No Child Left Behind: Methodological Challenges & Recommendations for Measuring Adequate Yearly Progress

CSE Tech Report 590

Yeow Meng Thum
CRESST / University of California, Los Angeles

March 2003

Center for the Study of Evaluation (CSE)
National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.1 Methodologies for Assessing Student Progress
Yeow Meng Thum, Project Director, CRESST/UCLA

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Institute of Education Sciences, or the U.S. Department of Education.

No Child Left Behind:
Methodological Challenges & Recommendations for Measuring Adequate Yearly Progress

Yeow Meng Thum[⋆]

Graduate School of Education and Information Studies
University of California, Los Angeles / CRESST

The recent federal push under the No Child Left Behind Act to finalize means for gauging school improvement has stoked public anxiety, but especially for public school officials who are charged with crafting and implementing defensible accountability systems. Other than the omnipresent political issues attending any proposal for shaping educational change, the methodological challenges of an accountability system are indeed daunting. Although not explicit in this paper in terms of specific references, this paper sketches a conceptual analysis and outlines recommendations for a viable accountability system that is based on recent research. Readers familiar with the issues will find a number of not-so-familiar answers to many of their concerns, including (a) what is value-added "productivity," (b) what data structure is sensible, (c) what might the minimum school-size be in order to attain the minimum level of precision, (d) how to define and estimate "adequate yearly progress", and (e) how to evaluate compliance under the so-called "safe-harbor" provisions.

1 Introduction

The *No Child Left Behind Act* of 2001 (NCLB, 2001) represents a much expanded federal role in public education as the new law outlined provisions for strengthening accountability for academic achievement. Besides requiring annual testing, the law seeks a method for judging school effectiveness, sets up a timetable for ultimate progress, and establishes a sequence of specific consequences

[⋆] The work reported herein was partially supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education. The core ideas explored in this paper are based in large part on earlier work described in Thum 2002a and Thum 2002b, studies supported respectively by the Milken Family Foundation and by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) respectively. Their applications in the context of NCLB are sparked by more recent conversations with supportive colleagues, in particular Dr. Arie van der Ploeg of the North Central Regional Educational Laboratory (NCREL). Arie van der Ploeg, Michael Young, and William H. Conrad, among others, provided invaluable feedback on earlier drafts. Later drafts have benefitted in particular from comments by Gary Cook, Joan Herman, Robert Linn, and Steve Raudenbush. 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@ucla.edu`.

for failure. With this legislation, the federal government appears to satisfy itself with the role of an arbiter of performance goals and progress, leaving the establishment of the evidentiary base for the judgments, including curriculum matters and the choice of assessment instruments, to the states. NCLB's immediate objective, it would seem, is a set of procedures that will help link assessments over time, across systems, and with external assessment components such as the *National Assessment of Educational Progress* (NAEP) to provide some validation of the system. A common measurement instrument for monitoring the education progress of the nation's children may be the key to producing a common currency for evaluating productivity, the lack of which is, to many, a major roadblock to fostering a coherent nationwide effort aimed at improving the debate on public education.

This paper dwells on the methodological challenges for student, school, district, and state accountability as formulated by the new law and provides some recommendations, based on recent research, for a viable approach for measuring progress of schools toward a set target. To further limit the scope of the paper, I will be leaving untouched problems related to the alignment of curriculum, standards, and tests, or to the choice of alternative tests and forms of testing. In particular, I focus on the following two sets of core issues of a useable accountability measurement scheme, namely, how to:

1. define, measure, and monitor the progress of *value-added* performance and productivity of multiple and nested reporting units (student subgroups, schools, districts, or states), and
2. define *adequate yearly progress* (AYP) under NCLB, and design a procedure to gauge and to compare progress of accountability units in terms of the AYP.

Conceived as a planning document, this paper aims to provide an analytic platform that will be transparent enough so that the discussion of the procedures for accountability measurement can be better de-coupled from the more contentious policy side of the current school accountability debate. It outlines the principal rationale for employing scale scores, using multiple outcomes, estimating value-added gains from student-level longitudinal performance data, requiring model-based aggregation, (e) requiring model-based inference, and keeping the *black-box* open in a viable accountability system. Within the same framework, it proposes a definition of what it means for a school to "make AYP" under NCLB. It shows that this notion of AYP, termed "AYP-NCLB", can be operationalized as a comparison at any point in time of a school's growth rate with a minimum growth required of that school if it is expected to be proficient by 2013-2014. The same analysis yields the proportion of the students in a school who are "proficient" each year, the primary interest of standards-referenced approaches to the assessment. Additionally, the proposed analytic strategy addresses directly issues pertaining to the precision of decisions, the choice of starting points for making evaluations, and the so-called "safe-harbor" provisions under NCLB.

As outlines and sketches go, please understand that the arguments in this paper are necessarily abbreviated. The proposals in this paper rely heavily on an ongoing study of the research of many scholars but, for the sake of readability, I will cite only the primary sources on which this document is based and leave it to the reader to inspect the references contained therein (Thum, 2002a; Thum, 2002b).

2 Using Test Scores

Like most accountability applications, NCLB's reliance on the use of student standardized assessment directs immediate attention to an ongoing debate on the proper use of test scores. There has been a long disquiet about the use of standardized test scores for making educational decisions that centers on issues of test score accuracy, even when only reliable and valid tests are used. Nevertheless, a test score can be useful if we carefully weigh its validity and its accuracy. See also the recent *Standards for Educational and Psychological Testing* (1999).

