockwood, Koretz et al.(2004) for general model specifications.

[17] According to TDOE (2008), the t-value of the teacher effect is the teacher effect estimate (relative to the district gain) divided by its standard error. This measure was used (1) to address concerns about lack of randomization (teachers are not randomly assigned to districts) and discrepancies in the numbers of students associated with each teacher and (2) to enable multi-grade comparisons.

[18] Only one year's data (2008) were included in estimating effect scores.

Table 1. Difference in T-Value Effect Scores for New Teachers, Grades 4-8

	Lowest Quintile	Highest Quintile	Difference
Mathematics	-3.8	4.9	8.7
Reading/ELA	.75	4.3	3.6
Science	-3.4	6.5	9.9
Social Studies	-3.1	5.6	8.7

While final decision-making about program quality included examination of data about placement and retention as well as Praxis scores, the TVAAS effect data served as one valuable source of information for exploring disparities in preparedness among teachers in this state.

A second recent example of research that capitalized on data collected through TVAAS was a 2007 comprehensive analysis of teacher experience and education levels that looked at the distribution of effective teachers across state schools (TDOE, 2007). This study was conducted as part of the Tennessee Teacher equity plan, approved by the TDOE. Of particular interest was comparing the qualifications of teachers in schools that serve high versus low proportions of students in poverty and of minority students. Two key findings were as follows:

- Beginning teachers are overrepresented in high-poverty schools and high-minority schools.
- Fewer teachers with master's degrees teach in high-poverty schools and high-minority schools.

Since research suggests that teacher experience and education alone may not be strong predictors of classroom effectiveness, analyses using TVAAS data also were conducted (Sanders, Wright, & Rivers, 2006; Sanders, Ashton, & Wright, 2005; Sanders, 2000; Sanders & Horn, 1998; Wright, Paul, & Sanders, 2007). Using teacher effect scores reported relative to the state mean for student growth for each grade and content area, teachers were categorized by teacher effect scores into those who were most effective, least effective, or between. Findings indicated that the state's most effective teachers made up a smaller percentage of the teaching staff (18% vs. 21%) and the least effective teachers made up a larger percentage of the teaching staff (24% vs. 16%) in high poverty/high minority schools. Some disparity also emerged in the mean effectiveness levels of teachers in low poverty/low minority schools vs. high poverty/high minority schools. The TDOE is using these data to educate state policy-makers about potential inequities in educational opportunity in their state.

Emerging Context for Data Use to Support Teacher Evaluation

In the announcement for Race to the Top funds (USED, July 2009), Proposed Priority 5–Invitational Priority–School Level Conditions for Reform and Innovation, the following eligibility requirement appears:

> A state must not have any legal, statutory, or regulatory barriers to linking student achievement or student growth data to teachers for the purpose of teacher and principal evaluation. Research indicates that teacher quality is a critical contributor to student learning and that there is dramatic variation in teacher quality. Yet it is difficult to predict teacher quality based on the qualifications that teachers bring to the job. Indeed, measures such as certification, master's degrees, and

> years of teaching experience have limited predictive power on this point. Therefore, one of the most effective ways to accurately assess teacher quality is to measure the growth in achievement of a teacher's students, and by aggregating the performance of students across teachers within a school, to assess principal quality…This capability is fundamental to Race to the Top reform and to the requirements in Section 14005(d)(2) of the ARRA that States take actions to improve teacher effectiveness…these plans must require LEAs and schools to determine which teachers and principals are effective using student achievement data. (p. 37811)

The limitations of existing tools for measuring (or predicting) teacher effectiveness remain a key incentive to Sanders and his team of researchers as they continue to promote value-added modeling as a component of teacher evaluations (Sanders, Wright, & Rivers, 2006; Sanders, Ashton, & Wright, 2005; Sanders, 2000; Sanders & Horn, 1998; Wright, Paul, & Sanders, 2007). Based on this funding requirement, it would appear that states who currently use or plan to use student achievement data to measure teacher effect will have an advantage in securing Race to the Top funds.

In addition, in the announcement for Race to the Top funds (USED, July 2009), an "effective teacher" is defined as follows:

> *Effective teacher* means a teacher whose students achieve acceptable rates (e.g., at least one grade level in an academic year) of student growth (as defined in this notice[19]). States may supplement this definition as they see fit so long as teacher effectiveness is judged, in significant measure, by student growth (as defined in this notice. (p. 37811)

This eligibility requirement and clarification of terms suggest that research associated with the TVAAS and other VAMs were carefully considered when developing the most recent federal guidelines for funding. Clearly, standards for best practice in data use are emerging as a result of the ground-breaking work in states like Tennessee.

[19] See pg. 5 for definition of *student growth*.

Section III: Using TVAAS Data to Examine School Effect

> …now we are using data more responsibly. We have added a tremendous amount of sophistication to our analyses. Now we can pinpoint which achievement level of kids in a particular teacher's classroom are doing very well and which ones are not doing so well, and we can furnish this information to the practitioner. This is information that teachers, principals, and other educational decision-makers need, if they are to do the best they can for every student in their school (Sanders, 2000, p. 338).

While the focus of data collection and reporting for TVAAS growth modeling is on teacher effect, evaluation of school effect is of primary interest for accountability purposes. As described in Sections I and II, in TVAAS analyses, demonstrated cumulative school gains are interpreted as related primarily to instructional factors rather than to racial and socioeconomic factors (Sanders, 1998; Ballou, Sanders & Wright, 2004). A portion of student performance gains are attributable to a teacher effect, with the aggregate effect across all classrooms averaged for a school effect. In short, the mean teacher effect becomes the de facto school effect. Non-instructional contextual factors specific to a school, such as extra-curricular activities offered or location, are assumed to be part of the teacher effect and so are not considered separately in TVAAS school effect estimations (McCaffrey, Lockwood, Koretz & Hamilton, 2003).

District-level administrators and school-level principals play powerful roles in determining the extent to which TVAAS data will be used to promote instructional improvement in Tennessee schools. According to TDOE (2009b),

> Support from leaders really matters. If superintendents, curriculum supervisors, and principals buy into data use, changes are happening that benefit students. It may take time, but those systems whose leaders are asking questions, studying data, and meeting with teachers about the meaning of TVAAS results are way ahead of the others.

Research supports this observation. School and district leaders, particularly principals, have been shown to play a pivotal role in the success of school reform efforts (Bryk, Sebring, Kerbow, Rollow & Easton, 1998; Fullan, 2001; McLaughlin & Talbert, 2001).

As discussed in Sections I and II, TVAAS does not explicitly incorporate student, teacher, *or* school-level covariates in calculating school effect (Sanders & Horn, 1994; Sanders & Wright, 2009).[20] This design feature has led to a number of questions about the validity of TVAAS data at all levels: Must all growth models take into account those factors known to covary with test performance (Amrein-Beardsley, 2008; Coleman, 1990; Raudenbush & Bryk, 2002; Goldstein, 1997; Kupermintz, 2003; McCaffrey et al., 2004)? According to Wright and Sanders (2009), the answer is no:

> …adjustment for group SES factors will over-adjust the estimates and can camouflage the fact that students in certain schools are not getting an equitable distribution of the teaching talent. The

[20] As described in Section I, TVAAS is a Layered Mixed Effects Model that uses multiple cohorts and panel data to adjust for prior achievement while examining "layered" gains (Sanders, Saxton, & Horn, 1997). Each student's most recent scores to be compared over time only to his or her own previous test scores (Sanders & Horn, 1998).

