

Designing Gross Productivity Indicators: A Proposal for Connecting Accountability Goals, Data, & Analysis

Yeow Meng Thum
College of Education,
Michigan State University,
East Lansing, MI

DRAFT: November 30, 2005

Abstract

As more and more educational agencies consider value-added analysis to anchor their accountability systems, recent research caution that, with important information about schools and programs still unavailable, value-added analysis may not be able to deliver the more ambitious policy goals of isolating effects due to teaching or school practices. Nevertheless, while educational agencies strive to improve information about teaching and school practices, I argue that state accountability systems could focus on the more modest but no less important questions concerning various conceptualizations of *gross productivity*, the measures of which are to provide the essential outcomes in subsequent explorations for evidence of teacher or school effectiveness. While attributions to teachers, programs, or schools may be limited, gross academic productivity indicators by design will at the very least allow monitoring agencies to determine if students currently attending a school, a grade-level, or a significant sub-group made progress, how much was learned on average, and if the improvement met specified standards. In this paper, I draw on recent applications of a school-specific Bayesian multivariate multi-cohort mixed-effects growth model and show how such measures may be readily culled from the longitudinal student data-block for each school. I consider several value-added hypotheses concerning the progress of a school framed in terms of two complementary indicators (1) cohort-to-cohort improvement and (2) grade-level improvement over time. Finally, we employ a Bayesian multivariate meta-analysis to arrive each school's productivity profile, construct reliability estimates for the indicators, as well as provide rankings and other comparisons of the schools.

Keywords: accountability, Bayesian meta-analysis, gross productivity, measuring progress, multivariate multi-cohort mixed-effects models.

Introduction

The debate on educational reform during the past four decades had given an increasing emphasis to monitoring public schools for improvement and accountability.¹ This trend culminated in the No Child Left Behind (NCLB) Act of 2001 (US, 2001).² Under the new law, each state must put into place an accountability process by the 2002-2003 school year to start gauging the progress its schools are making annually so as to ensure that *all* students are proficient by 2014.

By tying Title 1 funding to compliance, NCLB holds new and sweeping policy implications for states and school districts.³ States need to ensure that their assessments align with their curricular standards. States also need to produce a basis for defining and measuring *adequate yearly progress* (AYP) with respect to their individual standards.⁴ Thus, seemingly overnight, NCLB pushed the research on accountability analyses to the forefront of the discussion on standards-based school reform. Considering the state of unpreparedness many state accountability regimes were in prior to the passage of NCLB (Stevens, Estrada, & Parkes, 2000), it is fair to say that the need to have an accountability system rationalized and in operation almost immediately had caught many states off-guard.

As the states look to flesh out their existing accountability systems, increasing attention is given to the on-going research employing *value-added* analyses. Bryk and Weisberg (1976) had earlier introduced the notion of value-added by a treatment (or simply a treatment effect) in a program evaluation setting with quasi-experimental data, showing how it is an improvement for separating treatment effects from natural maturation (age being the primary covariate) over the more commonly applied “adjustment” strategies for a pre-post test design.⁵ It is important to note that the authors recognized their proposed analysis to be merely a stop-gap measure for separating treatment effects from maturation when only two time points are available. A more informative analysis would apply growth modeling to longitudinal data, in which the individual student’s past performance could serve as his immediate baseline when evaluating his current progress.⁶

The common purpose of value-added analysis in the accountability context is to order, or rank,

¹Willms (1992) traced the policies and their constituencies in the debate on school monitoring, and discussed approaches to reporting school performance.

²For a brief history of federal efforts in improving public education from the Elementary and Secondary Education Act of 1965 (ESEA) to the No Child Left Behind Act of 2001, consult National Conference of State Legislatures (2004).

³See, for example, National Coalition for Parent Involvement in Education (2004) for a brief overview.

⁴Two timely and informative reports on the progress in the states are Erpenbach, Forte-Fast, and Potts (2003) and Marion et al. (2002).

⁵In retrospect, this work, which expands on an earlier presentation of the same idea in Bryk and Weisberg (1974, August), proposed a definition of a causal “effect” in non-randomized studies that was also considered by Rubin (1978) and by Holland (1986).

⁶This assessment is now widely accepted. Essentially, growth modeling of student longitudinal observations improves precision by blocking on the individual subject in a repeated measurement designs – the idea behind the phrase “the subject serves as his own control.”

This is an update of a paper prepared earlier for the Presidential Invited Session *Empirical Investigations of Value-Added Modeling for Accountability* at the annual meetings of the American Education Research Association, Montreal, April 2005. The research reported here drew from an earlier study commissioned by the New American Schools, and from lessons learned in evaluations for the Los Angeles Unified School District and for the Milken Family Foundation of Santa Monica. Additional funding was received from the National Center for Research, Evaluation, and Student Testing (CRESST). Opinions expressed in this paper are however the sole responsibility of this author, as are all remaining errors. Please direct all correspondence to Y. M. Thum, at thum@msu.edu

schools in terms of their performance after taking into account meaningful and explicit performance baselines in the comparisons. Applied to school accountability data, schools are compared on their performance on large-scale testing outcomes, most often resulting in rankings, after adjusting for differences in student academic intake characteristics and socio-economic status. A value-added estimate is defined here as the difference between the actual performance of schools with the expected performance for certain groups of schools forming a comparison baseline.⁷ If successfully implemented, a value-added approach would produce arguably “fairer” rankings because the results would reflect better the overall learning that had occurred in a school, or for a teacher, free from the influence of their students’ prior academic attainment. The same cannot be said, however, of the simpler and more prevalent procedure that compares or monitors schools, or teachers, in terms of their average proficiency scores from year to year (Meyer, 1996; Raudenbush, 2004b).

Many studies implementing essentially the equivalent idea of an “adjusted comparison” exist today. Value-added analyses form a class of complex statistical procedures for digesting assessment data to serve accountability purposes, pioneered some 10 years before by Dr. Sanders and his colleagues at the Tennessee Value-Added Assessment System (TVAAS) (Sanders & Horn, 1994). There is greater use of longitudinal student data (*e.g.*, Sanders & Horn, 1994; Webster & Mendro, 1997; Gray, Jesson, Goldstein, Hedger, & Rabash, 1995; Meyer, 1996; Bryk, Thum, Easton, & Luppescu, 1998; Thum, 2003; Bryk, Raudenbush, & Ponsiciak, 2003; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Stevens & Moreno, 2004; Tekwe et al., 2004; Ballou, Sanders, & Wright, 2004), but the value-added principle is also applied to cross-sectional data (*e.g.*, Goldstein & Spiegelhalter, 1996; Harker & Nash, 1996; Tymms, 1999). Generally, some form of the linear mixed-effects model is employed. A small number of researchers have argued however for simpler multiple regression models to maintain a level of transparency that will retain public engagement (*e.g.*, Fitz-Gibbon & Tymms, 2002). Until recently, most applications considered criterion variables one at a time although Thum (2002a, 2003) also deployed mixed-effects models for multiple outcomes as well as their weighted composite in an index score problem. The predominant inference framework for models employed in value-added analyses is maximum likelihood but Bayesian approaches also existed. Thum (2002a, 2003) considered Markov chain Monte Carlo (MCMC) computations of a fully Bayesian model as well in order to facilitate inference on comparisons, rankings, and progress indicators that involved ratios of estimates. McCaffrey et al. (2004), in a significant attempt to reframe several earlier models, similarly favored a Bayesian inference framework.

The upsurge in interest among the policy and education research community in the value-added concept was duly noted by Lynn Olson who, writing in the November 17, 2004 issue of *Ed Week*, proclaimed in a headline that “*Value Added*” Models Gain in Popularity.⁸ Conferences on any number of issues raised by accountability under NCLB fanned across the country.⁹ The *Journal of Educational and Behavioral Statistics* devoted a special issue to the major technical and interpretive challenges of value-added accountability modeling (Wainer, 2004). As a final testament to this feeding frenzy, a casual count of the 2005 *American Educational Research Association* annual meeting program for Montreal totals 58 separate sessions that are devoted to some aspect

⁷Frequently called “benchmarks,” the appropriateness of baselines are of course subject to challenge. A benchmark may be relative, such as one derived from “similar schools,” or it may express an absolute standard, such as “100% proficient.” But no matter how it is derived, benchmarks carries with it a certain amount of arbitrariness.

⁸See http://www.oft-aft.org/In.the.news/News.articles/EdWk_ValueAdded_11.17.04.htm.

⁹An early conference is the *Student Achievement and School Accountability Conference* hosted by the US Department of Education in October 2002. To take just one example indicative of the interest on value-added procedures, there was the University of Maryland Conference on *Value-added modeling: Issues with theory and application* of October 21, 2004.

of accountability,¹⁰ up from 17 for the previous year.

But just when the value-added concept appears to have gained acceptance, its methodological properties and policy merits have come under increased scrutiny. Questions about its modeling assumptions, how does one assess the validity and reliability of its results, or how should we deal with the ethical issues when we use school rankings based on standardized assessments in high stakes decisions, though serious, are of course not new (*e.g.*, Lacey & Lawton, 1981; Goldstein & Myers, 1996; Thomas & Goldstein, 1995; Broadfoot, 1996; Rowe, 2000; Linn, 2000). Researchers are also concerned that questionable psychometric quality of standardized test scores and their scales, the principal outcome variables in many value-added analyses, may invalidate accountability decisions (Reckase, 2004). Although clearly critical to all attempts at measuring growth and change, Thum (2002b, 2003) suggested however that these issues concerning tests, what they measure and how well they are doing it, are more likely to be quality issues for psychometric research¹¹ and for psychometric practice, and less an issue with theories for measuring growth.

But it is the recent reminder regarding the validity of causal claims based on extant accountability data that has proven to be sobering for analysts and policy makers alike. Drawing upon the distinction between Type A (a whole-school effect) and Type B (effect of identifiable teacher or school practices) effects that Raudenbush and Willms (1995) had made earlier in the contexts of cross-sectional observational study, Raudenbush (2004a, 2004b) pointed to the lack of systematic information about teacher or school practices, and this critical information gap makes the determination of causal effects even more illusive. In short, results described by phrases such as “school effects” or “teacher effects” do not automatically merit causal interpretation (Rubin, Stuart, & Zanutto, 2004). In short, we must simply recognize that there are limits to what we can learn if we do not have the “right” data.