While it is clear that no test score will perfectly determine the student's achievement status, it is nonetheless based on a student's responses to a typically large enough sampling of domain tasks. As such, portrayals which only highlight the inherent inaccuracy of test results are alarmist if they pretend that test scores ought to be perfect to do their job. A test score, after all, is an estimate. It is merely an informed guess based on explicit albeit imperfect evidence of performance. For any reliable and valid test, a measure of how (in)accurate a particular score may be is found in its accompanying *standard error of measurement* (sem). Without explicitly taking the imprecision of scores into account, inferences about differences, whether positive or negative, may be biased and inferences will tend to be too liberal. The issue then is not that a test score can be off the mark, but whether it is biased in any particular way and by how much, and what impact will its imprecision have on individual decisions. For these reasons I recommend procedures that take explicit account of the standard error of measurement of scores.

3 Defining and Measuring Value-Added Productivity

A good understanding of the change in student learning is critical to improving public schooling. A value-added approach to the problem of measuring student learning seeks to locate that change within the student, isolated as best as possible from the many omnipresent factors related to the student's social and economic history, and to the makeup of his school and his community. While specific formulations of this basic idea may vary, and success cannot be guaranteed in any one instance due to relatively stringent technical and data-quality requirements (though necessary), a value-added approach holds the greatest promise for answering the question:

How are the kids in our schools learning?

Although the intensity of the current focus on measuring value-added productivity is relatively new, the methods for implementing the analysis are relatively well-established in the research methods literature. Thum (2002a) reviewed recently the methodological literature on the measurement of change and formulated arguments in support of some critical choices in constructing a statistical system for indexing academic progress. Some of the major conclusions are:

3.1 Metric Matters for Measuring Change

At the base of a viable system for measuring change is an interval scale outcome measured on a suitably equated metric. While the analytic work is being done on this underlying scale, categories that reflect ordered performance levels on the scale may be used for setting goals and for reporting results. Lacking an interval scaled on an equated metric, which represents a minimal metric requirement for a valid comparison of changes, will make a sound accountability analysis impossible – *value-added or not*. It is the responsibility of the test producers to continuously provide the necessary evidence for their scales, making explicit any shifts in procedures, conventions and modeling assumptions that are necessary components of standardization in measurement, so as to support the appropriate uses of their scales.

3.2 Multiple Outcomes Help

Multiple measures serve to replicate our readings on a performance construct, not merely as a hedge or a ruse to intentionally present a hazy target. When appropriately deployed, systems that employ multiple measures help us triangulate a more general performance construct that we understand is imperfectly represented by any one measure. Furthermore, the informational redundancy in multiple measures also helps to reduce the impact of measurement errors. A multivariate analysis, one which treats all the test scores as outcomes simultaneously, will provide a more coherent set of results when compared with attempts to rationally integrate separate analyses of individual test subjects.

3.3 To Measure Change, Estimate Gains

Of the several approaches to defining value-added, only the student-level gain gives a congruent conceptual mapping of learning change. The raw gain score is simply a linear composite of two positively correlated measures. Based on the widely accepted “true score” model, we may show that, if the component measures are relatively precise, the raw gain score has a variance that is smaller than the sum of the variances of each component measure, due to the correlation between true scores. Research has shown that the reliability of gains will depend not only on the precision of its components but also on the distribution of gains in the population, with the result that gain scores are not always less precise than either of their component scores. For example, if we clearly observe large gains but they are all equal in magnitude, the reliability of the observed gains – a normative measure of differences in gains above background noise – is zero.

And although the raw gain score may not be as inherently unreliable as previously thought, I recommend that accountability procedures *estimate* gains. This may be accomplished by putting all test scores on an equal footing as outcomes, rather than employ raw gain scores as the starting point for analysis. It is also easy to show that, because this particular value-added model employs the subject as his own control, individual-level factors (such as ethnicity and free-lunch status) that may have comparable impact on the student's performance at each testing no longer predict the gains he makes. However, classroom or school-level gains may be correlated with classroom or school-level measures of these same factors.

Finally, the estimated gain score has none of the conceptual and methodological difficulties that attend the residualized gain score obtained by regressing the student post-test score on his pre-test(s). Not only do its results depend critically on the particular make-up of the classroom or school, using the pre-test as a control when it is correlated with the outcome violates basic assumptions for linear regression (i.e., that predictors are fixed and known, and not correlated with the residual).

3.4 Require Model-Based Aggregation

Recent advances in school effectiveness research have shown that the story about school and student performance status and progress changes, sometimes irreconcilably, as we average test scores in various ways. It is well-known that, for example, the difference between the third- and fourth-grade test score means does not always equal the mean of individual student differences in third- and fourth-grade scores, unless the analysis involved the same students in both third and fourth grade. Not only does aggregation have a decided impact on the conclusions, it defines the conceptual unit of what is being measured and, as a result, whose change we are tracking. Reporting at the school, district, or state levels should be accomplished within a coherent yet flexible statistical modeling framework that begins with tracking individual student change and further allows proper inference of disaggregated results.

