

Running Head: MEASURES OF SCHOOL PERFORMANCE

From Federal AYP to Value Added Models:

What are the Most Meaningful and Valid Measures of School Performance?

Henry May

Michael Weiss

Brooke Snyder Taylor

Consortium for Policy Research in Education

University of Pennsylvania

Paper Presented at the Annual Meeting of the

American Education Research Association,

San Francisco, CA

April 11, 2006

## Abstract

This paper explores several techniques for estimating school performance that vary in complexity of the statistical model. Resistance from teachers and educators and their distrust of the “black-box” that accompanies complicated Value Added Assessment (VAA) modeling is not surprising. In order for VAA-based accountability systems to gain more support from teachers, the VAA model itself must be clear and straightforward. The option to choose a less complicated model would be even more attractive if the results produced were similar to those of more complicated VAA models. Results from analysis of data from a large Florida school district suggest that overly complicated VAA models may be unnecessary and produce results that are less understandable to non-statisticians.

### From Federal AYP to Value Added Models:

#### What are the Most Meaningful and Valid Measures of School Performance?

It has become widely accepted that most NCLB measures of Adequate Yearly Progress (AYP) are not good indicators of school performance. This is largely because federal AYP captures achievement status instead of gains. NCLB does have a measure of performance gains in the form of the “safe harbor” provision; however that approach calculates progress for each grade level by comparing achievement scores for that grade to the scores of a different cohort of students in that grade the year before. Research has shown that these estimates of change are heavily influenced by shifts in student characteristics from one year to the next (Kane & Staiger, 2002).

An intense amount of political pressure is building to force revision of current accountability systems to include better measures of growth in student performance (Alliance for Fair and Effective Accountability, 2004; Olsen, 2005), and several states have already adopted growth measures in their accountability systems (Olsen, 2004). Although considerable research has been conducted to develop and compare alternative methods for calculating school-level growth measures, nearly all of this work is heavily technical, making it largely meaningless to non-statisticians (for examples, see Sanders, Saxon, & Horn, 1997; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Tekwe et al., 2004). Building upon and extending this work, the purpose of this paper is to describe and compare, using actual data from a large urban school district, alternative approaches to measuring growth in school performance through analyses of student achievement test scores in such a way that maximizes the utility of information for policymakers. Our primary question is “Are complex statistical models for value-added

assessment helpful and necessary, or can useful and valid information be obtained with less complicated, less expensive models?”

Value-added assessment models (VAA) are a class of statistical models that link individual student achievement scores over time, with the primary goal of identifying the degree to which students' learning gains are a function of student characteristics versus characteristics of teachers or schools. The primary intent of VAA is to determine the causal impact of teachers or schools on the learning gains of their students (Raudenbush, 2004). When used as part of an accountability system, the ideal goal of a VAA model is to hold teachers and schools accountable for that part of their students' performance that they can influence through policy and practice, but not to hold them accountable for any part of their students' performance for which they have no control (e.g., that driven by neighborhood characteristics and social context). As argued by Raudenbush and Willms (1995) and by Rubin, Stuart, and Zanutto (2004), it is difficult if not impossible to estimate the true causal effects of teachers and schools with VAA models because it is hard to disentangle the effects of educational policy and practice from the effects of social context. Nevertheless, Rubin, Stuart, and Zanutto (2004) point out that the estimates from VAA models may still be useful if the accountability structures based on VAA results actually lead to improvements in educational outcomes. So, if we stop thinking about VAA effects as representing the causal effect of teachers and schools, and instead, we think about VAA effects as descriptive measures of student performance, then any improvement in educational outcomes that results from the use of VAA estimates is still a good thing.

This way of thinking is the impetus behind the research presented here. If VAA models will never be able to estimate the true causal effects of teachers and schools, then what is the “value added” of using complicated and overly elaborate statistical models for VAA that almost

no one can understand? Resistance from teachers and educators and their distrust of the “black-box” that accompanies VAA modeling is not surprising. In order for VAA-based accountability systems to gain more support from teachers, the VAA model itself must be clear and straightforward in its formulas for calculating student performance. The option to choose a less complicated VAA model would be even more attractive if the results produced by these models were similar to those from more complicated models. This paper explores several techniques for estimating school performance that vary in complexity of the statistical model and require different degrees of technical expertise from the analyst implementing them.