> answer to whether or not to adjust for group SES variables depends on where the risks are to be placed. Even though we advocate for no adjustment, we certainly can make SES group adjustments if states and districts elect (pp. 4-5).

Clearly, careful consideration of tradeoffs associated with this decision is critical in designing a school effects model that satisfies all statistical requirements for accountability purposes yet also is guided by consequential validity concerns. This dilemma is described below by a team of researchers led by Tekwe, Carter, Ma, and Algina (2004):

> One [school effects] model might be preferred in a low-stakes accountability system that provides incentives and resources for "less effective" schools to improve and does not base salary raises on the value-added measures. In a high stakes system, however, not adjusting for significant sociodemographic factors could encourage the flight of good teachers and administrators from schools with high percentages of poor or minority students. On the other hand, adjusting for those factors could institutionalize low expectations for poor or minority students and thereby limit their opportunities to achieve their full potential. (p. 31)
>
> …It should be noted that if schools are partly but not wholly responsible for the effects of covariates, then bias results from *either* including or excluding them. Assuming partial responsibility, the exclusion of student and school level covariates from our analyses produced a bias against schools with an overrepresentation of, for example, poverty or minority students. On the other hand, if schools were at least partially responsible for the effects of these covariates, then including them resulted in value-added measures that were biased against schools with an underrepresentation of minority or poverty students. (p. 31)

While questions remain about the ideal system for modeling growth for multiple purposes, Tennessee schools *are* using TVAAS data in a number of constructive ways to support student growth and school improvement (TDOE, 2009b). Schools use TVAAS data to project future performance and to identify those grade levels, classrooms, and/or content areas that are above or below the expected in terms of growth. These data are useful in selecting and planning curriculum and in targeting professional development so it is most effective. Teachers, programs, and practices that are found to contribute to above average student growth may be used as mentors or exemplars for staff experiencing less significant academic gains. School leadership teams may help teachers reach instructional goals by including TVAAS data at weekly meetings to help address specific existing or emerging needs at the teacher-, grade- and content area-levels. Within classrooms, teachers may analyze growth patterns or other value-added data about students in their classrooms and use these data formatively in refining their instructional strategies. Use of these data in this way supports teachers' efforts to meet the needs of all students more effectively and to support the individual academic growth of their students regardless of their prior test scores or ability level.

As when measuring teacher effect, responsible use of TVAAS data to monitor school effect is associated with a number of challenges. As a statistical procedure, distinctions among schools that are well above or below average are most reliable (Goldstein, 1997). The impact of school characteristics, such as percent of students

eligible for subsidized lunches or its racial/ethnic composition, are difficult to untangle from impact on academic achievement when standardized test scores are the measure (Berk, 1998; Braun, 2005; Fisher, 1996; Ladd & Walsh, 2002; Rubin, Stuart, & Zanutto, 2004; Thum & Bryk, 1997). Research in Tennessee to date suggests that a school's racial/ethnic composition, percentage of subsidized lunch-eligible students, and mean achievement level are not related to the cumulative growth in performance across all state schools (Ballou, Sanders, & Wright, 2004; Sanders & Horn, 1998). However, the TDOE is encouraged by recent requests for data and hopes that planned research activities will continue to shed light on this concern (TDOE, 2009b).

Conclusion: Strengths and Challenges Associated with the TVAAS

> Growth models make demanding assumptions and enforce strong requirements on both data and users. They attempt refined answers to very specific questions. The functional and logical relationships among data elements tightly constrain logic and inference, method and conclusions. Their precision is responsible for their value. …Growth models require principled knowledge. They force us to think hard and clearly about our questions, the evidence chain we need, and the instrumentalities including metrics of the data elements that comprise the evidence chain. That is a good and necessary result, although hardly an easy one. (CCSSO, 2007, p. 33)

TVAAS is complex and costly. It requires large data sets (multiple measures and student cohorts) and a sophisticated longitudinal database with the capacity to track students over time and across schools (Lockwood, Louis, & McCaffrey, 2002; McCaffrey et al., 2004; Noell, 2005). Estimates rely on psychometric assumptions about the state's assessment system and annual standardized achievement tests (Ballou, 2005; Doran & Cohen, 2005; Schmidt, Houang & McKnight, 2005). TVAAS was designed specifically by William Sanders in 1997 to meet Tennessee's needs and remains dependent on proprietary estimation procedures (SAS) to operate. For this reason, this model has not been viewed as having wide applicability in other states for accountability purposes.[21]

TVAAS administrators face ongoing challenges. These include continuing to communicate with state constituents who regularly benefit from access to these data that:

- value-added results are not available to all Tennessee teachers because state tests are administered only at certain grades and in certain content areas.
- the reporting of growth data in conjunction with status and improvement (Safe Harbor) for federal accountability purposes may not result in significant numbers of new schools meeting AYP targets.
- TVAAS will continue to meet educators' and policymakers' needs despite a changing state context in which new state content and performance standards are emerging.
- district and school leaders must remain vigilant in monitoring the ways in which students and teachers are assigned to classrooms.

In addition, as members of the educational research community—as well as other nations, states, and districts—continue to ask tough questions, TVAAS administrators have the responsibility of ensuring that

- as the primary measure of learning and proxy for effective teaching[22]—annual state assessments (1) are aligned to state standards and comparable over time in terms of content breadth, depth, and rigor; (2) meet the most stringent technical quality expectations for reliability, validity, and freedom from bias; (3) include sufficient numbers and types of items of varying difficulty levels to allow for effective discrimination among a range of achievement levels at each grade; and (4) are monitored regularly to ensure fidelity to standardized administration conditions and responsible test preparation practices.

[21] Nonetheless, Pennsylvania developed and recently piloted their PVAAS using comparable methodology.
[22] Description of test scores coined by Rabinowitz in 2004 presentation.

- sufficient annual growth at each level remains linked to attainment of a defensible, research- and data-supported performance standard or benchmark.
- an ambitious research agenda is pursued to collect evidence that that the findings that emerge from analyses are trustworthy for the purposes intended and that inferences drawn from findings are meaningful and appropriate.

Yet the pioneering work of the TDOE in implementing and refining the TVAAS has been the catalyst for a dynamic conversation within the educational policy community and a revitalized focus on the quality of student learning for all students. Because of the innovative and sophisticated tools developed in Tennessee to collect, analyze, and report student performance data focused on growth, attention to the need for instructionally-sensitive assessments and knowledge about the effectiveness of instructional programs and practices have increased, and meaningful incentives—not just sanctions—for school improvement have emerged (TDOE, 2009b). In Dr. Sander's words, "The use of summative value-added measures as one component of accountability systems is important; but in our view, the diagnostic information is of greater importance (Sanders & Wright, 2009, p. 8)..."Just looking at proficiency is not enough to get us where we want to go…"[23]

Value-added modeling in Tennessee has brought to light the degree to which effective teaching is distributed across all classrooms, schools, and districts in one state, thereby creating an ongoing conversation among stakeholders about the value of instructional staff and school leaders and about targeted, research-based support for struggling teachers and schools. TVAAS data are being used as descriptive feedback to guide instructional planning and to inform decision-making about what works best for certain students or groups of students. Reports are developed that present educators with progress rates for students at each achievement level, individual projections for each student relative to various academic goals, numbers of students on track to enter rigorous coursework at the next grade level, and longitudinal data to help untangle the factors associated with achievement gaps. Teachers and administrators have access to meaningful information about changes in student achievement over time and support in using these data appropriately and effectively. With endorsement from the state's largest professional teaching organization, teachers and principals now have a quantitative component in their annual evaluations that helps them better understand their short- and long-term impact on student achievement.