3.5 Require Model-Based Inference

Accountability systems, to be minimally useful, ought to supply defensible reliability estimates for their productivity scores. Calculations, projections, comparisons, and rankings of performance status and productivity not accompanied by explicit accounts of the various sources of measurement and sampling variability should be avoided because such uncertainties will impact the decisions based on raw estimates. All high-stake decisions should be qualified by properly constructed inferential statements in order to fully represent the extent of the usable information and the degree of precision. To be able to offer reliability estimates for productivity scores is essential to a defensible accountability system. Figure 1 and Figure 2 are examples, based on a cohort from the Long Beach

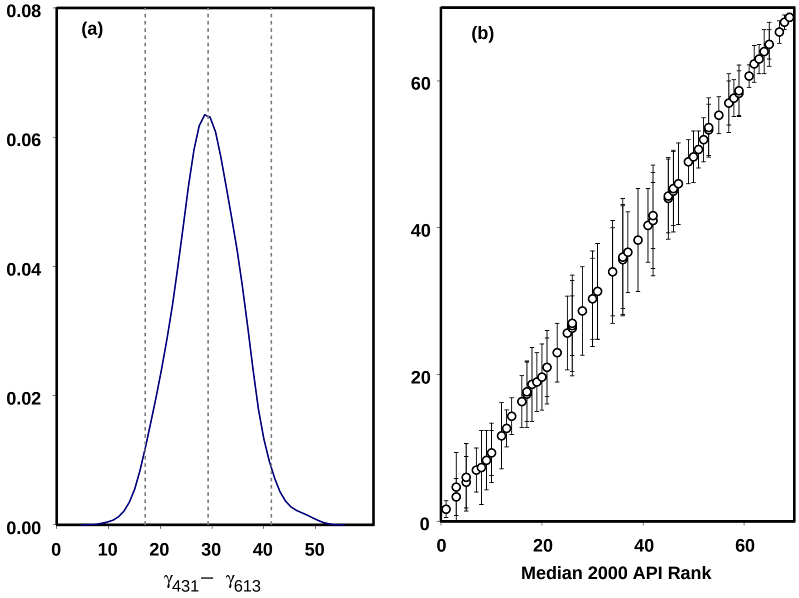

Fig. 1. In California, the API is a value-weighted composite of student performance. (a) Comparing School 431 and School 613 on their 2000 API gains, $\tilde{\gamma}_{431}$ and $\tilde{\gamma}_{613}$, respectively. Reference lines mark estimated mean difference at the 2.5%, the mean, and 97.5% points. (b) Ranking (with ties) of school median API's in 2000, set within their estimated 95% credibility intervals.

Unified School District (LBUSD), of a model-based comparison of the value-added gains between two schools and of the ranking of school status estimates, respectively. Productivity profiles, such as those depicted in Figure 3 and Figure 4, represent one approach to presenting how a teacher, school, or district is progressing *and* at what specific level of statistical confidence, *simultaneously*, given the available evidence (Thum, 2002a; Thum, 2002b).

3.6 Keep the Black-Box Open

To paraphrase Prof. Anita A. Summers of the Wharton School, not everyone needs to open the black-box of a potentially useful accountability model just as not every driver needs to understand how a car works to be comfortable with its use, *as long as someone is professionally charged with its design and safety.* We do however need to leave the key right on top of the black-box itself in order to facilitate ready access. Nothing should prevent the development of sound techniques as long as they are sufficiently open for peer review, professional evaluation, and systematic audit of their potential value or harm stemming from their deployment. And as the approach receives more and more realistic tests on the road, inadequacies of the methodology or myths regarding its practicality, both old and new, will be quickly identified and duly surmounted.

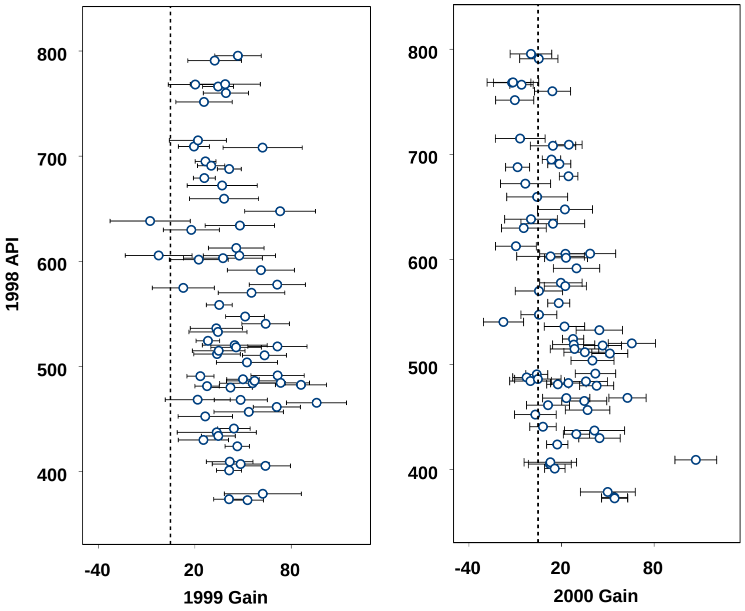

Fig. 2. LBUSD School API gain estimates and 95% interval estimates in 1999 and 2000 against their 1998 API status. 1999 API gain correlates -.31 (.12) with 1998 API status while 2000 gain correlates -.52 (.09) with 1998 status. Schools with the same 1998 status gain less in 2000 (-.10 points) than in 1999 (-.06 points) on average. Vertical reference lines marks 0 gains.

In the above, I have outlined the features of a more thoroughly rationalized approach, one that I expect will serve both diagnostic and perhaps accountability purposes better than many accountability systems currently in place. These same qualities make it a strong candidate as the core component of an accountability procedure. Unfortunately, this is also to suggest that the continued existence of most extant systems will be much harder to defend on conceptual and methodological grounds.