Although I am in full agreement with the conclusions drawn by Raudenbush (2004a, 2004b), I nevertheless argue that, for some of the available databases, we may still make sensible statements about the progress of classrooms, grades, and schools. This view, which I believe to be closer to the stated goals of many accountability agencies, represents a school monitoring perspective as opposed to one trained on the search for teacher or school effects. While, gross productivity analysis does not aim to pinpoint plausible sources of strengths or weakness of teachers, departments, or schools as *accountability units*, it seeks to provide measures of teachers, departments, or schools considered as *accounting units*. The purpose we have set for our accountability exercise is therefore admittedly limited; namely, in the spirit of what is most demanded of current accountability agencies, we will concern ourselves with estimating *gross productivity*. The initial focus of a value-added analysis in my view, therefore, should be to build descriptive measures of productivity and improvement, a conclusion also shared by Rubin et al. (2004).

In this paper, I outline an analysis for monitoring how a school is advancing on one, or on several, clearly stated criteria. Inference procedures, based on earlier proposals by Thum (2002a, 2002b), are detailed for a multitude of value-added hypotheses about a school’s progress in student learning in terms of two complementary school performance indicators. The first criterion relates to the improvement in productivity of a school in terms of the change in the growth in learning of succeeding student cohorts. After tracing the performance of each cohort within a school, I present evidence by way of a Bayesian multivariate meta-analysis of school-specific fixed-effect estimates and variance-components to suggest that, for our group of schools, productivity is improving on a whole. In the second, which I had termed *AYP-NCLB* earlier but is generally applicable for any

¹⁰These sessions involved some 155 individual papers declaring “accountability” in their titles!

¹¹Good psychometric research is certainly a vital partner to defensible accountability measurements.

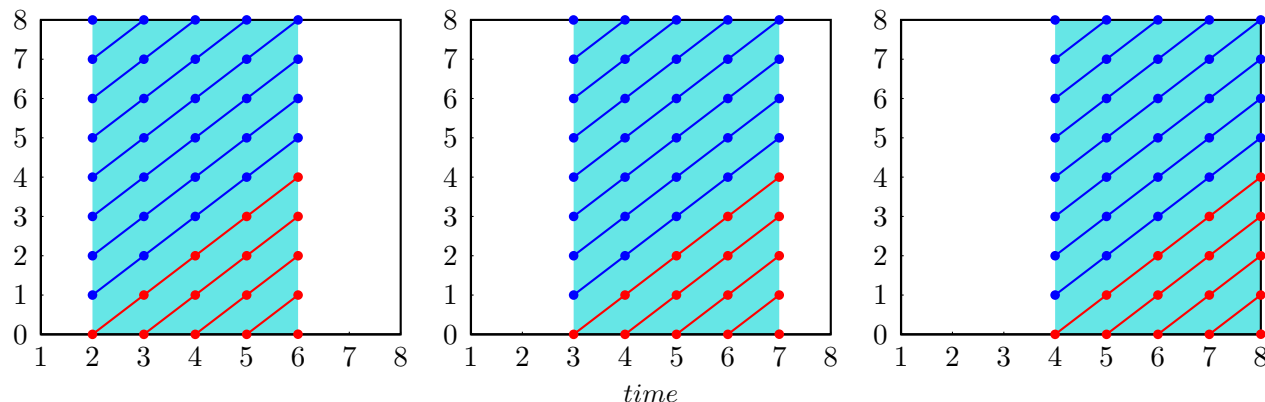


Figure 1. A five-year accountability data-block for kindergarten (0) through grade 8 (Thum, 2002b).

performance criterion anchored on a standard, I provide an estimate of how likely it is that given grade-level is producing learning at a rate that exceeds a minimum growth rate if students in that grade will, on average, be proficient (or better) in 2014. There is evidence, by way of a similar analysis, that the majority, almost 80% of the 96 elementary schools in an urban district, showed progress on this indicator with each school making AYP-NCLB at 95% chance or better.

Note: A roadmap to the paper is to be added here.

Gross Productivity Measurement

As in any attempt to measure a specific attribute of an object, and then to compare objects on the attribute, a substantial effort is required to *standardize* the instruments. Furthermore, attention must be given to standardize the conditions and procedures for obtaining the observations in order to secure relatively valid comparisons. There are many factors that impact the reliability and validity of measurements. Thum (2002b, 2003) discussed several features of the data to be used for accountability that would promote validity, reliability, and stability. For this illustration, we will highlight only the most critical of these.

Accountability Data-Block

The accountability data block, which delimits which students and for what time-frame the evaluation applies, must be explicated and be consistently employed over time (Thum, 2002b). Figure 1 shows an example of a five-year block for kindergarten to grade 8. This will add a “constant ballast” that will increase year-to-year stability to the evolving evidence base. For our illustration below, the data-block included the math and reading scores from all students attending grades 1 through 5 from 1998 through 2002. Going forward, we might move the block’s time limit ahead a year at a time, so that in 2003 we would use much of the same test information – grades 1 through 5 for 1999 through 2002 – and add grades 1 through 5 for 2003.

Outcomes

Outcomes for change measurement are ideally measured on an interval scale. While this scale property serves as an ideal, it may or may not be 100% achieved, and continual psychometric effort must be expended in order to assure the reasonableness of this assumption over time. Of course, we may discuss growth and change in terms of weaker measurement scales. However, there is a

limit to how much we can learn (and therefore say) about what is being learned in our classrooms and at what pace with, for example, ordinal scales because they convey much less information. And because ordinal scales contain less information, their use will be accompanied by greater uncertainty. From this perspective, therefore, analyses which begin with performance categories (e.g., advanced, proficient, etc.) are not a solution to accountability measurement in the absence of a usable interval scale.

Furthermore, if we are to detect change over grade-levels, score scales for each test should be *reasonably well-equated across grades*. Success in any scale equating effort is a matter of degree and thus a continual psychometric quality assurance effort is necessary to help maintain the comparability of our scales across years.

If performance standards on a test have been determined, e.g. through a standard setting study, the score scale will also support reporting in terms of performance categories. All analyses in this report are based on scale scores, which serves as the *analysis metric*. Reporting growth and change in terms of their corresponding percentile ranks and, more pertinent to NCLB reporting and inferences, in terms of performance categories (*reporting metric*), can be accomplished internally and in parallel. The percentage of students at and above the proficient performance standard (a statistic sometimes referred to as a “PAC,” or *percent at-or-above cut-point*), discards useful information when they refer to regions on an already well-mapped ability continuum. Of even greater concern is that the PAC is also insensitive to growth amounts that may be instructionally meaningful but do not push the student over the PAC threshold. Thus, calls to work directly with performance categories risk throwing out useful information in many systems that have at its base usable vertically-equated score scales.

Multiple measures convey a broader view of student learning. They are not an analytic obstacle, as we shall show below, as long as each outcome are vertically equated on a interval scale. Our procedure track growth simultaneously in multiple learning domains and construct appropriate inferences for single, or multi-subject, accountability hypotheses.

It is important to carefully weigh all the input information we have. Accountability analyses often begin with the student test score without indicating the well-known fact that not all scores are *equally precise* estimates of the student’s performance status. Not using this information about the relative imprecision of test scores means that we will not be able to separate out measurement errors from sampling variation. It also presumes that our estimates contain only sampling variation and no measurement error. We introduce the standard errors of measurement (sem) for each score into our models. This particular measure of score imprecision is the so-called conditional standard error of measurement, a by-product of an item response theory (IRT) scaling procedure in which student responses to individual test items are combined into measures. Other species of sem’s are possible, such as those inflated by the misfit of the individual’s responses to the scaling model (but these are not available for this analysis).

Analysis

Before we turn to the data, the analyses for individual schools, and the accountability results in the following sections, we note that, in general, employ models to construct a basis for making claims about regularity in the face of uncertain accountability information. What has “data massaging” to do with detecting progress in student learning? Skeptics of the value of statistical modeling in accountability should realize that whenever we make a summary statement about performance or about progress at grade-level, for example, we employ a model to split the observed information into what is signal and what is noise. As Thum (2002b) warned,

“Whether stated explicitly or not, it is important to recognize that every analysis involves a model and we should scrutinize with care any analysis that seemed to be free of one.”

A model “smoothes” the data for each school, thus enabling us to construct in a verifiable manner answers to well-framed questions regarding trends in performance. To help judge the usefulness of a model, we make explicit at every turn the model, and the modeling assumptions supporting it, that forms the analytic basis of even the most simplistic portrayal of student and school progress.

Data

Altogether, we processed a total of 217,106 annual student-matched SAT-9 reading and mathematics scale scores from 55,818 first through fifth graders who attended 96 schools in the district from 1998 through 2002. Each school retained in this analysis had students in every grade (1-5) and every year (1998-2002). Some schools may have no data for one or two grade-year combinations. We have argued earlier that a relatively uniform evidence-base is one aspect of standardization which leads to fairer comparisons. The evidence-base presented by a school with absent grade levels, or years, or both, will be relatively weaker than a school with a nearly filled data-block, resulting in greater difficulties in their comparison. Thus, for the purposes of this illustration, we have excluded schools which served fewer than the five grade levels, or schools which opened sometime between 1998 and 2002, or schools which exhibited both of these conditions. A small number of schools enrolled as low as 50 first through fifth graders a year while others enrolled as many as 200 a year. Most schools however enrolled totals from 400 to about 850 grades 1 through 5 students over the period from 1998 through 2002.

Variables

This accountability analysis employed student Stanford Achievement Test version 9 (SAT-9) mathematics and reading scale scores as performance outcomes. We account for student grade-level, the year tested, and the retention status of the score for each year and grade. We also identified the cohort to which a student belonged.