### *Indicators of School Performance*

The measurement of school performance in this paper is presented consisting of two components: a measure of current status and a measure of growth from the previous year to the current year. The two most common metrics for describing school performance status are used in this paper (percent proficient and average scale score) so that they can be correlated with growth measures and a clearer interpretation can be made for each measure of growth.

Six alternative measures of growth in performance are presented. Using the method for calculating gains under NCLB as the baseline, we present models of increasing complexity designed to deal with five specific issues relevant to calculating school performance growth. The five issues, expressed as policy questions, are:

1. Should gains be calculated using proficiency scores or scale scores?
2. Should gains be calculated by comparing school aggregate scores for sequential cohorts of students (the NCLB approach) or by comparing school aggregate scores for the same

cohort of students (e.g., comparing the percent proficient for 8<sup>th</sup> grade this year to the percent proficient for 7<sup>th</sup> grade last year)?

3. Should gains be calculated after linking individual student achievement scores over time so that students who are mobile can be tracked and so that school level scores are derived from the individual gains of students in that school during the time of testing?
4. Instead of attaching each student's gains to only the school in which they were enrolled at the time of testing, should student gains be apportioned to schools based upon the proportion of time a student was enrolled in that school during that school year?
5. Should individual gains be adjusted for prior performance and other student background characteristics? If so, how should this be done?

### *Sample and Measures*

For this analysis we use student-level vertically scaled scores from the Florida Comprehensive Assessment Test (FCAT) in reading for the 2002-03 through 2004-05 school years for all middle schools from Duval County Florida. The Duval County Public School system is the 20<sup>th</sup> largest school district in the United States and the seventh largest in Florida, serving over 129,500 students enrolled in 182 schools. The ethnic composition of the students in this urban school district is 46% white, 43% black, and 5% Hispanic with 49% of students receiving free or reduced-price lunch assistance.

The focus of this study is on 29 middle schools in Duval County containing at least 20 students in eighth grade in spring 2005. For the purposes of illustration, we calculate school performance and growth measures for 8<sup>th</sup> grade in each school for the 2004-05 school year. Growth measures were calculated using 2005 FCAT reading test scores from the criterion

referenced portion of the FCAT assessments and up to two years of prior FCAT reading test scores, depending on the requirements of the model used to calculate growth.

### *Methods*

Our analysis consists of a series of models which can be used to address the five policy questions presented above. The models fall into three categories, school-level current status, school-level change models, and student-level change models. Each model produces separate estimates for the eighth grade in each of the 29 schools for the 2004-05 school year. The models are as follows:<sup>1</sup>

#### *School-level Current Status.*

Current status measures provide a good picture of student achievement levels across schools. While easy to interpret, status measures do not provide information about change over time or student growth; as a result, status measures alone do not capture school effects on student learning since they ignore students' prior achievement. This begs the question, "Are current student achievement levels attributable to school effectiveness or simply a result of a school being fortunate to have students who entered the school with already high achievement levels?"

1. **Percent proficient** – the percentage of students scoring at or above the proficiency cut-score. This is the measure used under NCLB to determine if a school meets "Adequate Yearly Progress (AYP)". Any school with a proficiency rate below the minimum cutoff for the state for that year is identified as not meeting AYP, regardless of the amount of improvement in the school's scores from the previous year. However, NCLB does include a "safe harbor" stipulation designed to recognize growth (see method 3 below).

The most desirable property of percent proficient as a status measure is its ease of

---

<sup>1</sup> For a mathematical representation of each of these methods, see Appendix A.

interpretation and understandability for a wide audience. From a policy perspective, a school's proficiency rate can also help focus teachers on bringing students up to a minimum standard.

2. **Average scale score** – the average or mean scale score. This status measure differs from the proficiency category metric described above, and instead uses the linear scale scores produced by the state test. The scale of the scores is arbitrary, so familiarity with the scale is necessary to inform the interpretation of these scores. For example, the metric may be similar to that used for the SAT, with each subject having normally distributed scores with a mean of 500 and a standard deviation of 100. If the state test is vertically scaled, then the scores increase along a single scale from one grade to the next.

#### *School-level Change Models.*

School-level change models produce indicators of change, but because they are calculated as differences between the average score for a grade or cohort in a school from one year to the next, they are influenced by student mobility and shifts in student characteristics from year to year or cohort to cohort. As such, these school level indicators of change do not provide good estimates of student learning because they are influenced by changes in the student population.