4 Adequate Yearly Progress

NCLB requires that *all* third- through eighth-grade public school students become proficient in mathematics and reading by 2013-2014, the accountability mechanism to be activated in 2005-2006. This would suggest that irrespective of the type of tests or a particular grade level, the student body is to find itself in the proficient performance level in reading and mathematics in 12 years' time. Its rhetorical purpose served, what it truly means in practice is ambiguous. For instance, not every student will be served in the same time period. And does NCLB require that a student

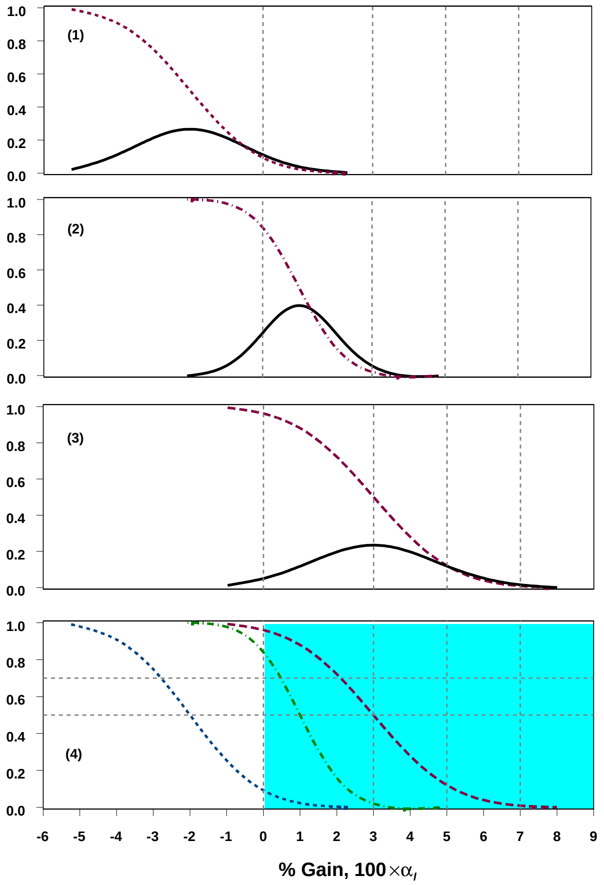

Fig. 3. The productivity profiles for three schools, given in Panels 1, 2, and 3, are overlayed in Panel 4 for easy comparison. Each is constructed from the simulated marginal posterior distribution of the school gain. A point on each line indicates the estimated $100 \times \alpha_\ell$ % gain made by a school towards a target attainment level (horizontal axis) and how likely a gain as large as $100 \times \alpha_\ell$ is observed in terms of a probability (vertical axis).

starting school in, say, 2012-2013 also be proficient in 2013-2014? Do students who will finish Grade 8 in 2006 need to be proficient at that time?

One reasonable interpretation would appear to be that NCLB intends for schools, not the individual student, to be on target in 2013-2014. In the interim, it is schools that need to show that they are be headed towards proficiency by 2013-2014; hence the importance of a clear definition of what we mean by *an accountability unit making AYP*. NCLB is primarily concerned with how each accountability unit is moving towards the target of 100% proficiency, or some number that is acceptably close to it, by 2013-2014. The important question of whom, among a school's student

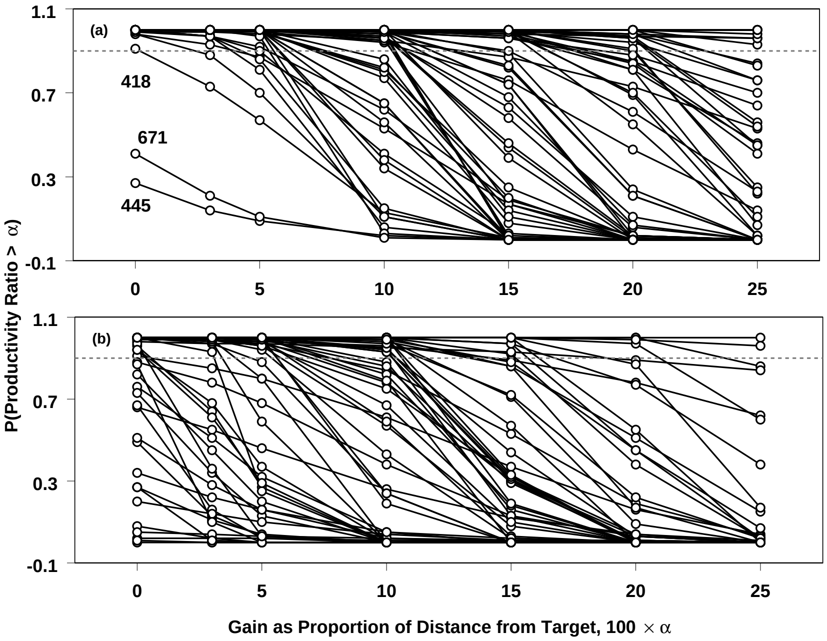

Fig. 4. LBUSD Productivity Profiles, reflecting progress in terms of the California PSAA ratio, for (a) 1999 and (b) 2000. Schools may now be more easily compared in terms of their productivity at any reasonably selected level of precision.

body, should be included in estimating its productivity at any point in time remains but it can be dealt with later.

Many current suggestions for AYP revolve around three ideas (Goertz, 2001). In Texas, schools must meet absolute thresholds on achievement and other criteria. Relative growth targets are employed in California. Michigan offers an example where a major goal is to decrease the proportion of students in the lower performance bands. These all make intuitive sense in their own way, and perhaps should all be monitored, as favored by some states, at least on some portion of the data as a part of a more comprehensive accountability strategy. Of intense interest to many state agencies, as a result of NCLB, is how each of their approaches may be re-articulated in NCLB terms.