Also, in keeping with the goal of this illustration, we have not employed student demographic characteristics, or common program indicators, such as those indicating bilingual home environment or limited proficiency in English. We have also not employed any information regarding the mobility rates, or school mobility change patterns. Another important factor relates to the preconditioning effects of prior classroom environment, principally identified with the contribution of individual teachers. All these factors may be easily accommodated with a more exhaustive preparation of the available database. A fuller accountability system will prepare and compare estimates and inferences from a series of progressively more elaborated models, with each subsequent model accommodating an expanded subset of these factors. We will provide in the future some procedures based on a generalized information criterion (see, for example, Wright and Wolfinger (1996)) to support model choices where relevant (Thum, 2002b). However, we note that the value of these analyses would not easily make up for the relatively weaker basis for attributing causal agency to one set of factors or another because they are largely based on quasi-experimental data.

Modeling

One compelling reason for considering the student as the natural unit of accountability analysis is that growth in learning occurs in the individual student. And because students move through

Table 1: Definition of Student Cohorts.

Grade	Year				
	1998	1999	2000	2001	2002
5	9	8	7	6	5
4	8	7	6	5	4
3	7	6	5	4	3
2	6	5	4	3	2
1	5	4	3	2	1

Cell entries are arbitrary cohort labels.

Table 2: Grade-Year-Cohort Data

Cohort	Year				
	1998	1999	2000	2001	2002
1					1
2				1	2
3			1	2	3
4		1	2	3	4
5	1	2	3	4	5
6	2	3	4	5	
7	3	4	5		
8	4	5			
9	5				

Cell entries are grade.

the school in cohorts, sharing a similar curriculum and perhaps even peer groups and teachers, students within a cohort tended to a greater degree than students between cohorts to share a common school experience. It is therefore important to separate out the student cohort effect as we trace individual student growth in our analysis.

As a reminder, each cohort consists of students who, no matter when their first scores show up in the accountability data-block, started first grade in the same year. Cohort 5 students ought to have started grade 1 in 1998, Cohort 3 in 2000, etc. Arbitrary cohort labels for our accountability data-block are given in Table 1. Table 2 is the conventional format for displaying cohort data, keyed on the two irreducible factors: cohort and year.

We also agree that individual teachers may have an impact on how and how much a student learns and thus a model that explicitly separates out teacher variability will be appropriate. While we have not identified teachers specifically in this analysis, we have, by following students within their cohorts, blocked student scores by the sequence of teachers and classrooms that students within their cohort shared. A further development of our approach will include a host of relevant time-varying or time-invariant teacher and program covariates introduced either as fixed or as random effects to account for their impact on student performance.

Modeling Student Learning Growth

Single Domain. For each student, subscripted $i = 1, 2, \dots, I_c$, in cohort $c = 1, 2, \dots, 9$, we pose the following linear model

$$y_{tic} = \pi_{0ic} + \pi_{1ic} \cdot \text{TIME}_{tic} + \pi_{2ic} \cdot \text{RETAIN}_{tic} + \epsilon_{tic} \quad (1)$$

for her scores, y_{tic} . This model suggests that the students performance is a linear function of the time (denoted by TIME_{tic} and subscript by $t = 1, 2, \dots, T_{ic}$). Note that the values of TIME_{tic} are specific to the student and her cohort. In fact, TIME_{tic} is simply the year (YEAR_{tic}) of the test score but defined differently for each cohort of students. We centered time so that the intercept for cohort c , π_{0ic} , estimated a cohort-dependent predicted final status of the student. For example, for students who belonged to cohort $c = 5$ (they entered grade 1 in 1998, see Table 1) we centered YEAR_{ti5} at 2002, or

$$\text{TIME}_{ti5} = (\text{YEAR}_{ti5} - 2002) ,$$

and thus π_{0i5} estimated the predicted status of this cohort 5 student when she was in grade 5 in 2002. Our choice of centering for the remaining cohorts had fixed their intercepts at the time each cohort left our data-block. For example, because we computed $\text{TIME}_{ti4} = (\text{YEAR}_{ti4} - 2002)$ for cohort $c = 4$ students, π_{0i4} estimated the predicted fourth grade status of cohort $c = 4$ student i in 2002. Accordingly, we defined for a student belonging to cohort $c = 6$, $\text{TIME}_{ti6} = (\text{YEAR}_{ti6} - 2001)$, so that π_{0i6} estimated her fifth grade score in 2001. Note that the interpretation of our student growth rates, π_{1ic} , were not affected by our choice of centering for the different cohorts.

A student's retention status varies over time. In this model, we defined a time-varying covariate in RETAIN_{tic} to estimate in π_{2ic} which is an adjustment (downwards on average) to the student's growth rate. See Raudenbush and Bryk (2001, p. 179) for a discussion of the rate change coding employed here. If the model fits the data, equation 1 would have helped us remove the effects of retention from our growth factors, π_{0ic} and π_{1ic} , by separating out the impact of retention on student growth. All subsequent accountability analyses are based on these retention-adjusted cohort growth profiles.

If our model for student i of cohort c fits his data, we expect the residual term, ϵ_{tic} , to be distributed as a normal random variable with mean 0 and variance σ^2 . Homogeneity of residual variance among test scores is unlikely given what we know about tests and testing. Under general testing conditions, the student score is only one estimate of his true ability and thus it determined with error. More likely therefore, the residual for each score has variance that is specific to each score, σ_{tic}^2 , as opposed to σ^2 .

Introducing the conditional standard error of measurement, \hat{s}_{tic} , for each of the observed scores into our analyses allows us to weigh observations according to their varying levels of precision. This procedure explicitly recognizes how in reality each single observation carries with it measurement as well as sampling components of error. We have argued that a more realistic representation of these sources of variation will help to produce better calibrated inferences, for example, Thum (2002b). We accomplish this, following Bryk et al. (1998), by weighting the entire equation 1 by the inverse of the score-specific sem estimate, resulting in the following:

1. Each outcome score, as well as the intercept term, TIME_{tic} , and RETAIN_{tic} are each rescaled by $1/\hat{s}_{tic}$.
2. If we assume that each residual ϵ_{tic} has variance σ_{tic}^2 and that \hat{s}_{tic}^2 is an adequate estimate of it, rescaling ϵ_{tic} by $1/\hat{s}_{tic}$ suggests that we fix σ_{tic}^2 at 1.0 for all observations.

Note that our analyses employed all usable test scores, including those who have only one set of test scores in the school. This is of course not in accord with the prevailing view for value-added estimation of student learning; as it is often suggested that if a school does not have an opportunity to affect a student for a full year (i.e., the student should have scores in the same school in two successive years), the school should not be held accountable for the performance of the student. This is very clearly an important empirical issues in any one study and for any one school. We do not disagree with this basic position, and, with proper tracking of student mobility, our approach represents a model-based, as opposed to a design-based, approach to the problem. A fuller analysis must explore the impact of a variety of mobility-related issues on the inferences on the progress of any single school. Lastly, in our current analysis, we are interested in learning growth at the grade level, and thus every single available test score contributes to a better estimate of grade-year score means.

Multiple Domains. In our current model, we estimated a simple linear growth trend for each cohort for both subject matter domains – reading and mathematics simultaneously. Recall that

we have employed the conditional standard errors of measurement of each input scale score so as to appropriately separate measurement errors (the uncertainty of test scores) from sampling errors towards the determination of trends. Furthermore, we have adjusted our growth estimates for those scores that appeared to have been received during retention. Accordingly, we may now write our multivariate student growth model as

$$y_{tic}^m = \pi_{0ic}^m + \pi_{1ic}^m \cdot \text{TIME}_{tic}^m + \pi_{2ic}^m \cdot \text{RETAIN}_{tic}^m + \epsilon_{tic}^m, \quad (2a)$$

$$y_{tic}^r = \pi_{0ic}^r + \pi_{1ic}^r \cdot \text{TIME}_{tic}^r + \pi_{2ic}^r \cdot \text{RETAIN}_{tic}^r + \epsilon_{tic}^r, \quad (2b)$$

where the superscript (and subscript where appropriate) m denotes the model for math scores and r denotes the model for reading scores. Tracing growth in multiple learning domains (math and reading) simultaneously would require that we consider a more general residual error model. For example, we may let σ_m^2 denote the residual variance for math, σ_r^2 be the residual variance for reading, and σ_{mr} be the covariance between the residuals. Weighting equation 2a and equation 2b by the corresponding sem, \hat{s}_{tic}^m for math and \hat{s}_{tic}^r for reading, fixed the residual variance terms to 1.0, but the covariance between the residual term remained unknown. Because its scale is now fixed to be between -1.0 and +1.0; i.e., we see that it is now an unknown error correlation to be estimated. For simplicity, we will assume that residual errors are correlated to a similar degree, ρ , across individual students and time within the school.

Between-Student Variation in Learning Growth

If students start with different competencies and develops individually, we would expect a significant degree of variation among the individual student growth factors ($\pi_{0ic}, \pi_{1ic}, \pi_{2ic}$). Suppressing the superscript denoting subject matter for the moment, we represent this view by the following set of three student-level equations:

$$\pi_{0ic} = \gamma_{00c} + r_{0ic}, \quad (3a)$$

$$\pi_{1ic} = \gamma_{10c} + r_{1ic}, \quad (3b)$$

$$\pi_{2ic} = \gamma_{20c}. \quad (3c)$$

Equation 3a states that the student status parameters vary around a cohort mean, γ_{00c} , with individual residual r_{0ic} . Similarly, equation 3b states that individual growth rates also vary around a cohort mean growth rate, γ_{10c} , with residual r_{1ic} . The adjustment to a student's growth rate, π_{2ic} in equation 3c, depended on the cohort to which the student belonged and did not vary among students in a cohort. This last assumption is not unreasonable, partly because there were insufficient numbers of retentions in cohort for a school to determine individually varying slope adjustments. On the other hand, an adjustment unique to each cohort appeared useful.

Lastly, we assume that the parameter residuals are distributed multivariate normal, with zero means for both the growth factors and the following variance-covariance terms: τ_{00} for individual final status (π_{0ic}), τ_{11} for growth rates (π_{1ic}), and τ_{01} for the covariance between final status and growth rate (π_{0ic}, π_{1ic}).