In her 2002 history of the standards-based accountability movement, Vaughan drew the following conclusions about TVAAS:

> Sanders's system has already demonstrated possibilities of truly transforming our nation's schools to serve all of our children in many areas of the country more productively. The system has additionally allowed many policy makers, taxpayers, and parents see how teachers are helping students to learn. Many research studies have found that teacher effectiveness has a greater impact on children's academic achievement and subsequent success than either poverty or perpupil expenditures. School success is directly related to this concept. (p.7)

[23] Comment captured at Roundtable discussion on value-added analysis sponsored by the Working Group on Teacher Quality.

APPENDIX A: Sample Value Added Report for a System

2009 TVAAS System Value Added Report
TCAP CRT Math

The Tennessee Department of Education has reset the growth standard to reflect the state's present student progress. Shading below is consistent with this new minimal expectation for systems and schools. The Help link above includes the specific details of this transition year.

Estimated System Mean NCE Gain								
Grade:	3	4	5	6	7	8	Mean NCE Gain over Grades Relative to	
Growth Standard:		0.0	0.0	0.0	0.0	0.0		
State 3-Yr-Avg:		-0.3	-0.1	0.3	0.1	-0.4	Growth Standard	State
2007 Mean NCE Gain:		0.9 G	-5.9 R*	-0.3 Y	-0.2 Y	0.8 G	-0.9	-0.9
Std Error:		1.3	1.4	1.3	1.1	1.2	0.6	0.6
2008 Mean NCE Gain:		-1.8 R	-5.7 R*	2.6 G*	-1.9 R	-1.2 R	-1.6	-1.5
Std Error:		1.5	1.3	1.4	1.2	1.2	0.6	0.6
2009 Mean NCE Gain:		-7.7 R*	-8.2 R*	4.1 G*	1.5 G*	1.1 G	-1.8	-1.8
Std Error:		1.4	1.5	1.3	1.2	1.2	0.6	0.6
3-Yr-Avg NCE Gain:		-2.9 R*	-6.6 R*	2.1 G*	-0.2 Y	0.2 G	-1.5	-1.4
Std Error:		0.8	0.8	0.8	0.7	0.7	0.2	0.2

Estimated System Mean NCE Scores						
Grade:	3	4	5	6	7	8
New State Baseline:	50.0	50.0	50.0	50.0	50.0	50.0
State 3-Yr-Avg:	48.9	48.4	48.3	47.9	48.1	47.9
2006 Mean:	49.1	49.3	49.7	51.9	45.6	42.4
2007 Mean:	49.9	50.0	43.5	49.4	51.7	46.4
2008 Mean:	51.8	48.2	44.3	46.1	47.5	50.5
2009 Mean:	50.3	44.1	40.0	48.4	47.6	48.6

G* - Estimated mean NCE gain above the growth standard by at least 1 standard error.

G - Estimated mean NCE gain equal to or greater than growth standard but by less than 1 standard error.

Y - Estimated mean NCE gain below the growth standard by 1 standard error or less.

R - Estimated mean NCE gain more than 1 standard error below the growth standard but by 2 standard errors or less.

R* - Estimated mean NCE gain below the growth standard by more than 2 standard errors.

To view additional reports, click on the underlined numbers or words.

APPENDIX B: Sample Value Added Report for a School

2009 TVAAS School Value Added Report
TCAP CRT Math

The Tennessee Department of Education has reset the growth standard to reflect the state's present student progress. Shading below is consistent with this new minimal expectation for systems and schools. The Help link above includes the specific details of this transition year.

Estimated School Mean NCE Gain					
Grade:	6	7	8	Mean NCE Gain over Grades Relative to	
Growth Standard:	0.0	0.0	0.0		
State 3-Yr-Avg:	0.3	0.1	-0.4	Growth Standard	State
2007 Mean NCE Gain:	-1.4 R*	-0.6 R	0.2 G	-0.6	-0.6
Std Error:	0.6	0.6	0.6	0.4	0.4
2008 Mean NCE Gain:	-2.9 R*	1.3 G*	3.8 G*	0.7	0.7
Std Error:	0.6	0.6	0.6	0.4	0.4
2009 Mean NCE Gain:	-2.4 R*	-0.9 R	2.6 G*	-0.2	-0.2
Std Error:	0.7	0.6	0.6	0.4	0.4
3-Yr-Avg NCE Gain:	-2.2 R*	-0.1 Y	2.2 G*	-0.0	-0.1
Std Error:	0.4	0.4	0.3	0.2	0.2

Estimated School Mean NCE Scores			
Grade:	6	7	8
New State Baseline:	50.0	50.0	50.0
State 3-Yr-Avg:	47.9	48.1	47.9
2006 Mean:	61.3	60.7	61.5
2007 Mean:	61.0	60.6	60.9
2008 Mean:	61.5	62.3	64.4
2009 Mean:	65.1	60.6	64.9

G* - Estimated mean NCE gain above the growth standard by at least 1 standard error.

G - Estimated mean NCE gain equal to or greater than growth standard but by less than 1 standard error.

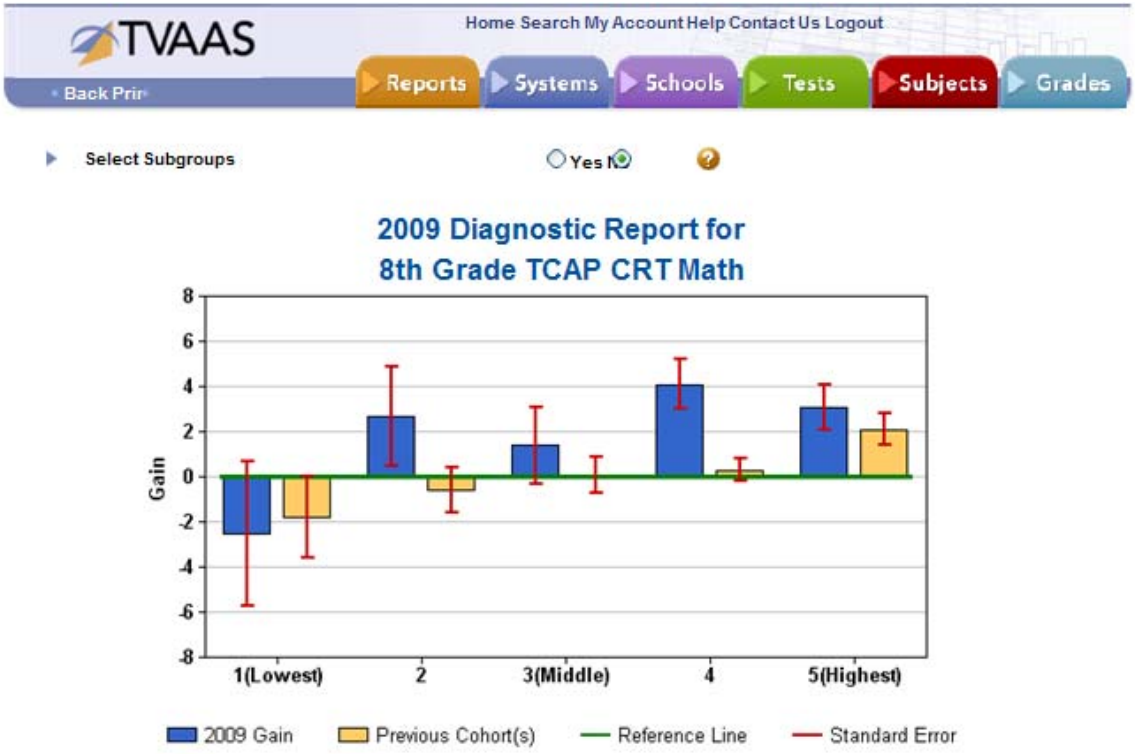
Y - Estimated mean NCE gain below the growth standard by 1 standard error or less.