3. **Change in percent proficient for successive cohorts of eighth-graders** – calculated as the difference in the percentage of 8th graders scoring proficient in spring 2004 subtracted from the percentage of 8th graders scoring proficient in spring 2005. This is the approach advocated by NCLB's "safe harbor" provision. Note that this model compares school-level scores for different cohorts of students, thus results may be highly affected by changing demographics and/or within-school cohort variability. Although

this measure of change is also easily interpreted like the status measure described above, proficiency measures do not account for any improvement unless the improvement leads to students surpassing the proficiency cut-score.

4. **Change in average scale scores for successive cohorts of eighth-graders** - calculated as the difference in the average scale score for 8th graders for spring 2004 subtracted from the average scale score for 8th graders for spring 2005. Like the previous model, this model compares school-level scores for different cohorts of students and is influenced by cohort to cohort differences. Unlike the previous model, changes in the average scale score do not require that students' scores surpass the proficiency cut-score in order to show improvement.
5. **Change in percent proficient for the same cohort of students** – calculated as the difference in the percentage of 7th graders scoring proficient in spring 2004 subtracted from the percentage of 8th graders scoring proficient in spring 2005. Note that this model compares school-level scores for the same cohort of students, helping to account for within-school cohort variability. Student mobility and retention in grade may still influence this metric, but it is much less sensitive to cohort shifts than methods using successive cohorts of students from the same grade. On the other hand, because this measure compares proficiency levels across adjacent grades it is sensitive to differences in how proficiency is defined for each grade. For example, if standards for 8<sup>th</sup> grade are more rigorous than those for 7<sup>th</sup> grade, then this metric would show a decrease in the percent proficient despite an increase in scores.
6. **Change in average scale scores for the same cohort of students** – calculated as the difference in the average scale score for 7th graders for spring 2004 subtracted from the

average scale score for 8th graders for spring 2005. Like the previous model, this model compares school-level scores for the same cohort of students. Unlike the previous model, the use of scale scores eliminates the problem of differences in the meaning of proficiency across grades. When test scores are vertically scaled, the change estimates produced with this approach may be used to approximate average learning gains, though mobility and grade retention can reduce accuracy. If mobility and retention rates are low, then this model is very similar to the next model, average individual change.

#### *Student-level Change Models.*

Student-level change models are based on gains in individual achievement scores, which are then aggregated to the school level. This produces a more direct measure of individual learning gains and is less influenced by student mobility and shifts in the characteristics of the student population.

7. **Average Individual Change** – After linking individual student scores from spring 2004 and spring 2005, each student's scale score from spring 2004 is subtracted from his/her scale score in spring 2005 to produce an individual gain score from 2004 to 2005. School performance gains are calculated as the average of the individual gain scores for all students enrolled in the 8<sup>th</sup> grade in spring 2005. Because this method captures annual gains for each student regardless of previous mobility or retention in grade, it is less sensitive to the problems affecting school-level change models. However, if a student changes schools during the school-year, then this model incorrectly ascribes that student's gains entirely to the new school. This is not a problem if mobility rates during the school year are low. For this method to work well, it is important that scores be measured on an interval scale so that a gain of  $x$  at one point on a scale is equivalent to a

gain of  $x$  at another point on that same scale. Most scaled scores from state assessments are designed to satisfy this requirement.

8. **Layered Value-Added Model** – After linking individual student scores for three years from spring 2003 to 2005, a statistical model designed by Sanders, Saxton, and Horn (1997) is used to estimate each school’s contribution to its students test score gains from 2004 to 2005 after controlling for previous test scores from 2003 (i.e., the model includes three years of test scores). These estimates are called “value-added” scores, and are identical to the like-named scores used in the Tennessee Value Added Assessment System (TVAAS). These scores represent achievement score growth relative to the average growth for all schools in the district, after adjusting for prior individual student achievement scores. The VAA scores are also able to credit proportionate amounts of a student’s gains to more than one school if they switch schools during the school year. Another nuance is that VAA scores for smaller groups of students are shrunken towards the mean VAA score, making it less likely that small schools will be identified as exceptionally high or low performing (for details, see McCaffrey, Lockwood, Koretz, & Hamilton, 2003, pp. 63-68). Furthermore, VAA effects are assumed to be normally distributed – an assumption which is not appropriate if there are substantially more low performing schools than high performing schools, or vice versa. The metric of VAA estimates can be represented as a school averaged gain (similar to the estimates from method 7 above) or as the deviation of the school averaged gain from the overall average gain where negative values suggest below average performance and positive values

suggest above average performance.<sup>2</sup> Our results present VAA estimates as deviations from the average gain divided by their standard errors. This allows VAA estimates to be quickly identified as statistically significantly above ( $>1.96$ ) or below ( $<-1.96$ ) the average VAA estimate with 95% confidence.

