Accounting before Accountability

In this article, we will review several issues relating to the definition of "adequate yearly progress" (AYP), and its measurement, and offer some reasons why it may be prudent to separate the definition and measurement of AYP from its use in accountability. The design of accountability system includes general prescriptions about what data are relevant, policy questions that will guide the formulation of progress and productivity indicators based on students' longitudinal test scores, and procedures for using the resulting indicators in accountability.

Keywords: accountability systems, productivity indicators, No Child Left Behind, Adecuate Yearly Progress.

Nd016

Yeow Meng Thum[1]

College of Education.
Michigan State University
MI, USA
thum@msu.edu

Contabilizar antes de rendir cuentas

En este artículo se van a revisar algunas cuestiones relacionadas con la definición de *Adecuate Yearly Progress* (AYP) y su medida en el contexto de la legislación norteamericana "No Child Left Behind Act" (NCLB). Se ofrecerán algunas razones por las cuáles es prudente separar la definición y medida de AYP de sus usos en los sistemas de rendición de cuentas en los sistemas educativos. El diseño de estos sistemas incluye prescripciones generales sobre qué datos son relevantes, qué cuestiones políticas guiarán la formulación de los indicadores de progreso y productividad basados en puntuaciones longitudinales de los estudiantes y qué procedimientos serán necesarios para usar los indicadores resultantes para la rendición de cuentas.

Palabras clave: sistemas de rendición de cuentas, indicadores de productividad, *No Child Left Behind, Adecuate Yearly Progress*.

[1] This article is a prolegomena to the following article No Child Left Behind: Methodological challenges and recommendations for measuring adequate yearly progress. Dr. Y. M. Thum may be contacted at thum@msu.edu.

AFTER YEARS OF POLITICAL WRANGLING[2], the US federal government finally pushed through in 2001 a landmark legislation called the No Child Left Behind Act (NCLB) in an effort to reverse the widely recognized declining trends in the achievement of public school children[3]. The legislation signed into law an explicit deadline of 2014 for attaining universal proficiency, that is, every state must bring all its children to attain academic proficiency, or better, in mathematics and reading by the 2013-2014 school year. In the interim, state are in compliance with the law each year if they demonstrate "adequate yearly progress" (AYP) towards the eventual goal. Failing to make AYP for two successive years identifies a school as needing improvement and key to NCLB's accountability provision; the school will be subject to sanctions. It is easy to see why the notion of AYP plays such a central role to standards-driven school accountability reform in the US. We review below several issues relating to the definition of AYP, and its measurement, and offer some reasons why it may be prudent to more diligently separate the definition and measurement of AYP from its use in accountability.

While achieving universal proficiency by 2014 may be a very laudable political goal, researchers have been quick to point out that it may be unrealistic educational policy given what we know about the prevailing rate of academic progress among students. For example, judging by the growth from 23% in 1996 to 29% in 2000 in the percentage of eighth graders who are proficient in mathematics on the National Assessment of Educational Progress (NAEP), Linn (2008) argued persuasively that universal proficiency is clearly unattainable and annual targets will be less and less realistic as we near 2013-2014. Linn (2008) also examined the eighth grade results in the 2003 Third International Mathematics and Science Study (TIMSS). He concluded that even Singapore, a country which sits consistently at the top of international comparison in mathematics, will find universal proficiency an unrealistic target.

And even if one leaves aside the question of whether the NCLB target of universal proficiency, or any substitute, is realistic or not, a second set of obstacles makes their attainment difficult in the US.

[2] See, e.g., Ravitch (2004) for one viewpoint.

[3] It is generally acknowledged that since the late 1970's student test scores in the US have generally declined, e.g., Stedman (1995, 1997), although there are some vocal non-believers, see e.g., Berliner and Biddle (1995). Agreement on the reasons for the downturn in performance has been more illusive.

Unfortunately, but understandably so when it comes to the US federal system of governance, NCLB left it up to individual states to define what "proficiency" meant and to provide their own individual annual performance target. Educational researchers from outside the US must be reminded that US students, by and large, do not all sit for a common set of examinations at any point in their time in school. Testing programs in the US not only varied by state, but that many states frequently changed their tests and testing schedules. As a consequence, the psychometric problems associated with the effort to vertically equate non-identical test forms across grade levels, a required exercise to derive a usable developmental scale, are extremely challenging. These conditions have together seriously complicated the intended comparisons of achievement results over any extended time period, as would be necessary under accountability systems such as NCLB. The lack of more specific guidance in the law on key notions such as "proficiency" or "AYP" notwithstanding, it is the absence of a common assessment system in the US that has made both the notion of universal proficient performance illusive and a common formulation of AYP impossible[4]. As a result, proficiency standards vary across states, and it should not be surprising to find that the percentage of students who are proficient according to a state's standard may bear little resemblance to the level of proficiency based on an external and common standard such as the NEAP for the state (Vu, 2007).