R - Estimated mean NCE gain more than 1 standard error below the growth standard but by 2 standard errors or less.

R* - Estimated mean NCE gain below the growth standard by more than 2 standard errors.

To view additional reports, click on the underlined numbers or words.

APPENDIX C: Sample Diagnostic Report for a School

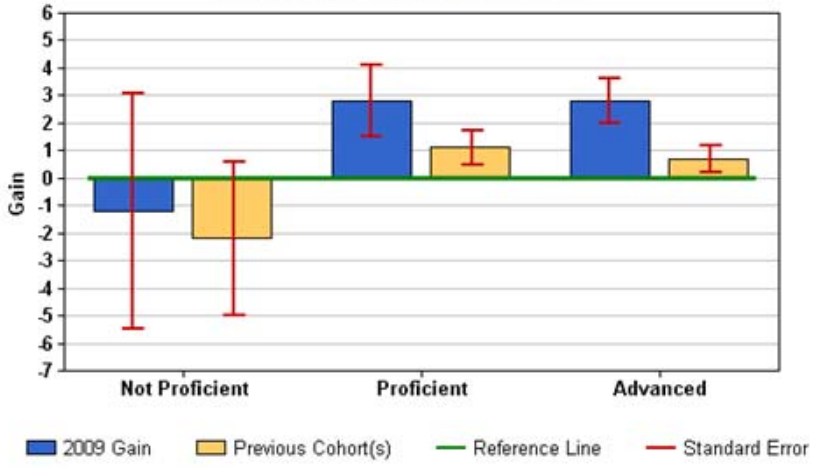


			Prior-Achievement Subgroups				
			1 (Lowest)	2	3 (Middle)	4	5 (Highest)
Math	Reference Line		0.0	0.0	0.0	0.0	0.0
	2009	Gain	-2.5	2.7	1.4	4.1	3.1
		Std Err	3.2	2.2	1.7	1.1	1.0
		Nr of Students	18	21	46	74	130
		%of Students	6.2	7.3	15.9	25.6	45.0
	Previous Cohort (s)	Gain	-1.8	-0.6	0.1	0.3	2.1
		Std Err	1.8	1.0	0.8	0.5	0.7
		Nr of Students	54	79	145	272	323
		%of Students	6.2	9.0	16.6	31.2	37.0

APPENDIX D: Sample Performance Diagnostic Report for a School

Select Subgroups

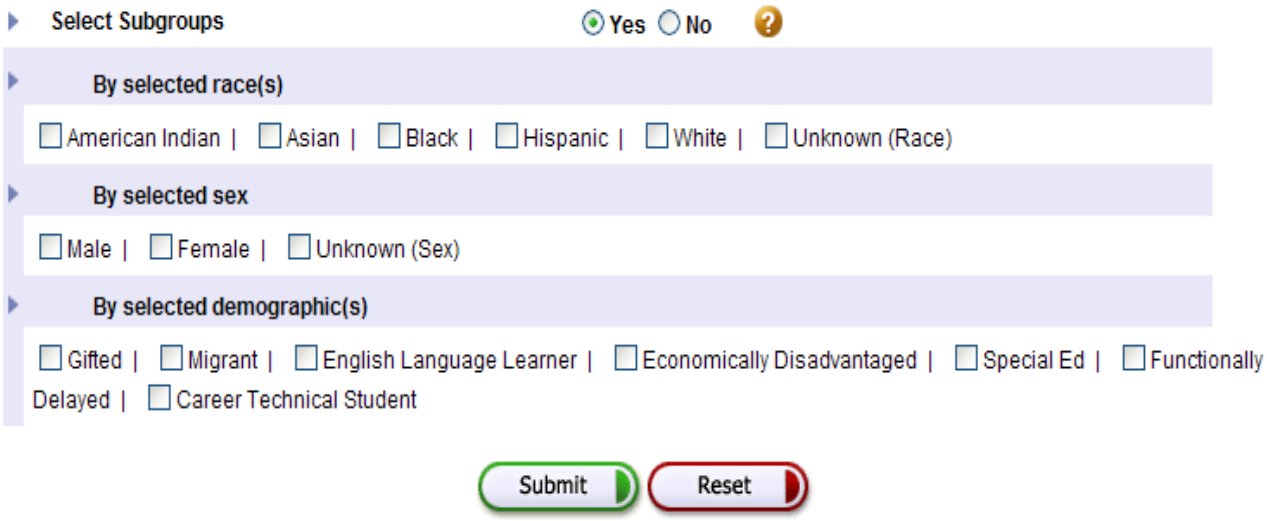
2009 Performance Diagnostic Report
8th Grade TCAP CRT Math