Additional Modeling Considerations. In this analysis, we estimated a separate parameter variance-covariance model for math and reading. As the reader will notice however, we have averaged the between-student variation in final status and growth rates over all cohorts. This seemed reasonable for a start, but the procedure will clearly support further relaxing this assumption when the data permits in the direction of cohort-specific covariance structure (as suggested by preliminary analyses).

Individual School Results

We fit a multivariate multi-cohort mixed-effects growth model (equations 2a, 2b, 3a – 3c) to the cohort data for each school. Due to insufficient data for cohorts $c = 8$ and $c = 9$, no slopes for these cohorts are estimated. Furthermore, we do not have data for cohort $c = 1$, and the model terms for cohort $c = 1$ are excluded. The properties of this solution is well-known, and we will refer the interested reader to its thorough treatment by Littell, Milliken, Stroup, and Wolfinger (1996). We documented the results for each school in a manner that facilitated their ready retrieval, with the belief that the model fit in *every* school needed a careful examination. values.

The validity of the residual assumptions, for both error residuals and for the random growth factors, for each school are evaluated using simple histogram and quantile-quantile (QQ) plots. Evidence in support of the multivariate normality assumption for the growth factors are gleaned by comparing the Mahalanobis distance for the four-component student residuals against a χ^2 deviate on 4 degrees of freedom in a QQ-plot.

Residual Error Correlation. Not unexpectedly, there were some degree of unevenness across schools in how well the residual assumptions hold. Normality of error residuals for both math and reading tend to hold, in that their estimated mean and variance are close to 0.0 and 1.0 respectively. We find a range in the estimated residual error correlations across schools. Correlations between math and reading residuals were positive, and hovered about 0.3. We also provided a test of univariate normality in the Cramer-von Mises statistic. There was generally little or no wild departures from normality for the error residuals.

Between-Student Variance Components. Checks revealed that student final status tended to normality but that student growth rates, though symmetric, tended to show longer (or heavier) tails. Comparing the Mahalanobis distances for the four student random effects within each school to the theoretical χ^2 deviate on 4 degrees of freedom indicated, quite generally across schools, marked departure from multivariate normality. We however note that this test will be conservative due to the small within-student sample of observations (most often far fewer than the maximum of 10) so that the actual degrees of freedom for the theoretical χ^2 deviate might be fewer the 4. In any case, while we do not expect our fixed-effects estimates for the school to be biased, their confidence interval estimates may tend to be too narrow.

The range was considerable for the 96 schools in terms of the variances of final status. We expect that a significant part of this variability might be a function of our choice of centering for each cohort. For example, while the final status for Cohorts 5 through 9 should hover about the grade 5 average test score, the final status for Cohort 2 estimated a much lower figure in the grade 2 test score average in 2002. Similar, final status for Cohort 3 estimates grade 3 average for 2002, and final status for Cohort 4 estimates grade 4 average for 2002. A new model may specify a common variance components model for Cohorts 5 through 9, but a different variances for Cohorts 2 through 4. Although we do not expect that the spread of the variance components for the cohort learning growth rates to be impacted by these proposed revisions to the covariance structure of the random effects within a school, we do expect some impact on the covariances between final status estimates and growth rates. However, we find only relatively small differences between the fixed-effects under these different covariance assumptions for a number of schools. Nevertheless, we strongly recommend further exploration of the fit of each within-school multilevel model to its data in lieu of using its results in an accountability analysis.

Bayesian Multivariate Meta-Analysis

If we think that comparing the growth factors, or some function of them, of each school to a target or standard is reasonable, we are implying that schools may also be compared to each other. It goes without saying that such comparisons must be done thoughtfully. Schools of course have different amount of information to begin with. Furthermore, it is certainly too much to expect that the same model would fit equally well the data from every school, so that when we examine the outputs from our separate school analyses, we find school estimates vary not only in terms of their values but also in their levels of precision. As a consequence, any direct comparison of school results will be fraught with methodological problems.

The conventional approach to this problem would be to model the data for the entire system of 96 schools by adding a third between-school model on the assumption that the separate two-level school models are *exchangeable*. This approach would be suitable when, for example, the within-school covariance structures are relatively simple, or if they do not vary between schools. But, as it is clear from an examination of the estimated school-specific between school variance-covariance components above, the within-school variance components for our set of schools are not likely to be homogeneous. Furthermore, we had also suggested that future analyses should seriously entertain the possibility that the covariance structure within a school differs between cohorts.

When the precision of the school-level results are sufficiently well-known, meta-analysis provides a vehicle for summarizing the results from individual schools (see Raudenbush and Bryk (1987)). Such an approach enables the analyst to estimate an aggregate result for the groups of school, after weighting the result from each individual school by its precision separately from its sampling variability. Another advantage of a meta-analysis is that it also improves on the individual school estimates by using the information from all the schools in the study. Furthermore, school-level covariates may also be introduced when appropriate into the model to explain differences among school performance profiles.

Following Raudenbush, Fotiu, and Cheong (1999) therefore, we propose using Bayesian multivariate meta-analysis with a vague prior for all the school-level parameters. Suppose that we denote the fixed-effects vector $(\hat{\gamma}_{00cj}^m, \hat{\gamma}_{00cj}^r, \hat{\gamma}_{10cj}^m, \hat{\gamma}_{10cj}^r)$ for each cohort (and both mathematics and reading tests) obtained from the multivariate multi-cohort mixed-effects model for each school by $\hat{\gamma}_{cj}$, and its variance-covariance matrix by $\hat{\mathbf{V}}_{cj}$.¹² We further note that the fixed-effects estimates for any two cohorts within a school are independent. Then, given the school performance estimator γ_{cj} for cohort c , its estimates $\hat{\gamma}_{cj}$ is distributed multivariate normal with a known variance-covariance matrix $\hat{\mathbf{V}}_{cj}$, or

$$\hat{\gamma}_{cj} | \gamma_{cj} \sim \mathcal{N}(\gamma_{cj}, \hat{\mathbf{V}}_{cj}) \quad , \quad (4)$$

and, given its mean, ζ_c , in the population of schools, the estimator γ_{cj} is distributed multivariate normal with mean ζ_c and variance-covariance matrix Φ_c , or

$$\gamma_{cj} | \zeta_c \sim \mathcal{N}(\zeta_c, \Phi_c) \quad . \quad (5)$$

Equations 4 and 5 together is an example of a multivariate true-score model.

¹²To keep the presentation relatively straightforward, we ignore the impact of retention that is specific to the test subject, cohort, and school.

Inference for Functions of Parameters

For accountability purposes, Thum (2002a) had argued that a Bayesian formulation of the model, if implemented via a Markov chain Monte Carlo (MCMC) solution, gives the analyst considerable power for making inferences. One distinct advantage with this approach in accountability applications is that inferences can be easily constructed for any functions, linear or otherwise, of parameters. In Thum (2003), for example, administrators wished to know if students in a particular classroom had reduced, as a group, by 5% the gap between their performance on a test the previous year and the performance target. Standard errors for this ratio involves a crude approximation via the delta method under likelihood inference, but Bayesian inference is straight-forwardly constructed when we are able to simulate draws of this ratio from its marginal posterior distribution. Thum (2002a) applied the same approach to construct inferences for the Academic Performance Index (API), a ratio performance standard favored by California's accountability program. We take a similar approach here, to be detailed in subsequent sections, when we propose procedures to evaluate whether a school-grade meets and exceeds a minimum growth threshold (*i.e.*, the school-grade's adequate yearly progress target).

Priors

We consider briefly priors for the unknowns (ζ_c, Φ_c) . For $s = 1, 2, 3, 4$, a non-informative prior for the system average growth factor ζ_{cs} is $\mathcal{N}(0.0, a_{cs})$, for some suitably large constant a_{cs} . We employed a conjugate vague prior for the system variance-covariance matrix of the fixed-effects: $\Phi_c \sim \mathcal{W}^{-1}(\Upsilon_c, \nu_c)$, *i.e.*, Φ_c is distributed as an inverse Wishart with prior precision Υ_c and degrees of freedom ν_c . Because the priors Υ_c for each cohort are independent, we set ν_c to be uniformly 4 to match the dimensionality of the variance-covariance components prior Φ_c . Most frequently, priors come from past results, selected with guidance from experienced analysts. In the absence of prior data, we may use construct a prior by keying on the range of plausible true-scores γ_{cj} given the data for a plausible guess of the prior precision for each random effect. Alternatively, and one may argue that this is not an entirely different use of the available data for the purposes of choosing priors, reasonable values for the prior precision matrix Υ_c , may be culled from among the observed variance-covariance matrices, $\hat{\mathbf{V}}_{cj}$. A persuasive rationale for looking to the data for prior information is that these priors lead to approximate likelihood inference (see also Wasserman (2000)).¹³

Initial Values and Convergence

Not unlike the rationale we described above for coming up with candidates for priors for the prior precision matrix, plausible starting values for this matrix can also come from the among the observed precision matrices. We employed WinBUGS (Spiegelhalter, Thomas, & Best, 1999) to estimate the meta-analysis model. For the 96 schools, we obtain rather stable results within 20,000 draws from the posteriors after 10,000 burn-in cycles. Thum (2003) reported similar experience as well in an earlier application involving some 60 schools.

Accountability Hypotheses & Results

Table 3 shows posterior mean (and posterior standard deviations) estimates for the system of 96 schools. The pattern of cohort final status means estimates are reasonable, with Cohorts 5-9 estimating overall fifth grade performance in the system (from 639.30 to 644.50 in mathematics,

¹³Priors for analyses in subsequent years will be readily available given out results this year's data block.

Table 3: Average System Cohort Regression Estimates(Posterior Means and Standard Deviations).

Cohort	Final Status				Growth Rate			
	Math		Read		Math		Read	
	mean	sd	mean	sd	mean	sd	mean	sd
2	574.30	2.36	580.50	2.27	37.08	1.98	42.59	1.56
3	599.80	2.25	608.10	2.24	32.22	0.90	35.88	1.04
4	627.00	1.98	633.30	2.14	31.63	0.58	31.60	0.53
5	644.50	2.11	651.30	2.18	27.46	0.43	27.03	0.47
6	641.30	2.00	643.80	2.04	25.99	0.46	22.54	0.41
7	637.80	2.08	640.50	2.06	22.50	0.78	18.03	0.66
8	639.30	1.94	642.70	2.01	25.90	1.11	16.53	1.14
9	639.50	2.04	647.10	2.21				

and 640.50 to 651.30 in reading). The steady increases in final status estimates from Cohorts 2 to 5 is also reasonable as they estimate, correspondingly, average system performance from grades 2 through 5.