Although almost all states set an intermediate (mostly annual) progress target in terms of an achievement level or rate, not all are clear about the time frame for attaining the eventual performance goal. For example, California requires schools to gain a fixed 5% each year given where they are on the API (academic performance index) from the interim statewide API target of 800, without however requiring a time frame for reaching it. NCLB, in contrast, sets a very clear timeline, 12 years, for all to attain the proficient performance level. For NCLB, therefore, AYP must involve a viable solution to the following central accountability question for schools:

> Given where you are at this point in time, are you improving at a pace that will put you on the specified target *in the remaining time frame?*

I will outline below how an explicit target and a specified deadline combine to suggested a notion of AYP that involves aspects of productivity and timeliness simultaneously. Specifically, my interpretation of NCLB suggests that

> AYP be defined as a minimum growth rate based on the amount of ground an accountability unit needs to make up in order to reach proficiency in the remaining time. At any point in time, the accountability unit is making AYP if it is improving at a rate that equals or exceeds its AYP.

I call this new composite “AYP-NCLB” to distinguish it from other AYPs in use currently. In AYP-NCLB, I evaluate progress towards a future target relative to a relevant performance baseline. As I will also show later, our analysis is easily modified to provide a direct assessment of whether a school may be expected to reach 100% proficiency by 2013-2014.

Besides its conceptual clarity, rates do not generally “bounce around” the way year-to-year gains do. Also *unique to my approach*, the accountability system is able to provide the following direct answer to the question posed by NCLB:

> We are P% confident that at this moment your school is making AYP-NCLB.

This statement can be clearly conveyed by the school’s productivity profile, as shown in Figure 3 above.

4.1 Data Is Part of Any AYP Definition

Given the high stakes involved in tracking school productivity, we need to be even clearer about the evidentiary base on which our estimate of the performance of a school rests. First, to avoid suggesting that a school’s productivity is a *trait* rather than a *state* (in that a trait is deemed less a transient quality than a state), we need to stress that a school’s performance is not only affected by the many factors relating to its composition and resources, it is especially critical to also recognize that *a school’s productivity is bounded in time.*

Furthermore, when describing a school’s productivity in 2006, for example, we need at minimum to be clear that our assessment is based on the relevant evidence available between 2002 and 2006. Another rule may be that a rolling block be employed, as shown in Figure 5, if we believe that older data may not be relevant given the school’s current conditions. I think that this practice is especially compelling because we expect that analytic models and data designs will vary across accountability systems. Specifically, I suggest that the 2002-2006 block of a school’s student assessment data be analyzed simultaneously with a multivariate mixed model that follows simultaneously all cohorts in the data block, each of which is represented by a lower-left-to-upper-right strand in Figure 5. After

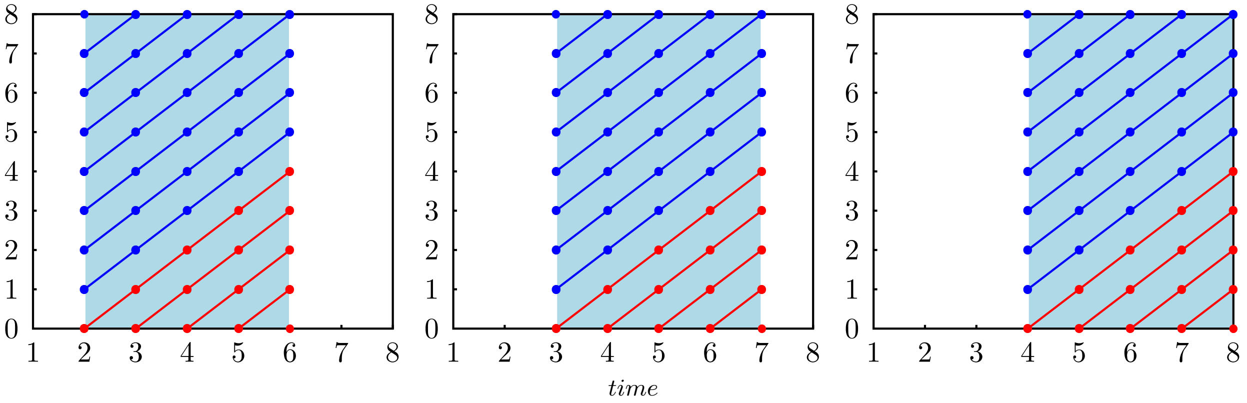

Fig. 5. A school's productivity at each point in time should be based on an explicitly designated (e.g., 5-year) block of the school's assessment database. The vertical axis is grade level, "0" being kindergarten. Note that while each strand represents a different longitudinal student cohort, real student test score vectors will contain missing components.

determining that our model adequately reproduces the data, we may then calculate fitted values along with estimates of their precision for all nodes, each representing subject matter, grade, and year (and sub-group) specific predicted status, gain, and growth rate. These results provide the necessary statistics for customized comparisons that respond to distinct accountability questions. For example, we may trace the annual achievement status made by the third grade in a school over the time to get a sense of progress for the third grade in the school. Change in productivity for third grade in the school can be assessed by comparing growth rates for the different longitudinal cohorts for each year. Furthermore, the procedure will also accommodate comparisons of school growth or productivity that take into account various student and school intake characteristics. Details of the model for employing this data structure for NCLB accountability will be forthcoming. In my view, this approach is also easily adapted for systems that prefer to track index scores, such as California's API.