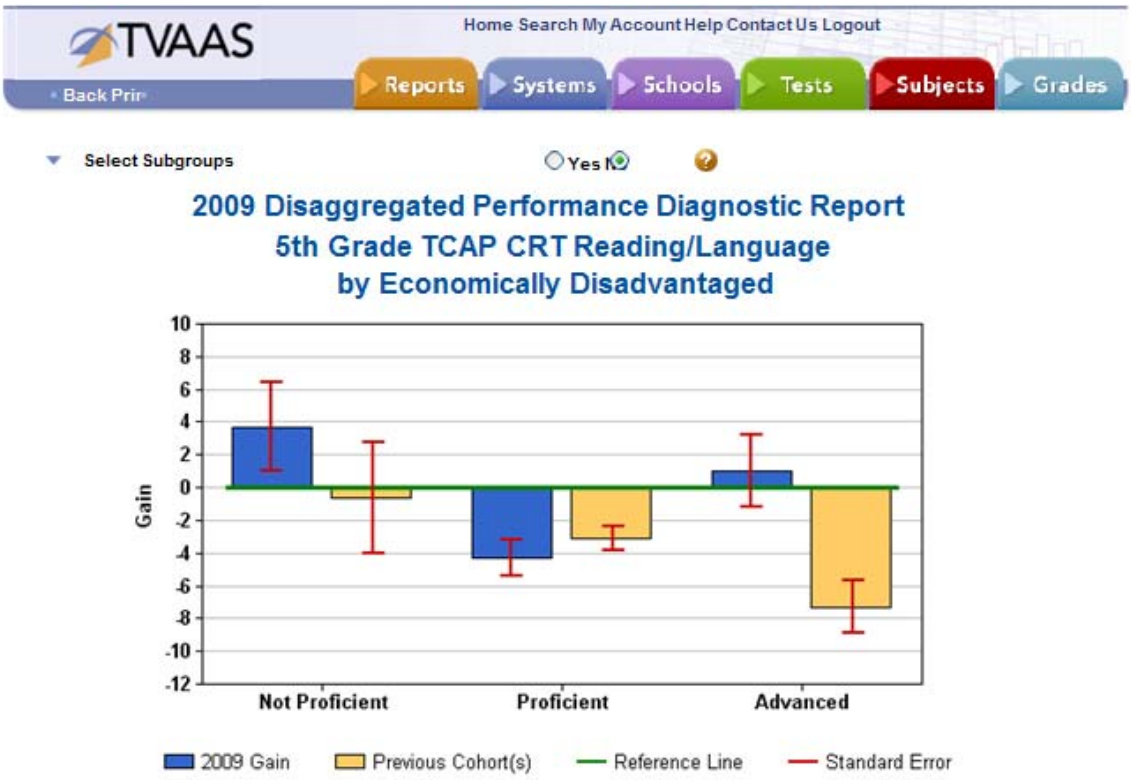
			Predicted Proficiency Group		
			Not Proficient	Proficient	Advanced
Math	Reference Line		0.0	0.0	0.0
	2009	Gain	-1.2	2.8	2.8
		Std Err	4.3	1.3	0.8
		Nr of Students	9	66	214
		%of Students	3.1	22.8	74.0
	Previous Cohort(s)	Gain	-2.2	1.1	0.7
		Std Err	2.8	0.6	0.5
		Nr of Students	24	269	579
		%of Students	2.8	30.8	66.4

APPENDIX E: “Select Subgroups” option within Diagnostic or Performance Diagnostic Report

Select any of the following subgroups of students to view Disaggregated Diagnostic or Performance Diagnostic report



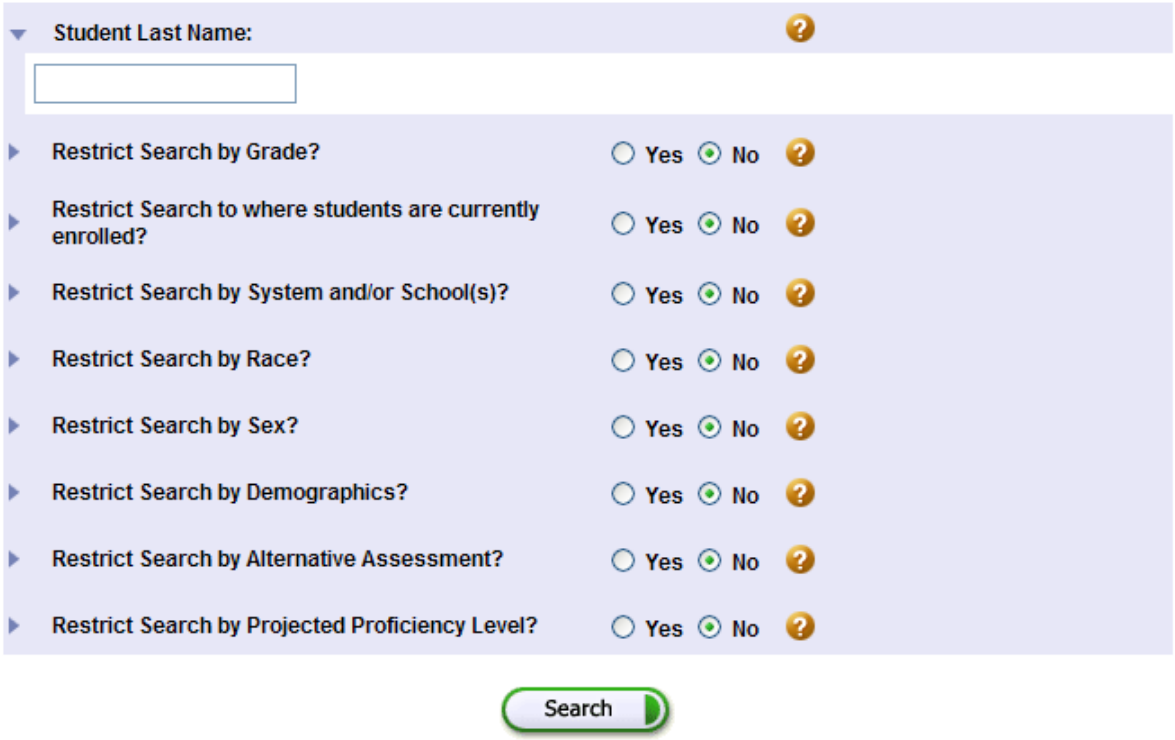
APPENDIX F: Sample Disaggregated Performance Diagnostic Report



			Predicted Proficiency Group		
			Not Proficient	Proficient	Advanced
Reading/Language	Reference Line		0.0	0.0	0.0
	2009	Gain	3.7	-4.3	1.0
		Std Err	2.7	1.1	2.2
		Nr of Students	11	120	17
		%of Students	7.4	81.1	11.5
	Previous Cohort(s)	Gain	-0.6	-3.1	-7.3
		Std Err	3.4	0.7	1.6
		Nr of Students	27	294	49
		%of Students	7.3	79.5	13.2

APPENDIX G: Creating Custom Student Reports

Create Custom Student Reports by selecting one or a combination of choices below.



APPENDIX H: Sample Saved Custom Student Report

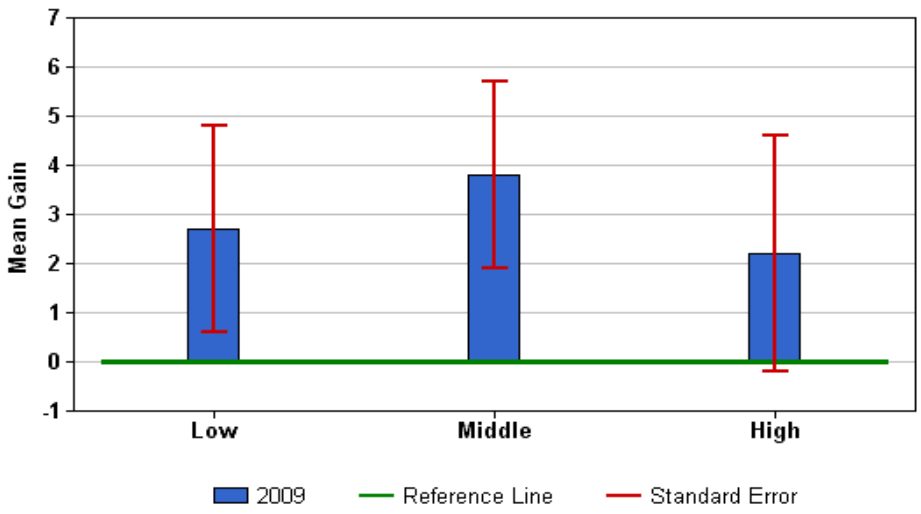
Gateway Algebra I (Proficient)