Although there seemed to be evidence of steady progress through the grades for all cohorts on average, the average gains for the more recent cohorts had tended to be larger in absolute terms. The results seemed equally positive for both mathematics and reading in the system. (More recent cohorts are labelled, in our discussion, numerically smaller.)

Note however how the size of posterior standard deviations were smaller for cohorts with a longer time horizon within the data-block. Confidence intervals for system averages of cohorts 5 and 6 growth rates, for example, were the most narrow. On the other hand, the cohorts occupying the extreme ends of the data-block had posterior standard deviation estimates three to over four times as large.

Cohort Improvement

To move beyond impressions formed from merely “eye-balling” the trends in the coefficients and their posterior standard deviations, we constructed and tested hypotheses regarding the cohort growth rates for the system. Focusing only on cohorts with at least three possible measurements (i.e., Cohorts 3 through 7), and separately for mathematics and for reading, we seek answers to the following two questions:

Hypothesis 1 Are more recent cohorts growing more rapidly? If this is affirmed, then there appears to be support for an optimistic view of recent improvement in student learning.

Hypothesis 2 Is the most recent cohort (Cohort 3) growing faster than expected? This will be an even more interesting hypothesis if we have reason to expect a boost in the learning rate for this cohort. For example, we may have put into place systemic initiatives at the start of the 2000-2001 school year that were aimed at improving instruction in mathematics and reading for example.

Figure 2 shows two ideal scenarios when comparing the growth rates of successive cohorts to assess improvement in academic productivity over time. In the left panel, learning rates appear to decrease

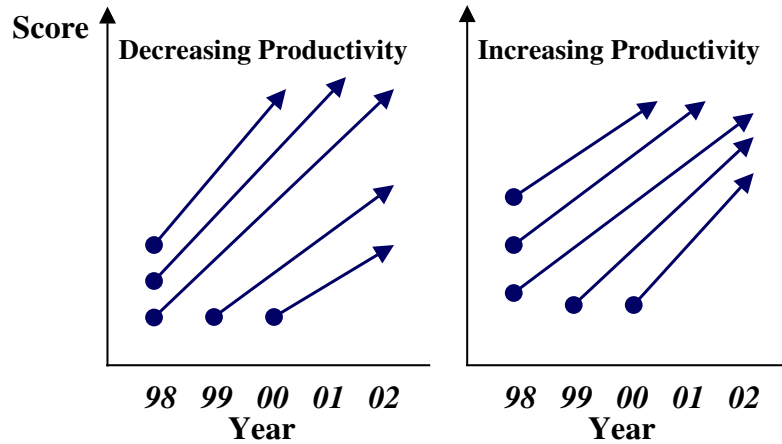


Figure 2. Comparing growth rates of successive cohorts to assess improvement.

over time and may signal dropping productivity. Improvement is likely the case depicted in the right panel when learning rates increase with each succeeding cohort.

Technically, Hypothesis 1 suggests that we pose a linear contrast,

$$\mathbf{h}_1 = \begin{bmatrix} -3 & -1 & 0 & 1 & 3 \end{bmatrix},$$

for the population parameters $\hat{\zeta}^{m'} = [\hat{\zeta}_7^m, \hat{\zeta}_6^m, \hat{\zeta}_5^m, \hat{\zeta}_4^m, \hat{\zeta}_3^m]$ in the case of mathematics and test if the trend is positive. In effect, we estimate the probability,

$$\text{Prob} \left(\mathbf{h}_1 \cdot \hat{\zeta}^m > 0 \mid \mathbf{Y} \right) \quad (6)$$

that, given the data and presuming that the model is adequate, the trend in cohort growth rates is positive. The evidence for successive improvement in mathematics and for reading growth over these cohort groups appear to be overwhelming. In the case of mathematics, the posterior mean for the gain in cohort learning growth rates is 2.52, with a posterior standard deviation of 0.26. The 95% highest posterior density (HPD) region for the overall increase in learning rate is 2.01 to 3.02 points per year. The results for reading is an increase of 4.48 points per year, with a posterior standard deviation of 0.26. The 95% HPD region ranged from 3.97 to 4.99 points per year.

A test for Hypothesis 2, which asked whether the most recent cohort (Cohort 3) out-performed the average of the previous Cohorts 4-7, takes a similar strategy with

$$\mathbf{h}_2 = \begin{bmatrix} -1 & -1 & -1 & -1 & 4 \end{bmatrix} .$$

The evidence again appear to be overwhelmingly in the affirmative. For each test subject, the probability that Hypothesis 2 is true nears 1.0. Of course an equally favorable picture is not likely for all schools. One indication is the marked level of heterogeneity among school cohort slopes as seen in Table 4. This is especially severe for Cohort 3, the most recent cohort. The mean estimate of Cohort 3's posterior standard deviation is 74.49 for mathematics, which is 1.5 to 3 times that of the preceding cohorts. The picture is even more exaggerated for reading where the mean of the posterior standard deviation of Cohort 3 is about 3 to 6 times that of the other cohorts.

Table 4: Between-School Variance Components of Cohort Regressions: Posterior Means and Standard Deviations of Standard Deviations and Correlations.

Cohort	Final Status				Growth Rate			
	Math		Read		Math		Read	
	mean	sd	mean	sd	mean	sd	mean	sd
2	551.70	87.19						
	0.94	0.01	503.20	80.04				
	0.54	0.08	0.42	0.09	371.90	57.45		
	0.39	0.09	0.39	0.09	0.80	0.04	222.70	35.04
3	450.40	71.05						
	0.92	0.02	449.50	71.21				
	0.53	0.08	0.51	0.08	74.49	12.06		
	0.27	0.10	0.44	0.09	0.76	0.05	101.30	15.91
4	406.10	64.16						
	0.93	0.02	476.10	74.22				
	0.38	0.10	0.29	0.10	27.12	4.64		
	0.30	0.11	0.37	0.10	0.78	0.05	22.12	3.96
5	399.60	62.02						
	0.96	0.01	426.40	65.80				
	0.46	0.09	0.45	0.09	14.31	2.57		
	0.31	0.10	0.38	0.10	0.86	0.04	17.79	3.15
6	401.60	62.93						
	0.95	0.01	411.40	64.45				
	0.34	0.10	0.26	0.11	16.47	2.98		
	0.13	0.12	0.20	0.12	0.55	0.09	12.06	2.46
7	394.30	62.48						
	0.95	0.01	385.10	61.18				
	0.27	0.10	0.15	0.11	54.09	9.02		
	0.08	0.11	0.05	0.12	0.73	0.06	36.57	6.69
8	348.00	55.70						
	0.94	0.02	379.40	60.83				
	0.24	0.11	0.10	0.12	101.10	17.05		
	0.06	0.12	0.07	0.12	0.65	0.07	106.80	18.36
9	365.50	61.62						
	0.94	0.01	430.90	71.31				

A simple summary of results for these two hypothesis, for mathematics and for reading, are displayed in Table 5, suggested that the majority of schools showed strong improvement gains of late when we compare the growth evidenced by successive cohorts according to the tests for either hypotheses. Of course an equally favorable picture is not likely for all schools. One indication is the marked level of heterogeneity among school cohort slopes as seen in Table 4. This is especially severe for Cohort 3, the most recent cohort.

Making AYP-NCLB

Under the current federal accountability regime, this seems to be the over-riding concern. We suggest that if our outcomes have clear (enough) psychometric properties, for example that they are vertically equated measures on an interval score scale with clearly established performance

Table 5: Number of schools with $\text{Prob}(\mathbf{h}_v \cdot \hat{\gamma}_{10}^m > 0 \mid \mathbf{Y}) = \alpha$ by cohort growth Hypothesis 1 and Hypothesis 2

α	Hypothesis			
	$v = 1$		$v = 2$	
	Math	Read	Math	Read
0.0	9	1	11	2
0.1	1	0	4	1
0.2	1	0	2	2
0.3	1	0	1	0
0.4	2	1	2	1
0.5	1	1	0	2
0.6	2	0	4	0
0.7	2	0	1	1
0.8	1	0	3	2
0.9	6	0	8	7
1.0	71	92	60	79

standards, the following are helpful in evaluating if we can expect a grade in a school, or the whole school itself, to reach the NCLB target of 100% proficiency on reading and/or mathematics by 2014.

Using the relationship between cohort (c), grade-level (g), and year (t)

$$c = (\max(t) - t) + (g - \min(g) + 1) ,$$

predicted means for each grade and year are easily recovered from the cohort-specific growth parameter estimates for each school. It is easily verified that the predicted means for each grade and year are given by

$$\hat{Y}_{c-(\max(t)-t),t} = \hat{\gamma}_{00c} + \hat{\gamma}_{10c} \times \text{TIME}_{tc} .$$

From the predicted means for the school's data-block, we may then test the performance of each grade-level for improvement over time. In particular, we are interested in assessing whether the improvement rate for a grade-level in a school meets or exceeds the minimum rate (AYP-NCLB) given our best guess of how the grade-level is currently performing if the average performance of students attending the grade at this school can be expected to be proficient by 2014.

We present arguments first introduced in Thum (2002b, 2003). Figure 3 sketches the rationale for this criterion. In Figure 3a, let us suppose that we have predicted means (circles) for the grade-level in years $t = 1, 2, 3, 4$ and we define $X_t = \text{YEAR}_t - 2002$. The range of scores that is judged to be "Proficient" for the test in question is from the lower cut-score, C_L , to upper cut-score, C_U . Proficiency cut-scores for each grade and subject matter were obtained from published performance standards for the SAT-9, see Table 6.