Despite the ambiguities of the law and the uneven state of testing systems across the country, most observers would agree however that the new legislation gave real impetus to previously uncoordinated efforts around the country to build capacity for monitoring student and school progress in learning. Until recently, state-wide databases that track students longitudinally over time are also relatively rare, and are usually under-utilized for the purpose of monitoring progress even when they are available. Many accountability systems have since been proposed (see, e.g., Goertz and Duffy, 2001), with newer ones keep arriving. But many of the accountability systems that have been proposed have a clear policy-drive disposition to show positive

[4] An exception to this general lack of co-ordination in testing of public school students in the US is the (NAEP). Until recent plans to extend the program to grade 12 and more subjects, NAEP assessments are given periodically to national cross-sectional sample of student volunteers from grade 4 and grade 8 on mathematics and reading. While NEAP may be helpful in validation studies, states must rely on their own assessment programs as the only evidence base pertinent to NCLB accountability. See http://nces.ed.gov/nationsreportcard/nclb.asp for additional information.

progress under the law rather than a desire to document objectively achievement trends, evidence that should form the basis for corrective actions. As an example, the formulation of AYP has been driven by the need to avoid sanctions for as many schools and for as along as possible. It is in fact not unusual in some policy circles to reject proposals for accountability systems based on whether the projected short-term results might or might not be educationally realistic, or politically palatable. While we accept that universal proficiency by 2014 is unrealistic educational policy, such attempts could undermine the rational design of accountability systems.

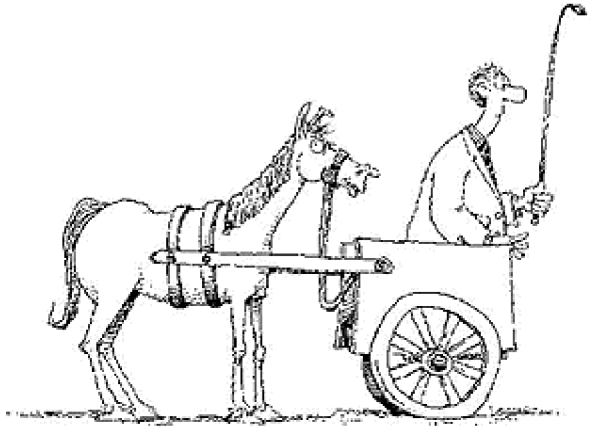
We have thus far dealt with systems that compared status over time. Another recent development in accountability models involved attempts to consider growth over time. It should be noted that these approaches generally do not adhere to more widely accepted approaches to the use of statistical evidence for decision-making. Even the recent clamor for including measures of student growth into accountability decisions have resulted only in layering into these existing "growth" models an additional set of conjunctive policy criteria (Ho, 2007). The result of chaining together multiple decision rules in these conjunctive models has made them less rather than more transparent when compared with applications of statistical growth models. The properties of such compound decision models are far less transparent when compared with more conventional statistical models. A caveat to remember in relation to modeling in general is that model estimates may be easy to come by but, without some idea about their properties, the estimates may be worthless.

In this paper, we think that one viable solution may be to clearly distinguish from the outset procedures that ensure the adequate description of student and school performance from procedures that deal with the policy goals in a working accountability system. Our rationale is quite simply that accounting should come before accountability. If the goal of an accountability system is to improve education, educators will appreciate that, as in any evaluation, objectivity would be harder to achieve if other peripheral political considerations are allowed to enter the evidence-marshalling measurement process prematurely. Allowing political concerns to shape the measurement or evaluation phase of school accountability would be a case of putting the cart before the horse[5]. Questions of

[5] The image in Figure 1 was retrieved from http://jilldenton.wordpress.com/2007/11/01/.

fairness and validity would surely arise if policy makers tweak the definition of proficiency or the standards for progress (AYP) when it is found that an unacceptably high number schools in a district fails to make AYP.

Figure 1.
Sound measurement, "the horse", ought to precede evaluation and accountability decisions, "the cart"



Aside from tinkering with the evaluation criteria, the temptation to game the system, for example by withholding lower-ability students from testing, has been hard to resist (Finn, 2007). Other behaviors that have been noted included teaching to the test, focusing instructional effort on those students who are more likely to show progress in assessments, resulting in "inflated" test scores. When stakes are real, loopholes always seem more seductive. But these challenges need not be considered flaws inherent to tests or to accountability systems, but rather professional failings on the part of a segment of educators and administrators. So, it would appear that an imperative first step for designing a defensible accountability system would be to very consciously separate the policy goals for student learning from its measurement.