### *Results*

Detailed results are shown in Tables 1 through 3 with status and change measures for each of the 29 middle schools are presented in Table 1, correlations between measures shown in Table 2, and school ranks under each measure presented in Table 3. Schools in Tables 1 and 3 are ordered from lowest to highest proficiency rate in terms of 2005 status.

The two things that are most evident in Table 1 are that (a) the metrics for each measure are vastly different, with very different interpretations, and (b) the columns that one might expect to be similar (e.g., percent change in proficiency for the sequential cohorts versus the same cohort) are often quite different. The first growth measure, the change in percent proficient for sequential cohorts (column C), shows about half of the schools improving (i.e., they have positive values) and about half of the schools losing ground (i.e., they have negative values). To some extent, this is a reflection of the demographic differences between the two cohorts involved in this calculation, and to some extent it is a reflection of actual school change that occurred that year. A slightly more positive trend appears in the next column where sequential cohorts are used to calculate changes in school-level scale scores (column D). Still, the presence of negative values may be interpreted as a decrease in average knowledge and skills of 8<sup>th</sup> grade students from one cohort to the next. Even if all the scores were positive, the main problem with the sequential cohort approach is that it is impossible to determine the degree to which this growth

---

<sup>2</sup> The term average is used loosely here, as the estimates are not arithmetic averages but precision weighted empirical Bayes estimates (see Raudenbush & Bryk, 2002).

measure reflects real changes in school performance versus changes in the student population. The large negative values in the scale score indicator suggest that there may be substantial bias due to differences between the sequential cohorts of students. We can begin to get a better idea of this by comparing the results for sequential cohorts to the results based on the same cohort followed from grade 7<sup>th</sup> to 8<sup>th</sup> grade.

The most dramatic difference between the changes in proficiency rates for sequential cohorts (column C) versus the same cohort (column E) is that the changes based on following the same cohort are even more negative. This may be a function of actual declines in performance from 7<sup>th</sup> to 8<sup>th</sup> grades, or it may be a reflection of differences in the proficiency standards from 7<sup>th</sup> to 8<sup>th</sup> grade. In the next column, the change in average scale score (column F) shows no negative values. This aligns more closely with the notion that students scores on a vertically scaled test should increase from one grade to the next. It also confirms that the negative values seen in the scale score differences for sequential cohorts of students are likely due to differences between the cohorts, instead of decreases in individual student performance (i.e., students losing knowledge and skills).

The column showing average individual change (column G) is also entirely positive, but its interpretation is dependent on knowing the amount of change in FCAT scores that is expected to happen between 7<sup>th</sup> and 8<sup>th</sup> grades. By subtracting the four cut scores for 8<sup>th</sup> grade from the corresponding cut scores for 7<sup>th</sup> grade (see <http://www.firn.edu/doe/sas/fcat.htm>) and then averaging, we see that the expected amount of change in scores from 7<sup>th</sup> to 8<sup>th</sup> grade is 131 points. We can also calculate the average individual change from the data in Table 1, which is similar at 133 points. By dividing each value in column G by 131, we can see the average gain score for each school proportionate to a years worth of growth. For example, school 155 had an

average gain of 169 points, which is equal to 127% of a years worth of growth. School 216 had an average gain of 115 points, which is equal to 88% of a years worth of growth.