	Remove	Student	System	School	Sex	Race	Grade	Gif	Mig	ELL	ED	SpED	Achievement Probability
1.	☐	AILES, THANH	Beta School District	Debby Middle School	F	W	7	N	N	N	N	Y	0.4
2.	☐	ALTROGGE, BUFORD	Beta School District	Debby Middle School	M	W	7	N	N	N	N	N	99.1
3.	☐	ARNZEN, COLE	Beta School District	Debby Middle School	M	W	7	Y	N	N	N	N	100.0
4.	☐	ASP, PETER	Beta School District	Debby Middle School	M	W	7	N	N	N	Y	Y	21.5
5.	☐	BEASON, KARA	Beta School District	Debby Middle School	F	W	7	N	N	N	Y	N	45.8
6.	☐	BELLUS, ALEC	Beta School District	Debby Middle School	M	B	7	N	N	N	Y	N	85.9
7.	☐	BICKFORD, JANET	Beta School District	Debby Middle School	F	B	7	N	N	N	Y	N	69.2
8.	☐	BORKOWSKI, ROSS	Beta School District	Debby Middle School	M	W	7	N	N	N	N	N	6.1
9.	☐	BORNER, RORY	Beta School District	Debby Middle School	M	W	7	N	N	N	N	N	95.9
10.	☐	BOTTO, ALVA	Beta School District	Debby Middle School	M	B	7	N	N	N	Y	N	88.0
												Mean	61.2
												Std Err	12.5

2009 TCAP CRT (Grade 4): Math Students

Select	Student Name	2008 State NCE	2009 State NCE	Avg State NCE	2009 Percentile	Perf Level	School Name
☐	BEGLEY, LETICIA	34	42	38.0	27	P	Chris Middle School
☐	BIGLOW, KRISTOFER	49	51	50.0	40	P	Chris Middle School
☐	BILLIEL, JULIANNE	47	31	39.0	11	NP	Chris Middle School
☐	BOREEN, ARCHIE	53	52	52.5	43	P	Chris Middle School
☐	BREAKELL, BETH	49	47	48.0	34	P	Chris Middle School
☐	BRICENO, LUCY	45	41	43.0	24	P	Chris Middle School
☐	BRISSETT, KATELYN	21	31	26.0	11	NP	Chris Middle School
☐	CANNEY, LYNNE	41	41	41.0	24	P	Chris Middle School

APPENDIX I: Sample Student Pattern List

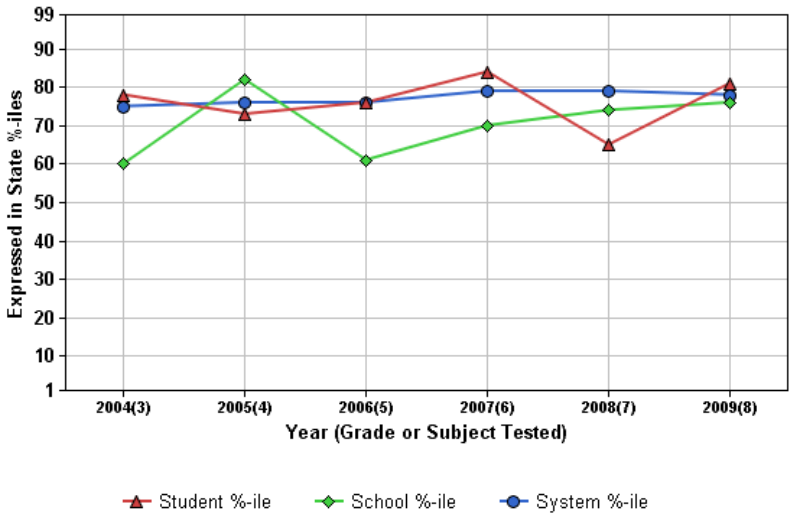


Mean Gain		
Low	Middle	High
2.7	3.8	2.2

Students by Subgroup		
Low	Middle	High
HORACIO GWALTNEY	HARLAN YZAGUIRRE	DEANN AUBIN
KENYA GEITNER	TROY ZOLLO	BERRY GUGEL
RICO CHUKES	ESMERALDA DORNFELD	ALMA HUDY
DEIDRA KEATE	SHANA BUEHRLE	DEON SCHWEBKE
CHELSEY MCDERMOTT	ISABELLE TUEY	ESTER FERANDEZ
MARGO LEPPER	JEWELL CLAEYS	MAXWELL JAHDE
TARYN FALTERMAN	JUNIOR BLUMENTHAL	EVERETTE DIMLER

APPENDIX J: Sample Student Report

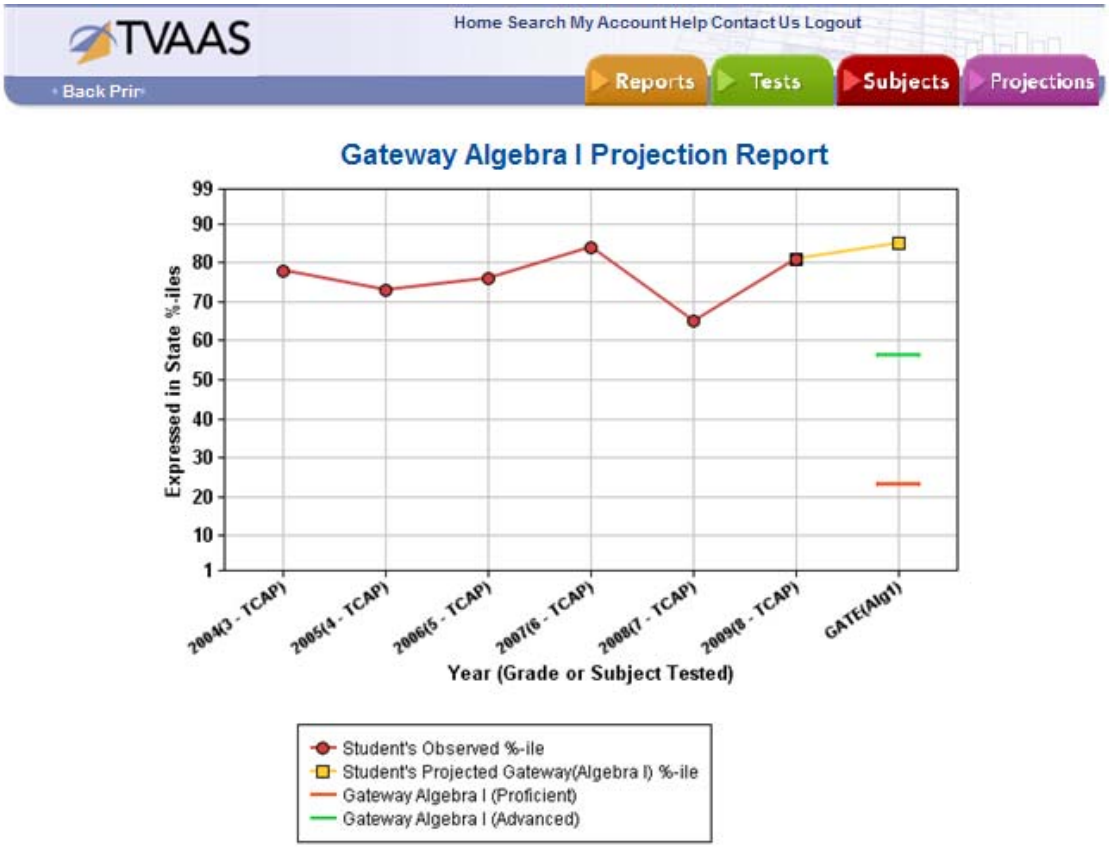
2009 TCAP CRT: Math Student Report



Subject: Math						
	Year (Grade or Subject Tested)					
	TCAP CRT (Math)					
	2004(3)	2005(4)	2006(5)	2007(6)	2008(7)	2009(8)
State NCE \ Score	67	63	65	72	58	69
%-ile	78	73	76	84	65	81
Perf Level	AD	AD	AD	AD	AD	AD