4.2 Analysis vs. Accountability Unit

And, finally, we need to address a very serious concern, but fortunately mistaken in my view, expressed by many proponents of longitudinal data for accountability. While my reading of NCLB identifies the school and its sub-units as the ultimate units of accountability, *it does not necessarily imply that only the school-level outcome score is relevant.* My proposed procedure, as outlined above, will employ a longitudinal student database to characterize AYP-NCLB, in which *the unit of analysis remains (appropriately) the individual student, and the accountability units are student-subgroups and the school.*

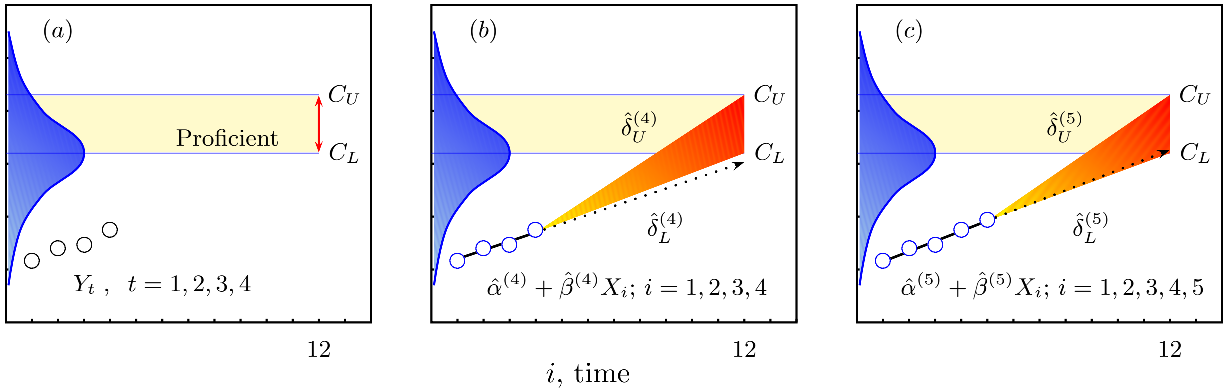

Fig. 6. At any point in time, t, a school's productivity is measured simultaneously along with its AYP. Attainment on scale scores are plotted on the vertical axis against time on the horizontal axis. (See text for further explanation.)

4.3 Making AYP Under NCLB

In order to measure productivity in terms of AYP-NCLB, I have extended an approach to measuring progress towards a target developed recently in Thum (2002a) and Thum (2002b). Here, I will provide only an outline, suppressing notation that would be necessary for representing multiple outcomes and multiple nesting units fully. For the limited purpose of this paper, the essential logic is sketched out, with the help of Panels (a), (b), and (c) in Figure 6, only for a school on a single test for a single grade. A full treatment of multiple measures and multiple criteria, including subtleties related to the form of the prediction function and heterogeneous error variances within a model for nested data, is forthcoming.

Scores and target. The distribution of scores for the population is shown in the backgrounds of each plot in Figure 6. Horizontal reference lines mark the lower and upper cut-scores for "proficiency" performance target, denoted by C_L and C_U respectively. Analysis will be performed using scale scores, even when *some mistaken readings of the legislation would suggest that the analysis implied needs to begin and end with performance categories.* Because we use performance standards that are defined on the original attainment score scale, we provide a more internally consistent evaluation of what it means, for example, to be "proficient" and, as a result, highlights the direct relevance of how performance standards are set. It also recommends an attractive, more internally consistent alternative when compared to performance standards that are imposed externally, such as the one-time fixed 5% annual growth on the API for California, that are not tied as closely to the domains being assessed.

Performance and productivity. In Panel (a) of Figure 6, we denote school progress on a single school-level outcome (Y_1, Y_2, Y_3, Y_4) for the first 4 years by "$\circ$". The deadline, according to NCLB, is $t = 12$. Using some simple linear model, for example, we may approximate how the school is performing at time, t, by estimating $\hat{Y}_t$ from

$$\hat{\alpha}^{(t)} + \hat{\beta}^{(t)} X_i \ .$$

If the predictor X_i encodes time in a way such that $X_i = i - t$, then $\hat{\alpha}^{(t)}$ gives a direct estimate of $\hat{Y}_t$. Note that the growth rate, $\hat{\beta}^{(t)}$, is simply an average gain estimate, which, as a performance measure, is less subject to wild fluctuations commonly observed for year-to-year gains. When we examine the behavior of $\hat{\beta}^{(t)}$ over time, we are studying the productivity of the school for the relevant time frame.

AYP-NCLB defined. Given the remaining time $12 - t$, the school will need to grow at a rate equaling

$$\hat{\delta}_L^{(t)} = \frac{C_L - \hat{Y}_t}{12 - t}$$

based on our best reading of where the school is at time t to make the target, bounded below by the cut-score C_L at $t = 12$. Similarly,

$$\hat{\delta}_U^{(t)} = \frac{C_U - \hat{Y}_t}{12 - t}$$

gives the rate needed to exceed proficiency. $\hat{\delta}_L^{(t)}$ and $\hat{\delta}_U^{(t)}$ form the lower and upper edges respectively of the "ray" from our best estimate of the school's current performance status to the target at time $t = 12$. Panel (b) and Panel (c) depict the school's growth rates and AYP-NCLB for $t = 4$ and $t = 5$, respectively. One immediate implication for almost all existing AYP formulations is that, because the ultimate target is fixed under NCLB, the intermediate target that is

> AYP-NCLB *changes over time.*

As we are able to better and better determine a school's progress over time, the common practice of setting *a fixed annual rate for an extended period of time makes little sense.* In addition to being specific to every point in time, my reasoning further suggests that AYP-NCLB is accountability unit-specific and is also specific to a test. Aggregating over tests and subgroup conditions will provide the appropriate estimates for assessing how the various accountability units satisfy NCLB.