We offer below some suggestions for the design of a basic accountability system at a single school system, as outlined in Thum (2006). An accountability system, according to Thum (2006), begins with specific questions about various aspects of student performance

over time. For example, policy makers, educators, and other stake holders may find it meaningful to know if the learning rate of successive age-cohorts of students in a school might have improved over time. Or, stake-holders may be interested in monitoring whether the growth over time for a school's lower grades might be stronger than that for the upper grades for indications differential effectiveness occurring between the two grade groups. Each of the preceding questions can be translated into testable statistical hypotheses, resulting in a family of *school productivity indicators*. Schools may then be compared with external policy benchmarks or with each other in terms of their indicators profiles.

As we shall see below, the design of accountability system includes general prescriptions to be jointly determined by stake-holders about what data are relevant, policy questions that will guide the formulation of progress and productivity indicators based on students' longitudinal test scores, and procedures for using the resulting indicators in accountability. To build progress and productivity indicators, we need to restrict our attention to school systems with testing programs and tests that support a description of how a school is performing in terms of the achievement measures of its students. Considering accounting systems that are based on a longitudinal student database at this juncture is timely for two principal reasons.

First, as it turns out, many of the accountability systems designed in response to NCLB thus far have been rather awkward, principally due to the lack of a suitable longitudinal database with assessments that have been placed on a psychometrically defensible developmental score scale. Most systems, in fact, employed only two years of data. When a developmental score scale is unavailable, many accountability systems would compare the proportion of students who are proficient in one testing with the proportion of students who are proficient in another. If the comparisons are to make sense, it would be necessarily assume that to be proficient on one testing has some relationship to being proficient on another testing, that is, we need to link the proficiency standards in some way. For virtually all tests, performance standards are set independent of each other and so their comparisons across tests are problematic. Recently, Lissitz and Huynh (2003) proposed a process for equating performance standards across tests that used a standard setting process to produce a vertically moderated "growth scale" for performance levels to enable their comparison over grade levels.

Interestingly, many systems that are currently in place have also opted to employ as basic input to accountability analysis student performance levels (e.g., proficient) even when scale scores are available. The rationale offered most often for this practice is that the NCLB AYP criterion is stated in terms of the percentage of proficient students. We point out that this practice may be flawed not only because focusing on the proficiency level discounts real growth that does not cross the proficiency cut-score, the use of a dichotomy as an outcome in place of the scale score makes it hard to detect growth due to the loss of information that occurs whenever one discretizes a more fine-grained scale. If it is believed that proficiency levels are adequately tied to the score scale, we would employ scale scores as the analysis metric and the report the results in terms of proficiency levels (reporting metric). Similarly, vertical articulation is unnecessary when a developmental score scale is available, as in the Stanford Achievement Test (SAT) series for example. Even when performance standards at different grade levels are determined by separate standard setting exercises, they may be directly compared by virtue of the fact that the performance cut-scores already reside on a common, vertically-equated score scale. We consider a vertically equated scale usable when scores across the range are interpretable, explanations of score differences are accepted, and understood, following the criteria for successful vertical articulation of performance standards given by Ferrara et al. (in press, footnote 2).

The second reason, and certainly a positive result of the legislation to many educational researchers, is that an increasing number of states are building databases with longitudinal student achievement results at its core, and thus making statistical growth models for accountability a more likely analytical option. With longer the time series, accountability results can be expected to be more stable, more persuasive, and consequently more helpful for gauging the health of schools and school systems.

We will limit the scope of our essay to two central methodological challenges for accountability proposals such as NCLB in the context of growth modeling with longitudinal student assessment data. First, we consider what it means from an analytical point of view to set a distant performance target and provide a statistical formulation for a test of whether aggregate student performance is on tract to achieving the eventual target in the given time line. We discuss in outline the principal rationale for the making the following choices: 1) employing scale scores, 2) using multiple outcomes, 3) estimating

value-added gains from student-level longitudinal performance data as opposed to other modeling approaches, 4) requiring model-based aggregation, 5) favoring model-based inference, and 6) keeping the analytic *black-box* open as part and parcel of a viable accountability system. Although necessarily limited if we were to place our "system" on-line as is and without further thoughts about implementing in parallel plausible validation procedures, we hope that the reader will find our proposals a useful point of departure for formulating an analytical platform for accountability measurement and clearly not naive.

Second, we provide a definition of AYP that, given current performance in relation to the eventual attainment target, reflects the effort required. It shows that our notion of AYP 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. We expect that this approach would apply equally well under NCLB, or any similar mandates. Readers familiar with the issues will also find a number of not-so-familiar answers to many of their concerns, including 1) what is meant by school "productivity," 2) what data structure is sensible, and 3) what might be the minimum school-size in order to attain the minimum level of precision. We note how the same analysis yields the proportion of the students in a school who are "proficient" each year, the preferred reporting metric for standards-referenced accountability measurement.