The scores from the Layered VAA model (column H) vary from  $-6.2$  to  $3.1$ . These values represent each school's relative effectiveness, where a score of 0 can be interpreted to mean that a school is of average effectiveness, scores above zero suggest above average effectiveness, and scores below zero suggest below average effectiveness. By setting a confidence level (e.g., 95%), we can identify a cutoff for declaring VAA scores as being significantly different from the average. This can be helpful in identifying high and/or low performing schools. These scores could also have been represented as adjusted gain scores that would look much like the gain scores in column G; however, the average VAA gain score cannot be interpreted as the expected annual growth because it is adjusted for more than one year of prior achievement. This precludes representing VAA scores as a proportion of a year's worth of growth using the method described above. The VAA scores are also skewed left ( $sk = -1.1$ ), suggesting that the VAA assumption of normally distributed school effects may not be appropriate here. The left skew also suggests that there are more very low performing schools than very high performing schools. There are also some anomalies where schools are quite similar on all metrics except their VAA scores. For example, schools 244 and 211 have virtually identical status measures (columns A and B), same-cohort gain scores (columns E and F), and individual gain scores (column G), but their VAA scores are quite different ( $+2.0$  vs.  $-0.1$ ). These two schools are also of similar size, so the difference in VAA scores can't be attributed to shrinkage. The difference may be due to differences in prior achievement scores from two years back (exploring this possibility is a next step in our analyses), but how can you rationalize giving two schools different VAA scores for this year given that they both started from about the same

point and produced similar gains this year? This complexity in the Layered VAA model is hard to explain, let alone justify.

Table 2 shows the correlations between each status and growth measure. The correlations reveal considerable similarity between some measures (evidenced by significant positive correlations), while other measures present opposing information (evidenced by significant negative correlations), or no relationship between measures. The correlations between the sequential-cohort measures and the same-cohort measures are modest for change in percent proficient ( $r = .55$ ) and stronger for change in the average scale score ( $r = .80$ ). While the modest correlation for percent proficient is not surprising and suggests a significant difference in the information conveyed by these metrics, the correlation for changes in average scale scores is quite high. The differences between these correlations suggests a compounded deleterious effect of comparisons of different cohorts combined with the reduced precision associated with the use of binary proficiency scores (i.e., proficient or not proficient) instead of continuous scale scores. Still, a correlation of .80 is not high enough to support substitution of a sequential-cohorts measure for one using the same cohort of students. Yet under the same-cohort approach, differences in the relative levels of the proficiency cutoffs can lead to negative estimates of growth for all schools (as is the case here because 7<sup>th</sup> grade 2004 proficiency rates were higher than 8<sup>th</sup> grade 2005 rates for all schools).

Another important finding that is evident in Table 2 is the very high correlation between average individual growth and the value-added estimates ( $r = .90$ ). This suggests that despite the arguments by Sanders, Saxton, & Horn (1997) that the TVAAS approach adjusts for student background, TVAAS does little more than produce school average gain scores. This finding corresponds with the finding by Tekwe et al. (2004) of practical redundancy between the

TVAAS estimates and unadjusted gain scores. Why then, would policymakers choose to implement a system that is very expensive and largely uninterpretable if the average individual growth model produces nearly identical results?

To further explore the implications of different models for specific schools we present Table 3, which shows ranks for each school under each measure of status and growth. What is most striking in this table is that most schools with low status measures show greater growth, and schools with high status measures show lesser growth. This may not be surprising from a measurement perspective because of ceiling and floor effects, or because change is often negatively associated with baseline scores, but it provides further support that status and growth should both be considered in an accountability system.

### *Discussion*

Our discussion is organized around the five policy questions presented earlier. The first question asked whether school change scores should be calculated using proficiency levels or scale scores. The answer is partially dependent on policy goals. If changes in the percent proficient are used to indicate change, then more focus is placed on getting students to a minimum standard – not proficient means not proficient, and that just doesn't cut it. Furthermore, when using changes proficiency rates, schools that have few students scoring below proficient have little chance of showing gains and therefore, little incentive to improve. What's worse is that in any school, teachers may have little incentive to push for further improvement from students who are already scoring above proficient.

If scale scores are used to calculate change, then growth will be evident at any point on the achievement continuum. This may help to avoid the problems associated with change in percent proficient, though some loss of interpretability is possible. Even so, representing scale

score changes relative to an expected annual growth rate can maintain interpretability. By combining the change in scale score with a percent proficient status measure, an accountability policy can focus attention on achieving minimum standards while also recognizing improvement towards proficiency or even higher levels of performance.