Performance Levels:
NP - Not Proficient
P - Proficient
AD - Advanced

APPENDIX K: Sample Projection Report for a Student



Projection: Gateway Algebra I		
Projected State Percentile	Probability of Success	
	Proficient	Advanced
85	99.8%	92.1%

Student's Testing History						
	Year (Grade or Subject Tested)					
	TCAP CRT (Math)					
	2004(3)	2005(4)	2006(5)	2007(6)	2008(7)	2009(8)
State NCE \ Score	67	63	65	72	58	69
%ile	78	73	76	84	65	81

References

Aaronson, D., Barrow, L, & Sander, W. (DePaul). (2007). Teachers and student achievement in the Chicago Public Schools. *Journal of Labor Economics, 25*(1), 95-135.

Amrein-Beardsley, A. (2008). *Methodological concerns about the Education Value-Added Assessment System*. Educational Researcher, 37(2), 65-76.

Ballou, D. (2002). Sizing up test scores. *Education Next, Summer 2002*, 10-15.

Ballou, D. (2005). Value-Added assessment: Lessons from Tennessee. In R. Lissitz (Ed.), *Value-Added Models in education: Theory and applications*. Maple Grove, MN: JAM Press.

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), 37-65.

Berk, R. (1998). Fifty reasons why student gain does not mean teacher effectiveness. *Journal of Personnel Evaluation in Education, 1*, 345-363.

Betebenner, D. (2008). Norm- and Criterion-Referenced Student Growth. *Nashua, NH: NCIEA.*

Black , P & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Bock, R. & Wolfe, R. (1996). *A review and analysis of the Tennessee Value-Added Assessment System (Part 1)*. Nashville, TN: Comptroller of the Treasury.

Bratton, S, Horn, S. & Wright, S. (1996). *Using and interpreting Tennessee's Value-Added Assessment System: A primer for teachers and principals.* Knoxville, TN: University of Tennessee.

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Educational Testing Service Policy Information Perspective. Princeton, NJ: ETS.

Bryk, A., Sebring, P., Kerbow, D., Rollow, S. & Easton, J. (1998). *Charting Chicago's School Reform*. Boulder, CO: Westview Press.

Bryk, A., Thum, Y., Easton, J. & Luppescu, S. (1998). Assessing school academic productivity: The case of Chicago school reform. *Social Psychology of Education, 2,* 103-142.

Campbell, D. & Kenny, D. (1999). *A primer on regression artifacts*. New York: Guilford Press.

Choi, K., Goldschmidt, P. & Yamashiro, K. (2005). Exploring models of school performance: From theory to practice. In J. Herman and E. Haertel (Eds.), *National Society for the Study of Education, 104*.

Choi, K., Seltzer, M., Herman, J., & Yamashiro, K. (2004). *Children left behind in AYP and Non-AYP schools: Using student progress and the distribution of student gains to validate AYP*. CSE Rep: No. 637. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Chrispeels, J., Brown, J. & Castillo, S. (2000). School Leadership Teams: Factors that influence their development and effectiveness. *Understanding Schools as Intelligent Systems*, Vol. 4, 39-73, JAI Press.

Council of Chief State School Officers. (2007). *Implementer's Guide to Growth*. Paper Commissioned by the CCSSO Accountability Systems and Reporting State Collaborative on Assessment and Student Standards.

Darling-Hammond., L. & Post, L. (2000). Inequality in teaching and schooling: Supporting high quality teaching and leadership in low-income schools. In R.D. Kahlenberg (Ed.), *A notion at risk: Preserving public education as a engine for social mobility*. New York: Century Foundation.

Darlington, R. (1997). The Tennessee Value-Added Assessment System: A challenge to familiar assessment methods. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* 163-168.

Doran, H. & Cohen, J. (2005). The confounding effects of linking bias on gains estimated from Value-Added Models. In R. Lissitz (Ed.), *Value-Added Models in education: Theory and applications*, (pp. 80-104). Maple Grove, MN: JAM Press.

Fisher, T. (1996). *A review and analysis of the Tennessee Value-Added Assessment System (Part 2).* Nashville, TN: Comptroller of the Treasury.

Fullan, M. (2001). *The new meaning of educational change, 3rd Ed*. New York City: Teachers College Press.

Goe, L., Bell, C. & Little, O. (2008) *Approaches to Evaluating Teacher Effectiveness: A research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goldschmidt, P. (2004). *Growth Models*. Paper presented at the CCSSO Conference on Large Scale Assessment, Boston, MA. June.

Goldschmidt, P. & Choi, K. (2006). *The practical benefits of growth models for accountability and the limitations under NCLB*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Goldschmidt, P., Roschewski, P., Choi, Auty, W., Hebbler, Blank, R. & Williams, 2005). *Policymakers' guide to growth models for school accountability: How do*

accountability models differ. A paper commissioned by the CCSSO Accountability Systems and Reporting State Collaborative on Assessment and Student Standards. Washington, DC: CCSSO.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement, 8*(4), 369-395.

Gong, B., Perie, M. & Dunn, J. (2006). *Using student longitudinal growth measures for school accountability under No Child Left Behind: An update to inform design decisions*. Accessed August 7, 2009 from ww.nciea.org.

Hanushek, E., Rivkin, S. & Taylor, L. (1996). Aggregation and the estimated effects of school. *The Review of Economics and Statistics, 78*(4), 611-627.

Heneman, H., Kimball, S. & Milanowski, A. (2006). *The teacher sense of efficacy scale: Validation evidence and behavioral prediction*. WCER Working Paper No. 2006-7. Madison, WI: Wisconsin Center for Educational Research.

Heritage, M. & Yeagley, R. (2005). Data use and school improvement: Challenges and prospects. In J. L. Herman & E. H. Haertel, (Eds.), *Uses and misuses of data for educational accountability and improvement. National Society for the Study of Education. Yearbook of the National Society for the Study of Education Vol. 104*(2), pp. 320–339. Chicago: Blackwell Publishing.

Hershberg, T., Simon, V. & Lea-Kruger, B. (2004). *Measuring what matters*. American School Board Journal, 19(2), 27-31.

Herman, J., Yamashiro, K., Lefkowitz, S. & Trusela, L. (2008). *Exploring data use and school performance in an urban public school district.* CRESST Report 702. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17(*3), 207-219.

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

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25*(3), 287-298.