If, for example in Figure 3b, we represent with the straight line, $\hat{\eta}_0^{(4)} + \hat{\eta}_1^{(4)} X_t$, $t = 1, 2, 3, 4$, our best estimate of the progress based on the available information, AYP-NCLB is the minimum growth rate

$$\hat{\delta}_L^{(4)} = \frac{C_L - \hat{\eta}_0^{(4)}}{12 - 4}$$

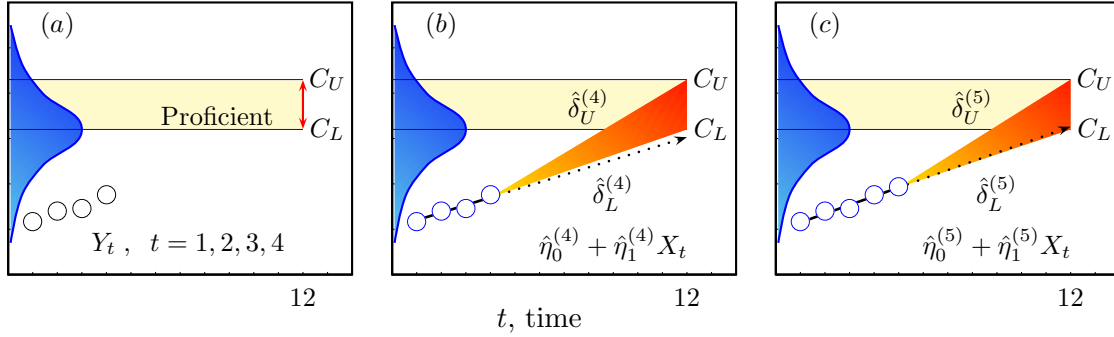


Figure 3. At any point in time, t , a school’s productivity is measured simultaneously along with its AYP. Attainment on scale scores are plotted on the vertical axis against time on the horizontal axis (Thum, 2002b).

Table 6: Cut-scores for Proficient Performance Standard for Grades 1 through 5 on the SAT-9.

	Grade				
Subject	1	2	3	4	5
Math	538	586	618	640	667
Read	539	601	624	646	669

required of this grade between $t = 4$ (the current year) and $t = 12$ (the target year) for the projected average attainment to meet the lower cut-score for proficiency in 2014. Therefore, a grade “makes AYP-NCLB” with probability α if the present growth rate, $\hat{\eta}_1^{(4)}$, exceeds AYP-NCLB, $\hat{\delta}_L^{(4)}$, or

$$\text{Prob}(\hat{\eta}_1^{(4)} > \hat{\delta}_L^{(4)} | \mathbf{Y}) = \alpha . \quad (7)$$

For the case depicted in Figure 3b, the available evidence suggest that we should not expect the projected grade performance average to exceed the cut-score during the target year. The case looks better in Figure 3c, because the projected growth rate appears to be co-incident with the required growth rate. We do not however to have much confidence in the result however because this is expected to occur at about the chance level. A more defensible policy consideration might require that $\alpha \geq 0.80$, for example, as part of the “makes AYP-NCLB” criterion. We also note, anticipating our illustration below for the individual school, compound hypotheses such as whether a grade-level, or the whole school, makes AYP-NCLB for both mathematics and reading are easily tested.

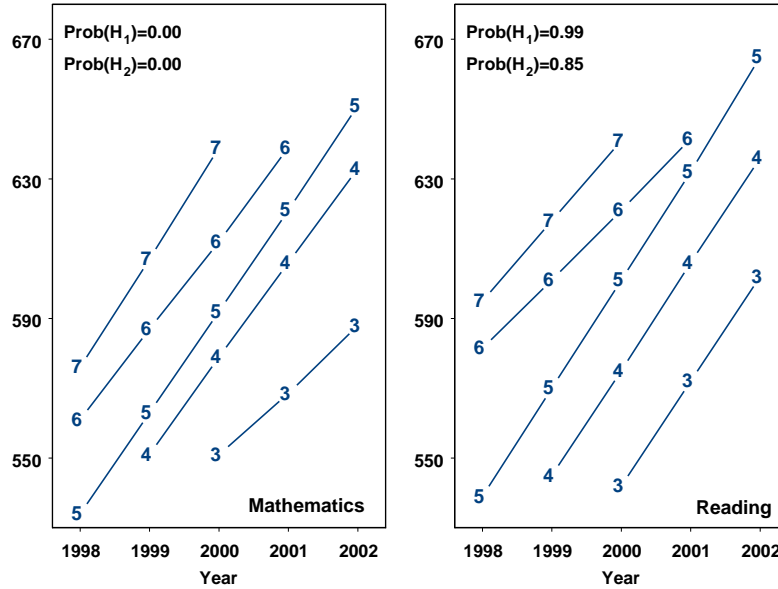
Individual School Productivity & School Improvement

We now turn to analogous hypotheses regarding school progress and improvement for the individual school.

Is your school getting more effective?

As we have seen above, a compelling question is how successive cohorts of students are progressing over time in a school. In particular, something positive is happening in a school if more recent cohorts are learning more (i.e., average scores for the same grades are higher) and perhaps also faster (i.e., slopes are steeper). We focus on the second issue in this round of analysis, and

Figure 4. Cohort Regressions for School 202. Plot symbols are cohort labels.



provide displays and assessments of two specific hypotheses (many others might be useful of course) comparing the estimated growth rates of the cohorts in the available accountability data-block.

Note that, in the above figures, the number labelling each point indicates Cohorts $c = 3$ to $c = 7$. The sub-figures in Figure 4 show the estimated cohort regressions for cohorts $c = 3$ to $c = 7$, on SAT-9 reading and mathematics for a sample school. Tests of the hypotheses, detailed below and also tabulated in Table 7, are reported to the bottom right of each graph.

As we have done earlier for the system, we may test changes in cohort growth rates over time within each school. Writing the school parameters as $\hat{\gamma}_{10}^m = [\hat{\gamma}_{107}^m, \hat{\gamma}_{106}^m, \hat{\gamma}_{105}^m, \hat{\gamma}_{104}^m, \hat{\gamma}_{103}^m]$ in the case of mathematics, we ask the questions:

Hypothesis 1: *Are more recent Cohorts growing faster?*

We estimate the linear rate of change in the cohort growth rates, and test $\mathbf{h}_1 \cdot \hat{\gamma}_{10}^m$ to see if this rate is likely to be greater than zero. The test statistic for reading suggest a 99% probability, see Table 5.3, that there is cohort growth rates are improving over time. There is no support at all for the same hypothesis with regards to mathematics.

Hypothesis 2: *Is the latest Cohort growing faster than expected?*

We contrast the growth rate of the most recent cohort (Cohort 3), with the average growth rates of the preceding cohorts, and tests $\mathbf{h}_2 \cdot \hat{\gamma}_{10}^m$ if this contrast is likely to be greater than zero. There is some, admittedly “weak”, support for this hypothesis in the case of reading but none whatsoever for mathematics.

Here, we return to the performance at each school. Figure 5 provides some details of this calculation with an example for grade 3 mathematics at a school in our study using all five years of data. Our best estimate for the present growth rate is 4.40 (sd=1.79). With the cut-score for

Table 7: Tests of Cohort Hypotheses 1 and 2 for School 202.

Hypothesis	Mathematics			Reading		
	mean	sd	conf.	mean	sd	conf.
1	20.9	1.5	0	31.5	1.7	99
2	-2.5	0.6	0	2.2	0.7	85

Confidence (%) of 80-89 is shaded blue, 90 or greater is green.

proficiency in grade 3 mathematics on the SAT-9 at 618, AYP-NCLB for 2002 is estimated to be 1.64 (sd=0.32) based on our best estimate of the grade-level performance in 2003 (598.30). At the pace determined for 2002, we estimate that the grade 3 mathematics average will exceed the cut-score in 2014 with 94% probability (all else being equal).

The tables in our individual school reports display estimates and tests for each grade and subject matter, evaluating simple or composite hypotheses. An example from School 2002 is given in Table 8 below.

For these displays,¹⁴ we overlay for each grade-level the predicted grade-level means (black dots), their current expected course (solid blue line), the projected course (green dots), the necessary course under AYP-NCLB (red dots and dashes), and the pertinent cut-score (gray dashes). Lastly, we provide an assessment of the likelihood (as percentage in the table and as probability in the graphs) that the school “makes AYP-NCLB” given our best estimate of its current progress. Returning to evidence presented in Table 8, the depicted school appears, under the AYP-NCLB standard for progress, to be reasonably healthy in grades other than 2 and 3.



Expected Proportion Proficient by Grade-level and Year by Subject Matter.

As we have discussed earlier, NCLB traces school progress in terms of PACs. Our analyses support the reporting of performance in terms of PAC, without incurring the loss of information and without sacrificing analytic coherence that would have resulted had we begun our analyses with the performance standard scores themselves.

Our tables, such as the example in Table 9, routinely provide the relevant estimates for each grade-level and subject matter, and results for the combinations of grades and subject-matter (i.e., for the entire school based on the information provided) will be available in the next rendition of the analyses. In this table, we also color-code grade-year performances that represented a 10% increase over the previous year, i.e., performances meeting the so-called “safe-harbor” provision. Again, probability estimates for this assessment will follow in future analyses. As we have noted before, and this is also the conclusion of many analysts, PACs are generally far more stringent measures of growth and this is clearly the case when we compare the implications of the evidence presented in Table 9 on this standard of progress with that offered above in Table 8.

Value-Added Progress Indicators

We have at the very outset limited our analysis to obtaining a description of the performance of a school in terms of its gross productivity. In one comparison setting, we provide an assessment

¹⁴Separate links (click on the icon  to graph and on the icon  to print) lead to the AYP-NCLB graphs for reading and mathematics.