The second policy question focused on the use of same-cohort versus sequential-cohort comparisons when using school-level scores to estimate growth. Even though it may seem obvious that same-cohort comparisons are closer to representing student learning, this choice is really one of the lesser of two evils. The same-cohort approach helps to reduce the influence of shifts in student demographics, but it can still produce confusing results if proficiency scores are used to calculate change because the definition of proficiency may shift from grade to grade. Therefore, if school-level scores must be used to measure change, then the use of same-cohort changes in scale scores (i.e., if they are vertically scaled) seems to be the best option.

The next policy question is: If change is calculated using the same cohort of students, can changes in school average scale scores be used to reflect learning gains? Table 1 shows remarkable similarity between changes in the school average scale score for the same cohort and the average individual change measure for all but four schools. Interestingly, these four schools are comprised of the two lowest achievement schools and the two highest achievement schools. Given that the main reason for differences between these two metrics is student mobility, these results would not be surprising if these schools had high mobility rates. This would be especially true if students in the two lowest performing schools used school choice under NCLB to move to one of the highest performing schools (exploring this possibility is a next step in our analyses).

The moderate correlations between same-cohort scale score change and individual growth measures ( $r = .66$  with individual growth,  $r = .64$  with VAA) suggest that changes in school average scale scores cannot serve as an equivalent substitute for measures based upon individual growth. However, if the four outlying schools identified in the previous paragraph are removed and these correlations are recalculated, the similarity increases ( $r = .88$  with individual growth,  $r = .80$  with VAA). This suggests that same-cohort changes in school average scale scores are quite similar for most schools. Of course, in a high-stakes accountability system, accuracy must be obtained for all schools.

This leads to the question of whether indicators of individual student growth are worth the extra effort required to link and then analyze annual student-level test data. Our conclusion at this point is yes, methods of individual student growth are preferred given their ability to greatly reduce the influence of shifts in student populations over time, and given that simpler alternative indicators did not provide similar enough results. Recognizing both the complexity and financial implications of maintaining student-level databases, the same-cohort scale scores are still a valuable alternative and a vast improvement over the sequential cohort method currently in use. However, when feasible, individual growth measures are preferable.

The fourth policy question focused on whether student gains should be attributed to more than one school if the student switched schools during the school year. The very high correlation between the individual growth measure and VAA scores ( $r = .90$ ) suggests that this adjustment does not make much of a difference. Still, if resources are available to make such a modification, it can only improve the accuracy of the measures. A planned step in our ongoing analyses is to alter the calculation of average individual growth to include weights that allow student gains to be attributed to schools proportionate to the amount of time spent in the school. This is similar to

the method used to handle mobility in the VAA model, but it maintains the interpretability of average individual growth relative to expected rates of growth. It will be interesting to see how the correlation between average individual growth and VAA scores changes after weighting for student mobility.

Lastly, and possibly the most important policy question addressed by this research is: Should the more complicated Layered VAA model be preferred over simpler models. In short, we find little evidence that the VAA model produces different results from a model of average individual growth. The latter is less complicated to implement, and its results are easier to explain. Furthermore, the model of average individual growth imposes fewer assumptions on the data, and is less likely to under identify small schools as being low or high performing. Of course, the model of average individual growth may produce less precise estimates for small schools, but this can be overcome by averaging effects over multiple years to identify consistent trends in performance. An important area of future research in estimating school performance is that focusing on the precision of estimates for making policy decisions and the amount of data, or number of years of data, required to produce sufficiently reliable estimates.

## References

- Alliance for Fair and Effective Accountability. (2004, October). *Joint organizational statement on 'No Child Left Behind' act*. Retrieved July 29, 2005 from <http://www.nea.org/presscenter/nclbjointstatement.html>
- Kane, T. J. & Staiger, D. O. (2002) Volatility in school test scores: Implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings papers on education policy 2002*. Washington, DC: Brookings Institution.
- McCaffrey, D., Lockwood, J.R., Koretz, D, and Hamilton , L. (2003). *Evaluating value-added models for teacher accountability*. Washington, D.C.: RAND.
- Olsen, L. (2004, November 17). 'Value added' models gain in popularity. *Education Week*, pp. 1, 14-15.
- Olsen, L. (2005, February 2). States revive efforts to coax NCLB changes. *Education Week*, pp. 1, 29.
- Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29(1), 121-129.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.
- Sanders, W. L., Saxton, A. M., Horn, S. P. (1997). The Tennessee Value-Added Assessment System: A quantitative, outcomes-based approach to educational assessment. In J. Millman, (Ed.), *Grading teachers, grading schools. Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin.