Making AYP-NCLB. Although we may at any point in time calculate the observed proportion of students attaining proficiency, I feel that this picture is unrepresentative of the school growth trend and suggests the following alternative instead. Given our best productivity estimate at time t (i.e. $\hat{\beta}^{(t)}$) the school is making AYP-NCLB if

$$\hat{\beta}^{(t)} \geq \hat{\delta}_L^{(t)} \ .$$

Thus, whether a school makes AYP-NCLB involves a comparison of growth rates. This comparison yields an estimate of the school's eventual performance. That is, we expect the school to be performing at the proficient level by 2013-2014 if

$$\hat{\delta}_U^{(t)} \geq \hat{\beta}^{(t)} \geq \hat{\delta}_L^{(t)} \ .$$

In Panel (b) and Panel (c), projections for eventually reaching the target are given by black dotted lines. In our example, at $t = 4$, the school appears not to be making AYP-NCLB. By $t = 5$, the school appears to be making AYP-NCLB. Had the school fallen behind again at $t = 5$, it may be subject to some form of intervention. It is clear that, in AYP-NCLB, we continuously evaluate progress towards a future fixed performance target relative to a relevant performance baseline. It should also be clear that *comparisons to multiple targets*, for example as defined by other levels of performance or by sub-group specific performance levels, are also easily implemented.

Confidence of decision. Our results thus far relied on *comparing estimated growth rates*, one representing how the school is performing, $\hat{\beta}^{(t)}$, and the other serving as the interim AYP-NCLB benchmark, $\hat{\delta}_L^{(t)}$. Because both are estimates, it is important, as I have argued throughout, to characterize the level of certainty attached to their comparison after taking into account all known sources of variation. This is especially clear for our example above for $t = 5$, where the result seems – to some stake-holders at least – too close to call.

Specifically, we wish to know how likely it is that the school makes AYP-NCLB at time t, or

$$\text{Prob}(\hat{\beta}^{(t)} \geq \hat{\delta}_L^{(t)} | \mathbf{Y}) \ .$$

Similarly,

$$\text{Prob}(\hat{\delta}_U^{(t)} \geq \hat{\beta}^{(t)} \geq \hat{\delta}_L^{(t)} | \mathbf{Y})$$

gives the probability that the school is expected to be proficient when the deadline arrives at $t = 12$. Mindful of the high stakes that are attached to our accountability decisions and given some reasonable consensus about how confident we need to be before making the pronouncement that a school makes or does not make AYP-NCLB, we may select a range of confidence levels (e.g. 70%, 80%, or 90%) to help us arrive at a decision based on the data. If we wish to be more certain that a school makes AYP-NCLB at time t, we may select a 90% confidence level. In this case, we are at least 90% confident that the school makes AYP-NCLB if

$$\text{Prob}(\hat{\beta}^{(t)} \geq \hat{\delta}_L^{(t)} | \mathbf{Y}) \geq .90 \ .$$

If we are however uncomfortable about selecting any one particular confidence level, we may employ the productivity profile described earlier in Figure 3 to help us convey this more precise statement about the statistical confidence of our decision for a suitable range of confidence levels.

4.4 Evaluating Award Programs

Accountability decisions are difficult, and therefore careful data analyses are critical, precisely because a school's true improvement status is unknown. However, when school productivity estimates are qualified by confidence estimates in the manner described above, they support firmer accountability actions (awards or sanctions) at the school level. For example, if we make an award to a school whenever we are at least 90% certain that the school attained or exceeded its growth target, we would have provided a sound statistical basis for how a school is rewarded. Similar calculations may be performed for any significant subgroups, as well as for any arbitrary combination of sub-groups. Interestingly, this procedure may also serve as a *gold standard*, as no other is readily available, of various award regimes. As an obvious but intriguing example, we may then evaluate the screening accuracy of an alternative award program that does not take into account the precision of a school's productivity estimate whether it is based on comparing the school's performance to an explicit target using either a mean or a percent above cut-score (PAC) estimate. Lastly, I suggest without elaboration that this approach to setting and estimating the precision of an accountability decision circumvents direct speculation regarding a magic number for group-size, the so-called "minimum group-size," required for making a "statistically reliable" decision.

4.5 Alternative Starting Points

Several alternatives have been proposed in the legislation for use as baselines. For example, using 2001-2002 data, we may consider employing the predicted performance status at time t of the lowest performing subgroup within the school, or the lowest performing school in district, or the predicted average status of a subgroup of schools in the system in place of $\hat{Y}_t$ in defining $\hat{\delta}_L^{(t)}$. In fact, comparing each school to multiple baselines, some of which may be specific to important student subgroups, poses no additional burden for our analysis.

4.6 AYP-NCLB & % Proficient

Although I have designed in AYP-NCLB a prognostic tool in terms of the estimated average performance of students in a school at a point in time, we may easily *report* the analysis in terms of the estimated performance status for the individual student. Suppose that we denote by $\hat{Y}_{it}$ the predicted score of student i in the school at time t. Then student i has a probability