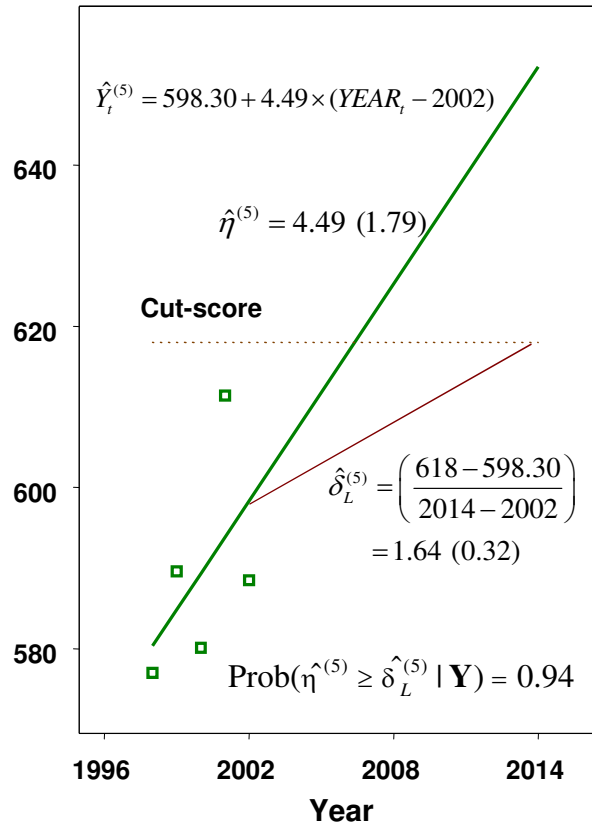


Figure 5. Third grade mathematics scores at a School 202 in our study are expected to be proficient on average with probability 0.94 in 2014.

of how likely each grade in a school appears to be faring, based on its past and present performance, in meeting an explicitly identified long-term performance target. In this instance, the growth rate estimate for a school based on t time points is $\hat{\eta}_1^{(t)}$ from Equation 7. We may consider $\hat{\eta}_1^{(t)}$ to be an estimate of the average gain up until time point t and the estimate of the AYP-NCLB growth standard, $\hat{\delta}_L^{(t)}$, to be the minimum gain for a school if it is to “make AYP-NCLB.” We provide a comparison in terms of the ratio of learning progress relative to an instantaneous standard specific to the school-grade (and subject matter), the result of which is a form of value-added indicator of progress towards a standard. In another setting, we compare the performance of succeeding cohorts in a school. Our Hypothesis 2, see Equation 6, refers to a different sort of value-added progress indicator; one which measures how much (and how likely) the most recent cohort is performing when compared to the average performance of past cohorts.

Other indicators are plausible, but we will just very briefly consider one other interesting idea offered by Bryk and Weisberg (1976). Working with one of our cohort regressions above, it may be useful to test the outcome we observe for year, $y^{(t)}$, with the projected output for that year $\gamma_0^{(t)}$, based on data observed prior to t . We may implement this indicator simply by posing the

appropriate design on time (leaving out the residual terms), for example,

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 & -4 & 0 \\ 1 & -3 & 0 \\ 1 & -2 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \xi_0^{(5)} \\ \xi_1^{(5)} \\ \Delta^{(5)} \end{bmatrix},$$

for a growth model where $\Delta^{(5)}$ is the value-added estimator of interest at $t = 5$ (see also Raudenbush (2001, pp. 520–521)). The idea here is to fix the comparison to each cohort, asking how much more learning have occurred at t over and above what we expect (due simply to maturation perhaps) as we start year t . A sketch of the basic logic is given in Figure 6 for $t = 5$. We note that the same analysis may be applied to any questions regarding growth or change for an accounting unit over time, such as a school, a grade, or a teacher.

Some Additional Tools for School Comparisons

Reliability of Estimates

A Bayesian multivariate meta-analysis of school-specific fixed-effects has several additional advantages. It produces in a straight-forward manner estimates of the reliability of each school-

Table 8: Does School 202 “makes AYP-NCLB” on reading and/or mathematics?

	Math 📊 📄				Read 📊 📄				Confidence (%)			
	Growth Rate		AYP-NCLB		Growth Rate		AYP-NCLB		Makes AYP-NCLB			
Grade	mean	sd	mean	sd	mean	sd	mean	sd	Math	Read	OR	AND
1	3.27	1.99	-1.33	0.44	-0.47	2.14	-0.11	0.45	97	44	97	44
2	-1.37	1.40	1.95	0.30	-4.41	1.49	3.12	0.31	2	0	2	0
3	4.23	1.29	1.66	0.26	2.09	1.35	1.52	0.28	95	64	95	64
4	8.06	1.38	0.73	0.26	7.72	1.44	0.54	0.28	100	100	100	100
5	4.88	1.74	1.65	0.28	5.36	1.89	1.04	0.31	94	97	98	94
									100	100	100	

Confidence (%) of 80-89 is shaded blue, 90 or greater is green.

Table 9: Estimates of the percentage of students at School 202 who are “proficient”.

Grade	Mathematics					Reading				
	1998	1999	2000	2001	2002	1998	1999	2000	2001	2002
1	53	61	66	66	82	53	54	53	52	52
2	24	11	9	15	16	23	7	4	7	6
3	4	2	1	8	19	11	8	4	14	22
4	3	1	4	14	33	5	2	7	19	38
5	4	4	2	9	17	9	9	8	19	29

The proportion proficient is magenta when it exceeds the previous year’s result by 10 % points.

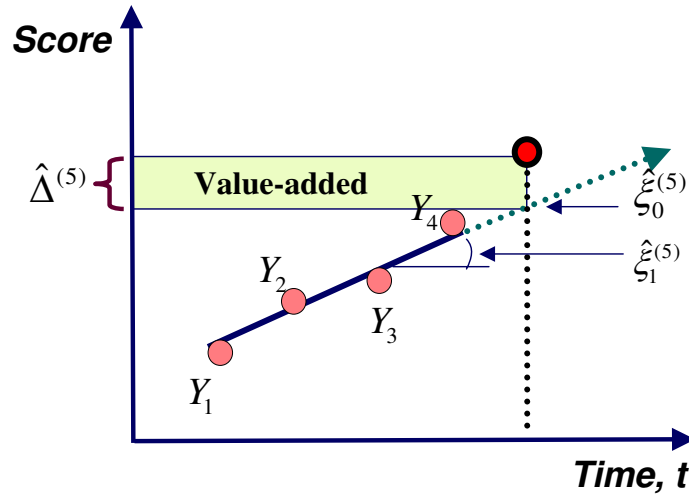


Figure 6. Value-added gain at $t = 5$ as the distance of the observed status relative to its predicted status based on past performance.

cohort gain and its approximate standard errors. For the Long Beach data, (Thum, 2002a) reported that two-year gains are relatively high and uniform, as indicated by their estimated posterior means (SDs) of .81 (.03) and .88 (.02) for 1999 and 2000 respectively.¹⁵ When reliability estimates are uniformly high, we expect shrinkage to be relatively mild and, consequently, the chance of observing dramatic re-ordering of schools based on their empirical Bayes estimate will be low.

Ranking and Comparisons

Several authors had argued against simple comparisons and rankings of estimates because these quantities are estimate with unequal precision (Goldstein & Healy, 1995, Goldstein & Spiegelhalter, 1996, Laird & Louis, 1989, Lockwood, Louis, & McCaffrey, 2002, Thum, 2002a). A Bayesian approach to inference for these quantities would produce more honest assessments of the level of precision attached to any such comparisons. As an example, Panel a of Figure 7 depicts the comparison between the California API gains of two schools (431 and 613), drawn from an earlier study employing data from a large urban southern California school district (Thum, 2002a). School 413 gained more, estimated at about 30 API points in 2000 with 90% credibility. Panel b plots the empirical median API rank estimates of school gains along with their 95% credibility intervals.

Productivity Profiles

Additionally, estimates may be compared on how they meet or exceed a set of graduated performance standards in terms of, in the context of school accountability, *productivity profiles* (Thum, 2003). A productivity profile is constructed from the simulated marginal posterior distribution of the school gain estimate. In Figure 8, 1999 (Panel a) and 2000 (Panel b) school API gains from the district may be compared with selected attainment levels (expressed as a percentage) of the

¹⁵We caution that this interpretation of reliability will nonetheless depend on the particular collection of schools employed in the analysis. It is relatively unproblematic when the set of schools approaches a simple random sample, or the schools make up the “population.”

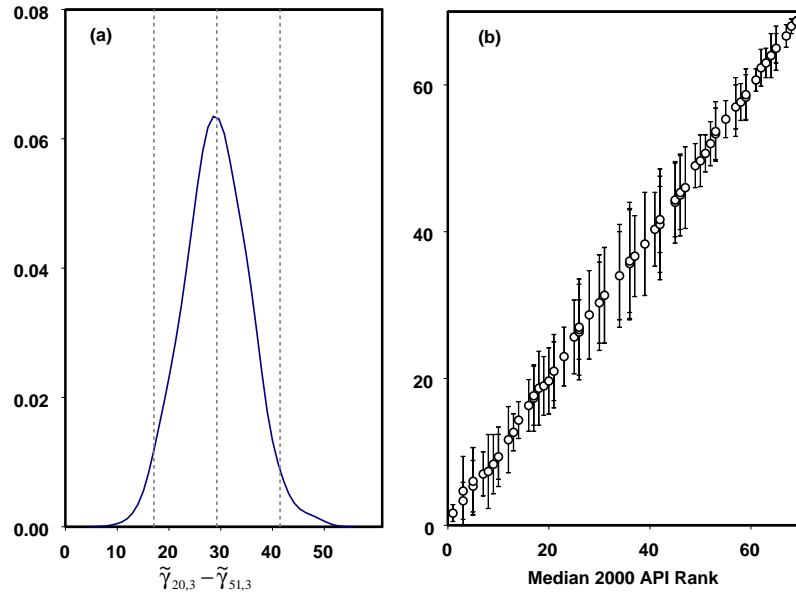


Figure 7. (a) Comparing School 431 and School 613 on their 2000 API gains, $\tilde{\gamma}_{431}$ and $\tilde{\gamma}_{613}$, respectively. (b) Ranking (with ties) of school median API's in 2000, set within their estimated 95% credibility intervals (Thum, 2002a).