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

Table 1.

*Alternative Measures of School Achievement Status and Change for Twenty-Nine Schools.*

School ID	Number of Students	STATUS 2005		CHANGE 2004-2005					
		Percent Proficient 2005	Avg. Score 2005	Sequential Cohorts		Same Cohort		Average Individual Change	Value Added Score
				Change in % Proficient	Change in Avg. Score	Change in % Proficient	Change in Avg. Score		
A	B	C	D	E	F	G	H		
14	28	0	1551	-6	104	-13	131	33	-2.5
6	72	7	1662	4	184	-4	211	73	-3.0
155	281	7	1656	-3	65	-5	170	169	-0.4
212	168	8	1670	0	62	-5	166	160	0.2
146	191	8	1679	1	154	-1	231	205	2.4
168	143	10	1681	3	134	-2	165	175	0.9
92	225	11	1702	0	84	-5	174	143	1.2
102	99	11	1739	5	138	-2	232	205	3.1
69	305	14	1735	-7	-1	-6	174	173	1.5
244	348	16	1743	-3	27	-12	127	154	2.0
211	353	18	1746	0	75	-12	145	159	-0.1
216	528	19	1766	-3	24	-15	101	115	-1.7
238	338	20	1758	-8	7	-16	123	126	-0.6
213	269	20	1770	1	72	-8	153	153	1.2
31	259	22	1776	0	54	-6	130	138	0.4
219	412	22	1784	-1	28	-11	167	162	2.6
207	320	23	1787	5	94	-7	191	153	1.9
66	360	29	1823	2	61	-8	140	156	1.7
25	259	31	1833	-1	5	-10	120	127	0.3
279	443	31	1831	-9	-23	-10	106	127	-0.4
253	587	31	1850	-8	-11	-10	134	118	-1.3
254	252	34	1850	3	75	-4	156	170	2.1
38	141	36	1853	10	93	-6	191	140	0.4
256	489	38	1879	4	51	-7	155	144	1.6
259	562	48	1914	-1	0	-13	88	100	-1.8
63	423	49	1927	4	34	-7	112	114	-0.5
267	329	55	1973	-3	-8	-6	127	118	0.4
145	383	67	2031	-2	-27	-7	68	43	-4.3
152	341	73	2048	5	-2	-6	70	32	-6.2



Table 3.

*School Ranks Under Alternative Measures of School Achievement Status and Change for Twenty-Nine Schools.*

School ID	Number of Students	STATUS 2005		CHANGE 2004-2005					
		Percent Proficient 2005	Avg. Score 2005	Sequential Cohorts		Same Cohort		Average Individual Change	Value Added Score
				Change in % Proficient	Change in Avg. Score	Change in % Proficient	Change in Avg. Score		
A	B	C	D	E	F	G	H		
14	28	29	29	25	5	27	18	28	26
6	72	28	27	5	1	5	3	26	27
155	281	27	28	23	12	6	8	6	19
212	168	26	26	14	13	7	10	8	17
146	191	25	25	12	2	1	2	1	3
168	143	24	24	9	4	2	11	3	12
92	225	23	23	15	8	8	6	15	10
102	99	22	21	2	3	3	1	2	1
69	305	21	22	26	24	12	7	4	9
244	348	20	20	22	19	25	20	11	5
211	353	19	19	13	9	24	15	9	18
216	528	18	17	24	20	28	26	23	24
238	338	17	18	27	21	29	22	20	22
213	269	16	16	11	11	18	14	13	11
31	259	15	15	16	15	13	19	17	14
219	412	14	14	19	18	23	9	7	2
207	320	13	13	3	6	14	4	12	6
66	360	12	12	10	14	19	16	10	7
25	259	11	10	18	22	22	23	19	16
279	443	10	11	29	28	21	25	18	20
253	587	9	8	28	27	20	17	22	23
254	252	8	9	8	10	4	12	5	4
38	141	7	7	1	7	11	5	16	15
256	489	6	6	6	16	15	13	14	8
259	562	5	5	17	23	26	27	25	25
63	423	4	4	7	17	16	24	24	21
267	329	3	3	21	26	10	21	21	13
145	383	2	2	20	29	17	29	27	28
152	341	1	1	4	25	9	28	29	29

## Appendix A

## Mathematical Representations of Models of School Performance

*School-level Current Status.*

1. **Percent Proficient** – the percentage of students scoring at or above the proficiency cut-score.