Ladd, H. & Walsh, R. (2002). Implementing value-added measures of school effectiveness: *Getting the incentives right. Economics of Education Review, 21*, 1-17.

Lockwood, J., Louis T. & McCaffrey, D. (2002). Uncertainly in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics, 27*(3), 255-70.

Lockwood, J., McCaffrey, D., Mariano, L. and Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics, 32*, 125-130.

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

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), 67-101.

McLaughlin, M. & Talbert, J. (2001). *Professional communities and the work of high school teaching*. Chicago: University of Chicago Press.

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

Millman, J. & Schalock, H. (1997). Beginnings and introduction. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid measure?* (pp. 3-10). Thousand Oaks, CA: Corwin Press.

Noell, G. (2005). *Assessing teacher preparation program effectiveness: A pilot examination of value-added approaches.* Accessed August 12, 2009 from www.asa.regents.state.la.us/TE/digest.pdf.

Osgood, W. & Smith, G. (1995). Applying hierarchical linear modeling to extended longitudinal evaluations. *Evaluation Review, 19*(1), 3-39.

Phelps, R. (2008). *Correcting fallacies about educational and psychological testing*. Washington, DC: American Psychological Association.

Rabinowitz, S. (2004). *Vertically articulated grade-level expectations and test specifications*. Paper presented at the annual Reidy Interactive Lecture Series, Nashua, NH.

Raudenbush, S. & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed)*. Thousand Oaks, CA: Sage.

Rivkin, S. & Ishi, J. (2008). *Impediments to the estimation of teacher value-added*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI.

Rivkin, S., Hanushek, E. & Kain, J. (1998). *Teachers, schools, and academic achievement*. National Bureau of Economic Research Working Paper #6691.

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

Rothstein, J. (2008). *Do value-added models add value? Tracking, fixed effects, and causal inferences.* Paper presented at the National Conference on Value-Added Modeling, Madison, WI.

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

Rumberger, R. W. & Palardy, G. J. (2004). Multilevel models for school effectiveness research. In D. Kaplan (Ed.), *Handbook on quantitative methodology for the social sciences* (235-258). Thousand Oaks, CA: Sage. Accessed September 9, 2009 from http://www.facultydirectory.ucr.edu/cgi-bin/pub/public_individual.pl?faculty=3193.

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

Sanders, W. (2000). Value-added assessment from student achievement data: opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14*(4), 329-339.

Sanders, W. (2006). *Comparisons among various educational assessment value-added models.* Paper presented at the Power of Two–National Value-Added Conference, Columbus, Ohio.

Sanders, W., Ashton, J. & Wright, S. (2005). *Comparison of the effects of NBPTS certified teachers with other teachers on the rate of student academic progress*. Arlington, VA: National Board for Professional Teaching Standards. Accessed August 9, 2009, from http://www.nbpts.org/UserFiles/File/SAS_final_NBPTS_report_D_-_Sanders.pdf

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

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

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

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

Sanders, W. & Wright, P. (2009). *A response to Amrein-Beardsley (2008), 'Methodological concerns about the education value-added system*.' Accessed from Tennessee Department of Education on September 24, 2009.

Sanders, W., Wright, P. & Rivers, J. (2006). Measurement of academic growth of individual students toward variable and meaningful academic standards. In R.W. Lissitz (Ed.), *Longitudinal and value-added models of student performance.* Maple Grove, MN: Journal of Applied Measurement Press.

Schmidt, W., Houang, R. & McKnight, C. (2005). Value-Added research: Right idea but wrong solution? In R. Lissitz (Ed.), *Value-Added Models in education: Theory and applications,* (pp.145-164). Maple Grove, MN: JAM Press.

Seltzer, M., Choi, K., & Thum, Y. (2003). Examining relationships between where students start and how rapidly they progress: Implications for conducting analyses that help illuminate the distribution of achievement within schools. *Educational Evaluation and Policy Analysis, 25*(3), 263-286.

Shrinkfield, A. & Stuffelbeam, D. (1995). *Teacher evaluation: guide to effective practice*. Boston, MA: Kluwer Academic Publishers.

Singer, J. & Willet, J. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: University Press.

Stone, J. (2002). *The value-added achievement gains of NBPTS-board certified teachers in Tennessee: A brief report*. Accessed September 1, 2009 at http://www.education-consumers.com/oldsite/briefs/stoneNBPTS.shtm.

Stringfield, S., Wayman, J. & Yakimowski, M. (2005). Scaling up data use in classrooms, schools and districts. In C. Dede, J. Honan & L. Peters, (Eds.), *Scaling up success: Lessons learned from technology-based educational innovation.* San Francisco: Jossey-Bass.

Talbert, J. & McLaughlin, M. (1993). Understanding teaching in context. In D.K. Cohen, M.W. McLaughlin and J.E. Talbert (Eds.), *Teaching for understanding: Challenges for practice, research, and policy*, (pp. 167-206). New York: Jossey-Bass.

Tennessee Department of Education. (2009a). *Academic preparedness in Tennessee*. Accessed August 7, 2009 from http://tennessee.gov/education/doc/TN_Academic_Preparedness.pdf.

Tennessee Department of Education. (2009b). Semi-structured telephone interview conducted on September 8, 2009.

Tennessee Department of Education. (2007). Tennessee's most effective teachers: Are they assigned to the schools that need them most? Research Brief. Accessed September 5, 2009 from http://tennessee.gov/education/nclb/doc/TeacherEffectiveness2007_03.pdf

Tennessee State Board of Education. (2008). *Report card on the effectiveness of teacher training programs.* Report prepared in conjunction with the TDOE and the Tennessee Higher Education Commission.

Tekwe, C., Carter, L., Ma, C., Algina, J., Lucas, M., Roth, J., Ariet, M., Fisher, T., & Resnick, M. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Staistics*, 29(1), 11-35.

Thum, Y. (2003). *No Child Left Behind: Methodological challenges and recommendations for measuring adequate yearly progress*. CSE Technical Report 590. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Thum, Y. & Bryk, A. (1997). Value-added productivity indicators: The Dallas system. In Jason Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid measure?* (pp.100-119). Thousand Oaks, CA: Corwin Press.

United States Department of Education. (2009). Part III, Race to the Top Fund. *Federal Register, July 29, 2009*. Washington, DC: USED.

Valli, L., Croninger, R. & Walters, K. (2007). Who (else) is the teacher? Cautionary notes on teacher accountability systems. *American Journal of Education, 113*(4), 635-662.

Vaughan, A. (2002). Standards, accountability, and the determination of school success. *The Education Forum, 66*(3), 206-213.

Wayman, J. (2005). Involving teachers in data-driven decision-making: Using computer data systems to support teacher inquiry and reflection. *Journal of Education for Students Placed At-Risk, 10*(3), 295-308.

Weisberg, H. (1979). *Statistical adjustments and uncontrolled studies. Psychological Bulletin, 86*(5), 1149-1164.

Wright, S., Horn, S. & Sanders, W. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*(1), 57-67.

Wright, S., Sanders, W. & Rivers, J. (2006). Measurement of academic growth of individual students toward variable and meaningful academic standards. In R. Lissitz (Ed.), *Longitudinal and value-added models of student performance*. Maple Grove, MN: JAM Press.