$$\text{Prob}(\hat{Y}_{it} \geq C_L | \mathbf{Y})$$

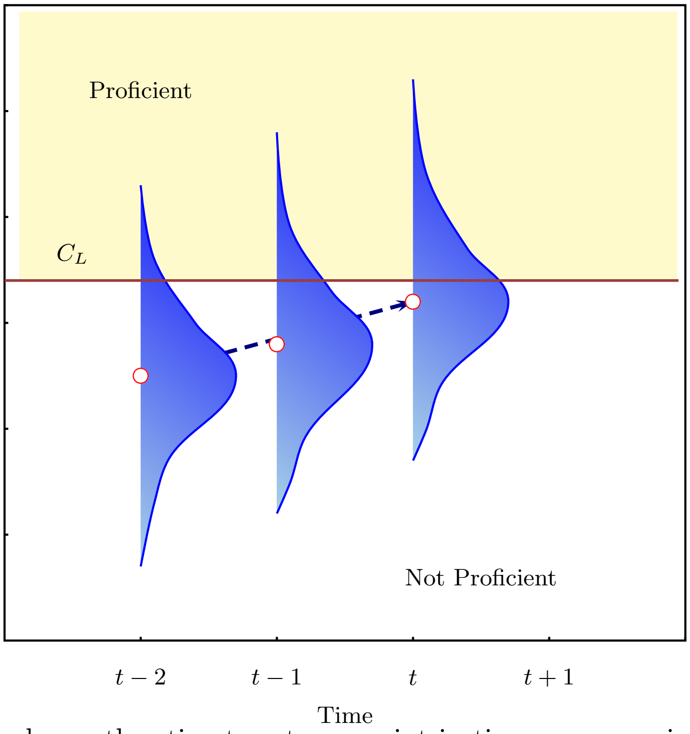

Fig. 7. Given our school growth estimates at any point in time, we may simultaneously track the reduction in the percentage of non-proficient students of a school. For each year, the proportion of non-proficient students is the proportion of the distribution below C_L.

of being at least "proficient" and the estimated percentage of students in the school at time t who are at least "proficient" is simply

$$\hat{P}_{jt} = \frac{100}{n_t} \sum_{i=1}^{n_t} \mathrm{Prob}(\hat{Y}_{it} \geq C_L | \mathbf{Y}) \ ,$$

where n_t is the number of students in the school at time t (see Figure 7). Using the school performance distribution for 2 consecutive years, t and $t-1$, we may address directly the so-called "safe harbor" provision by estimating

$$\mathrm{Prob}(\hat{P}_{j,t-1} - \hat{P}_{jt} \geq 10\% \ | \mathbf{Y}) \ ,$$

to assess what the likelihood may be that there has been a decrease of at least 10% of non-proficient students in the school over any 2-year period. Needless to say, conclusions based on AYP-NCLB and the various % proficient criteria can in some cases diverge. I expect that the latter is the more conservative and harder to estimate with reasonable levels of accuracy.

Because much of the language employed in the legislation to describe progress is in terms of this distribution, many agencies have understandably held up this summary as the starting and ending point for their accountability system. We already know that when we employ performance categories that are obtained by discretizing the original scores scale we are throwing away important information, rendering smaller, but in many cases meaningful, learning changes undetectable. As is also evident now, the appropriate analysis may begin with longitudinal student scale scores and aspects of the results may be presented in terms of the school performance distribution. Starting out with the observed cross-sectional school performance distributions would have grossly misrepresented the longitudinal character of growth and change in student learning.

5 Summary and Conclusion

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. In this article, I have provided a sketch of an accountability procedure that is responsive to NCLB. The proposed system would be best served by beginning with longitudinal student attainment data. To require that we start a standards-referenced test with attainment expressed on an interval scale is not a real restriction, in the absence of real alternatives. Test developers, who should be playing a more active role in support of the renewed nationwide accountability effort, should provide periodic assurances that their test scores are equated over grades and over time. They should also help to clarify the performance standards for their tests. For example, what are the performance correlates for "proficiency" on their tests? More guidance must also be directed at fostering the proper use of their test scores. With a sensible test measured on a suitable scale, elements of tested multivariate multilevel models may form the analytical core for estimating growth in performance, on which I overlay newly formulated procedures for making reasonable inferences about whether or not a school "makes AYP." I conclude by briefly addressing the following three widely expressed concerns:

1. While NCLB casts its attainment goals in terms of performance categories, I argue that this merely eases communication but does not alter our analytic focus on continuous student outcomes. For reasons given above, I have serious questions regarding the conceptual foundations of recent recommendations of so-called *standards-referenced assessments* over more "traditional" approaches (e.g., Schwarz et al., 2001). Why work with the degraded information in performance categories when we have multiple readings of the student's achievement that are summarized by a continuous interval measure? We should not confuse basic analysis principles with mere reporting convenience in building a sound accountability system. Performance on a criteria and performance relative to a set of norms are complementary information that are both important for assessing the health of public education.
2. We also need to avoid naively aggregating student results to the school level because, without explicitly modeling the nested nature of the data, important information regarding test and sampling sources of variation are irretrievably lost. And as I have noted earlier, this is quite

contrary to many current readings of NCLB: Just because NCLB explicitly targets schools for action does not mean that school-level results must be the starting point for analysis. I recommend estimating school-level summaries using student longitudinal data and then making inferences on functions of these statistics.

3. High standards are well and good, but preliminary analyses across the country question if the NCLB goals are unattainable in practice. And this is not to say we only set goals that we judge to be attainable, because one may easily question their credibility. Note that, as a standard, AYP-NCLB sets only a pace for reaching a performance standard in a specified amount of time. And while we may take the push for high standards for its motivational value, and even though AYP-NCLB provided a useful standard for progress in our view, certainly more needs to be known about what level of learning gains are possible in the many contexts of assessments in order to arrive at reasonable growth standards. At present, I agree that gain norms, as follow-ups to level norms, are obvious pieces still missing from the accountability puzzle.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Goertz, M. E., & Duffy, M., with Le Floch (2001). *Assessment and accountability systems in the 50 states: 1999-2000.* A CPRE Report.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Schwarz, R. D., Yen, W. M., & Schafer, W. D. (2001). The challenge and attainability of goals for adequate yearly progress. *Educational Measurement: Issues and Practice*, 20, 26–33.

Thum, Y. M. (2002a). *Measuring progress towards a goal: Estimating teacher productivity using a multivariate multilevel model for value-added analysis.* A Milken Famility Foundation Report.

Thum, Y. M. (2002b). *Measuring student and school progress with the California API.* CSE Tech. Rep. No. 578. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing (CRESST).