$Y_{tgs}$  = % of students at time  $t$  in grade  $g$  in school  $s$  who scored at or above proficient

2. **Average Scale Score** – the average or mean scale score

$$Z_{tgs} = \frac{\sum_{i=1}^{n_{tgs}} X_{itgs}}{n_{tgs}}$$

$Z_{tgs}$  = Average scale score at time  $t$  in grade  $g$  in school  $s$

$X_{itgs}$  = Scale score at time  $t$  for individual  $i$  in grade  $g$  in school  $s$

$n_{tgs}$  = Number of students at time  $t$  in grade  $g$  in school  $s$

*School-level Change Models.*

3. **Change in Percent Proficient for successive cohorts of eighth-graders** – calculated as the difference in the percentage of 8th graders scoring proficient in spring 2004 subtracted from the percentage of 8th graders scoring proficient in spring 2005.

$$Y_{tgs} - Y_{(t-1)gs}$$

4. **Change in Average Scale Scores for successive cohorts of eighth-graders** - calculated as the difference in the average scale score for 8th graders for spring 2004 subtracted from the average scale score for 8th graders for spring 2005.

$$Z_{tgs} - Z_{(t-1)gs}$$

5. **Change in Percent Proficient for the same cohort of students** – calculated as the difference in the percentage of 7th graders scoring proficient in spring 2004 subtracted from the percentage of 8th graders scoring proficient in spring 2005.

$$Y_{tgs} - Y_{(t-1)(g-1)s}$$

6. **Change in Average Scale Scores for the same cohort of students** – calculated as the difference in the average scale score for 7th graders for spring 2004 subtracted from the average scale score for 8th graders for spring 2005.

$$Z_{tgs} - Z_{(t-1)(g-1)s}$$

*Student-level Change Models.*

7. **Average Individual Change** – After linking individual student scores from spring 2004 and spring 2005, each student’s scale score from spring 2004 is subtracted from his/her scale score in spring 2005 to produce an individual gain score from 2004 to 2005. School performance gains are calculated as the average of the individual gain scores for all students enrolled in the 8<sup>th</sup> grade in spring 2005.

$$\frac{\sum_{i=1}^{n_{tgs}} (X_{tig_t s_t} - X_{(t-1)ig_{t-1}s_{t-1}})}{n_{tgs}}$$

$X_{ig_t, s_t}$  = Scale score at time  $t$  for individual  $i$  who was in grade  $g$  at time  $t$  and in school  $s$  at time  $t$

$X_{(t-1)ig_{t-1}, s_{t-1}}$  = Scale score at time  $t - 1$  for individual  $i$  who was in grade  $g$  at time  $t - 1$  in school  $s$  at time  $t - 1$

8. **Layered Value-Added Model** – After linking individual student scores for three years from spring 2003 to 2005, a statistical model designed by Sanders, Saxton, and Horn (1997) is used to estimate each school’s contribution to its students test score gains from 2004 to 2005 after controlling for previous test scores from 2003. These estimates are called “value-added” scores, and are identical to the like-named scores used in the Tennessee Value Added Assessment System (TVAAS). These scores can be interpreted as relative to the average growth for all schools in the district.

$$x_{tis_t} = \mu_t + \sum_{l=1}^t \sum_{s_{t-1}=1}^{29} P_{is, s_{t-1}l} u_{s_{t-1}l} + \epsilon_{tis_t}$$

$x_{tis_t}$  = the test score at time  $t$  for the  $i^{th}$  student who was in school  $s$  at time  $t$

$\mu_t$  = the population mean parameter at time  $t$

$P_{is, s_{t-1}l}$  = the proportion of academic year time spent by the  $i^{th}$  student, who was in the  $s^{th}$  school at time  $t$  test, in the  $s_{t-1}$  school during the year prior to the test at time  $l, l = 1, t$

$u_{s_{t-1}l}$  = the random effect of the  $s_{t-1}$  school on test scores at time  $l$

$\epsilon_{tis_t}$  = random within school error for the  $i^{th}$  student in the  $s^{th}$  school at time  $t$