Although many of the main points have been made elsewhere, e.g., Thum (2003) and elaborated later in Thum (2006), we think that highlighting the central goals of accountability measurement in a way that is less constrained by the specific language of the legislation NCLB would be helpful. Koretz (2008) recently made a similar plea for building a better accountability systems based on assembling more systematic and richer evidence base and designing methods for their analysis rather than continuing to tinker with the details of NCLB. All we are concerned with in this and another paper published in *Revista de Educación* (see Thum, 2009) is the formulation and use of a set of productivity indicators based on longitudinal student test scores for monitoring student and school learning progress. The reader will notice that we have also avoid addressing another constellation of troublesome issues for learning about how schools and their students progressed that stems from inadequate inter-alignment of students' curriculum, instruction,

standards, and testing (Webb, 1997; Porter & Smithson, 2001) that very much characterizes many places within the US public education system. We have focused solely on measurement, description, and statistical inference for another reason. The temptation to draw causal attribution from within and between school comparisons of indicators is great, but we agree that valid causal attributions are not support by the very nature of the prevailing assessment design (Rubin, Stuart and Zanutto, 2004). These are important issues, especially in the context of the US, but their treatment is beyond the scope of this paper.■

Manuscript received: September 3rd, 2008
Revised manuscript received: November 15th, 2008

References

- Berliner, D. C. & Biddle, B. J. (1995). T*he manufactured crisis: Myths, fraud, and the attack on America's public schools*. Reading, MA: Addison-Wesley.
- Ferrara, S., Phillips, G. W., Williams, P. L., Leinwand, S., Mahoney, S. & Ahad, S. (in press). Vertically articulated performance standards: An exploratory study of inferences about achievement and growth. In R. L. Lissitz (Ed.), *Assessing and modeling cognitive development in schools: Intellectual growth and standard setting*. Maple Growth, MN: JAM Press.
- Finn, C. (2007, October 5). Dumbing Education Down. *The Wall Street Journal*. Retrieved May, 16, 2008 from the World Wide Web http://online.wsj.com/article/SB119154392619949671.html
- Goertz, M. E. & Duffy, M. (with Le Floch, K. C.). (2001). *Assessment and accountability systems in the 50 states: 1999-2000*. CPRE Report Series. Philadelphia, PA: Consortium for Policy Research in Education, Graduate School of Education, University of Pennsylvania.
- Ho, A. (December 2007). Growth models under NCLB: Back to basics [Electronic version]. *NCME Newsletter, 4*(15), 5-7.
- Koretz, D. (2008). The pending re-authorization of NCLB: An opportunity to rethink the basic strategy. In G. L. Sunderman (Ed.), *Holding NCLB accountable: Achieving accountability, equity, and school reform* (pp. 9-26). Thousand Oaks, CA: Corwin Press.
- Linn, R. L. (2008). Toward a more effective definition of adequate yearly progress. In G. L. Sunderman (Ed.), *Holding NCLB accountable: Achieving accountability, equity, and school reform* (pp. 27-38). Thousand Oaks, CA: Corwin Press.
- Lissitz, R. W. & Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research and Evaluation, 8* (10). Retrieved April 2, 2008, from http://PAREonline.net/getvn.asp?v=8&n=10
- No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).
- Porter, A. C. & Smithson, J. L. (2001). Are content standards being implemented in the classroom? A methodology and some tentative answers. In S. H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states One hundredth yearbook of the National Society for the Study of Education, Part II* (pp. 60-80). Chicago: University of Chicago Press.
- Ravitch, D. (2002). A brief history of testing and accountability [Electronic version], *Hoover Digest, 4*, Retrieved August, 8, 2008 from http://www.hoover.org/publications/digest/4495866.html
- 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.
- 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*(2), 153-207.
- Thum, Y. M. (2006). Designing gross productivity indicators: A proposal for connecting accountability goals, data, and analysis. In R. Lissitz (Ed.), *Longitudinal and value-added models of student performance* (pp. 436-479). Maple Grove, MN: JAM Press.
- Thum, Y. M. (2009). No Child Left Behind: retos metodológicos y recomendaciones para la medida del progreso anual adecuado. *Revista de Educación, 348* (enero-abril), 67-90.
- Vu, P. (2007). *Lake Wobegon, U.S.A. – where all the children are above average*. Retrieved May, 20, 2008 from the Stateline Site: http://www.stateline.org/live/details/story?contentId=172668
- Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education.* (Research Monograph No. 6). Madison: University of Wisconsin-Madison, National Institute for Science Education.