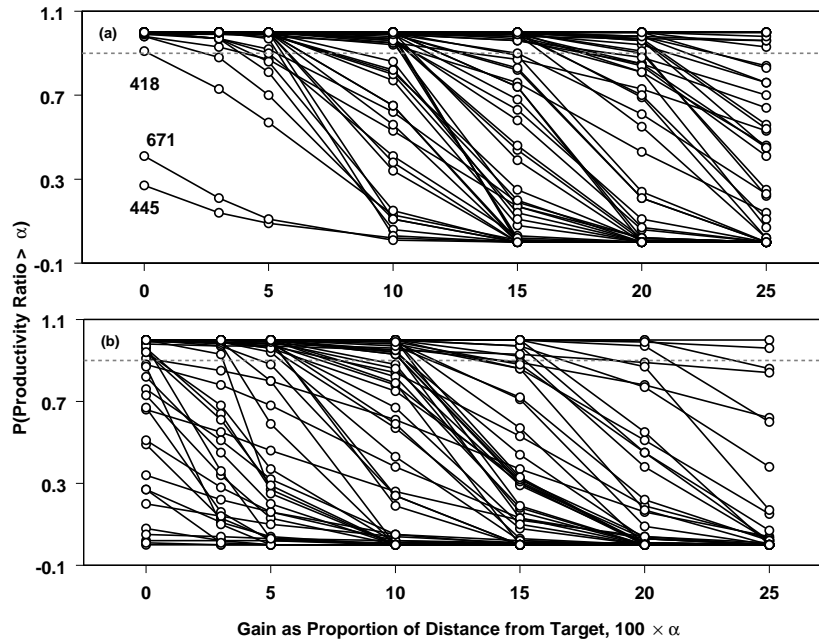


Figure 8. Productivity profiles, reflecting relative progress in terms of estimated gains against individual progress standard, for (a) 1999 and (b) 2000 (Thum, 2002a)

state-mandated performance target of 800 API points, controlling for their difference in the levels of precision of their estimates. Schools may now be more easily compared in terms of their productivity at any reasonable level of precision. A point on each line indicates the estimated percentage gain made by a school towards a target attainment level (horizontal axis) and how likely a gain as large as each target is observed in terms of a probability (vertical axis). We conclude that, other than schools 418, 445, and 671, most of the displayed schools made some gains from 1998 to 1999. The situation for the same schools appear different between 1999 and 2000.

Next Steps

We believe that, together, the above progress indicators provide an useful portrayal of the health of a school. At the very least, and this is the principal reason for our interest in these indicators, they help us identify schools that may be struggling. Note however how each of these indicators begin with a model that employs the student as his own control. This design serves usefully to control for the more stable impact due to student-level covariates. If the assumption is reasonable, each of these indicators provide one account of gross learning productivity. Clearly, this also means that the value of these indicators is severely circumscribed by unstated assumptions about the stability over time of the effects for other factors, such as changes in school composition, instructional practices, or school organization, that may impact the comparison we are making within and between schools.

References

- 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), 37–66.
- Broadfoot, P. (1996). Assessment and learning: Power or partnership? In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues*. Chichester, UK: Wiley.
- Bryk, A. S., Raudenbush, S. W., & Ponsiciak, S. (2003). *A value-added model for assessing improvements in school productivity: Results from the Washington, DC public schools and an analysis of their statistical conclusion validity*. Paper presented at the annual meeting of American Educational Research Association, San Diego, CA.
- Bryk, A. S., Thum, Y. M., Easton, J. Q., & Luppescu, S. (1998). Assessing school academic productivity: The case of Chicago School Reform. *Social Psychology of Education*, 2, 103–142.
- Bryk, A. S., & Weisberg, H. I. (1974, August). *A new approach to analyzing quasi-experimental data: value-added analysis*. Proceedings of the Social Statistics Section, American Statistical Association.
- Bryk, A. S., & Weisberg, H. I. (1976). Value-added analysis: A dynamic approach to the estimation of treatment affects. *Journal of Educational Statistics*, 1, 127–155.
- Erpenbach, W. J., Forte-Fast, E., & Potts, A. (2003). *Statewide educational accountability under nclb*. Retrieved November 1, 2003, from <http://www.ccsso.org/content/pdfs/StatewideEducationalAccountabilityUnderNCLB.pdf>. Washington, DC: Council of Chief State School Officers.
- Fitz-Gibbon, C. T., & Tymms, P. (2002). Technical and ethical issues in indicator systems: Doing things right and doing wrong things. *Education Policy Analysis Archives*, 10. Retrieved March 10, 2005, from <http://epaa.asu.edu/epaa/v10n6/>.
- Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, Series A*, 158, 175–177.

- Goldstein, H., & Myers, K. (1996). *Freedom of information: Towards a code of ethnics for performance indicators* (Tech. Rep.). London, UK: Institute of Education, University of London.
- Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A*, 159, 384-443.
- Gray, J., Jesson, D., Goldstein, H., Hedger, K., & Rabash, J. (1995). A multi-level analysis of school improvement: Changes in school's performance over time. *School Effectiveness and School Improvement*, 6, 97-114.
- Harker, R., & Nash, R. (1996). Academic outcomes and school effectiveness: Type "A" and type "B" effects. *New Zealand Journal of Educational Studies*, 32, 143-170.
- Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945-960.
- Lacey, C., & Lawton, D. (1981). *Issues in evaluation and accountability*. London, UK: Methuen.
- Laird, N. M., & Louis, T. A. (1989). Empirical Bayes ranking methods. *Journal of Educational Statistics*, 14, 29-46.
- Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 29, 4-16.
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute.
- Lockwood, J. R., Louis, T., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics*, 27, 255.
- Marion, S. F., White, C., Carlson, D., Erpenbach, W. J., Rabinowitz, S., & Sheinker, J. (2002). *Making valid and reliable decisions in the determination of adequate yearly progress*. Retrieved July 10, 2003, from <http://www.ccsso.org/content/pdfs/AYPpaper.pdf>. Washington, DC: Council of Chief State School Officers.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29(1), 67-101.
- Meyer, R. H. (1996). Value-added indicators of school performance. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 197-223). Washington, DC: National Academic Press.
- National Coalition for Parent Involvement in Education. (2004). *No child left behind act of 2001: An overview*. Website. (Retrieved March 23, 2005, from <http://www.ncpie.org/nclbaction/nclboverview.html>)
- National Conference of State Legislatures. (2004). *No child left behind: History*. Website. (Retrieved March 23, 2005, from <http://www.ncsl.org/programs/educ/NCLBHistory.htm>)
- O'Hagan, A., Stevens, J. W., & Montmartin, J. (2000). Inference for the cost-effectiveness acceptability curve and the cost-effectiveness ratio. *Pharmacoeconomics*, 17, 339-349.
- Raudenbush, S. W. (2001). Toward a coherent framework for comparing trajectories of individual change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change (decade of behavior)* (pp. 35-64). Washington, D.C.: American Psychological Association.
- Raudenbush, S. W. (2004a). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29(1).
- Raudenbush, S. W. (2004b). *Schooling, statistics, and poverty: Can we measure school improvement?* The ninth annual William H. Angoff Memorial Lecture. Princeton, NJ: Educational Testing Service.

- Raudenbush, S. W., & Bryk, A. S. (1987). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10*, 75–98.
- Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical linear models: Applications and data analysis methods* (2 ed.). Newbury Park, CA: Sage.
- Raudenbush, S. W., Fotiu, R. P., & Cheong, Y. F. (1999). Synthesizing results from the trial state assessment. *Journal of Educational and Behavioral Statistics, 24*, 413–438.
- Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*, 307–335.
- Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics, 29*(1).
- Rowe, K. J. (2000). Assessment, league tables and school effectiveness: Consider the issues and “let’s get real”. *Journal of Educational Enquiry, 1*, 73–98.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics, 6*, 34–58.
- Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103–116.
- Sanders, W. L., & Horn, S. P. (1994). The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation, 8*, 299–311.
- Simon, H. A. (1982a). *The sciences of the artificial* (second ed.). Cambridge, MA: MIT Press.
- Simon, H. A. (1982b). Theories of bounded rationality. In H. A. Simon (Ed.), *Models of bounded rationality: Behavioral economics and business organization* (Vol. 2, p. 408-423). Cambridge, MA: MIT Press.
- Spiegelhalter, D. J., Thomas, A., & Best, N. G. (1999). *WinBUGS: Bayesian inference using Gibbs sampling* [Computer program]. Cambridge, UK: MRC Biostatistics Unit.
- Stevens, J., Estrada, S., & Parkes, J. (2000). *Measurement issues in the design of state accountability systems*. Paper presented at the April 2000 Annual Meetings of the American Educational Research Association, New Orleans, CA.
- Stevens, J., & Moreno, R. (2004). *Multilevel, longitudinal analysis of ethnic differences in children mathematics achievement*. Paper presented at the April 2004 Annual Meetings of the American Educational Research Association, San Diego, CA.
- Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., et al. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29*(1).
- Thomas, S., & Goldstein, H. (1995). Questionable value. *Education, 185*, 17.
- Thum, Y. M. (2002a). *Measuring student and school progress with the California API* (CSE Technical Report No. 578). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Thum, Y. M. (2002b). *No Child Left Behind: Methodological challenges and recommendations for measuring adequate yearly progress* (CSE Technical Report No. 590). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
- Thum, Y. M. (2003). Measuring progress towards a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis. *Sociological Methods & Research, 32*, 153–207.
- Tymms, P. (1999). Baseline assessment, value-added and the prediction of reading. *Journal of Research in Reading, 22*, 27–36.

- US. (2001). *No Child Left Behind Act (2001)* [Legislation]. Pub. L. No. 107-110, 115 Stat. 1425.
- Wainer, H. (2004). Value-added assessment [special issue]. *Journal of Educational and Behavioral Statistics*, 29(1).
- Wasserman, L. A. (2000). Asymptotic inference for mixture models using data dependent priors. *Journal of the Royal Statistical Society, Ser. B*, 62, 159–180.
- Webster, W. J., & Mendro, R. L. (1997). The Dallas value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 81–99). Thousand Oaks, CA: Corwin.
- Willms, J. D. (1992). *Monitoring school performance: A guide for educators*. Washington, DC: Falmer Press.
- Wright, S. P., & Wolfinger, R. D. (1996). Repeated measures analysis using mixed models: Some simulation results. In T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren, & R. D. Wolfinger (Eds.), *Modelling longitudinal and spatially correlated data*. New York: Springer-Verlag.