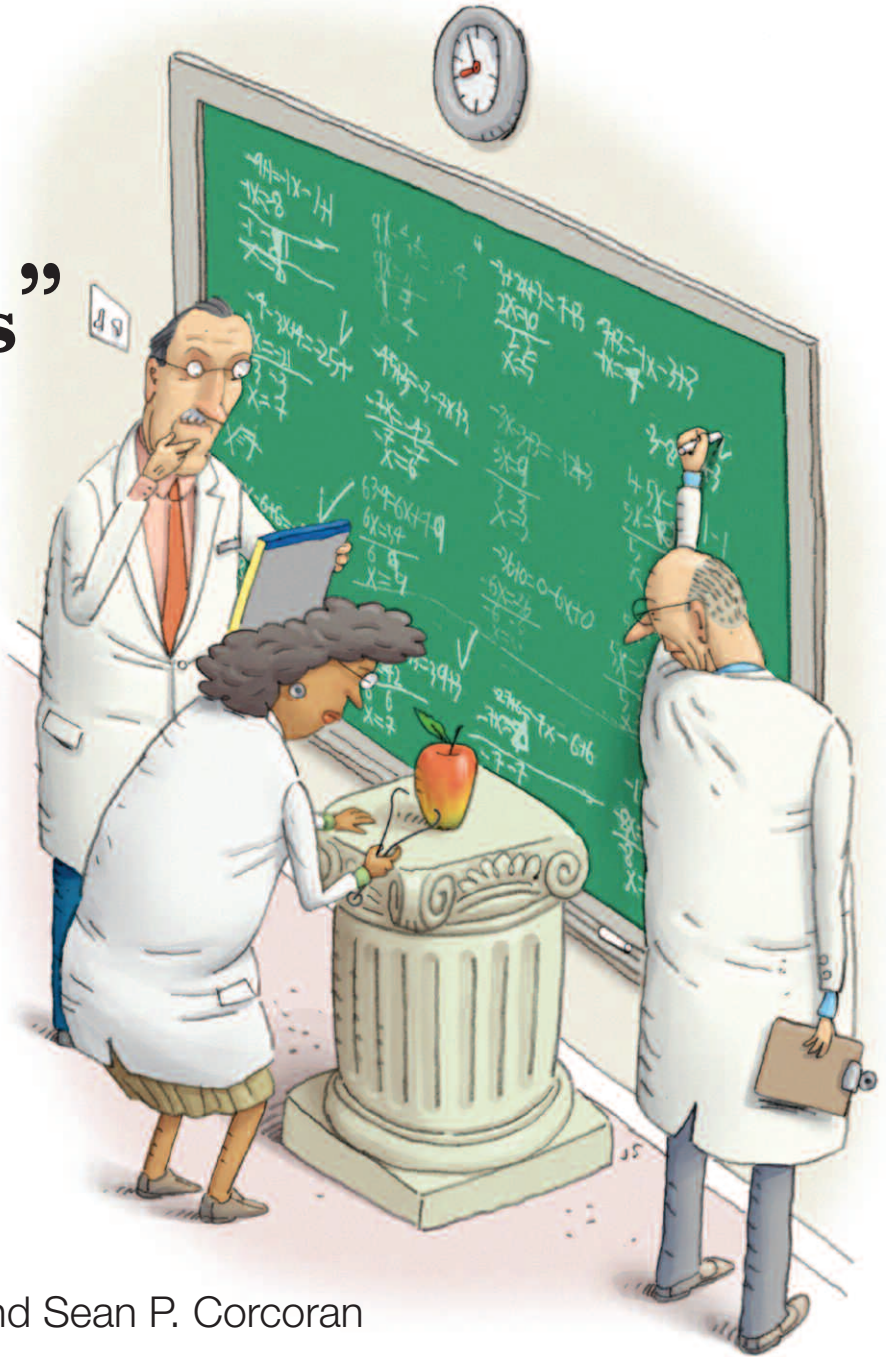


“Beware of Geeks Bearing Formulas”

Reflections on
Growth Models for
School Accountability

Growth models are being touted as the next answer in school accountability. But a poorly designed growth model is no better than the poorly designed proficiency model we have now.



By Jennifer L. Jennings and Sean P. Corcoran

In fall 2007, Ellen Foote, the principal of Intermediate School 289, got an unexpected surprise when she opened her mail. The New York City Department of Education had graded its schools on an A-F scale for the first time, using a method it would later call “the most sophisticated accountability system in the country” (Gootman 2008). What made New York City’s approach so different from No Child Left Behind-style accountability was that the primary factor dictating a

school’s grade was students’ year-to-year growth on standardized tests. To further emphasize “apples to apples” comparisons between schools, schools were compared not only with all schools in the city, but also with schools demographically similar to their own.

■ *JENNIFER L. JENNINGS is a doctoral candidate in sociology at Columbia University, New York City. SEAN P. CORCORAN is assistant professor of educational economics at the Steinhardt School of Culture, Education, and Human Development at New York University.*

Yet, Foote's decorated school — the only federally recognized Blue Ribbon middle school in the city and a regular on the "best schools" list in a popular school guidebook — had been stamped with an unsavory D (Medina and Gebeloff 2008).

One year later, the Department of Education issued its school progress reports a second time. And again, New Yorkers scratched their heads trying to make sense of the results. What does it mean, for example, when 77% of the elementary and middle schools that received an F in 2007 jump to an A or a B in 2008? Mayor Michael Bloomberg had a resolute answer to this question: "Not a single school failed again. . . . The fact of the matter is it's working" (Medina and Gebeloff 2008).

It makes no sense to label some schools as "in need of improvement" simply because their students were lower performing to begin with.

But Ellen Foote, whose school had received an A the second time around, wasn't so sure. Her school was the same, but her grade was very different. "A school doesn't move from a D to an A in one year unless there is a flaw in the measurement or the standardized test itself," she said. "We have not done anything differently, certainly not in response to the progress report" (Medina and Gebeloff 2008). Even less enthusiastic were the education measurement experts asked to comment on the wild grade swings between 2007 and 2008. Wrote Daniel Koretz (2008*b*), an educational psychologist at the Harvard Graduate School of Education, "[The New York City results show] far more instability from one year to the next than could credibly reflect true changes in performance. . . . It does not make sense for parents to choose schools, or for policy makers to praise or berate schools, for a rating that is so strongly influenced by error." Walt Haney, an education professor at Boston College, was similarly perplexed: "These [grades] are showing dramatic changes that can have nothing to do with what is actually happening" (Medina and Gebeloff 2008).

Here was a district that set out to combat the central problem with the federal accountability system: over-reliance on simple passing rates and the conflation of school and out-of-school influences on students' test performance. In short, New York City's system of measuring growth was intended to provide a better measure of school performance. Why, then, was this new system producing such dramatically different results from one year to the next?

The finer points of accountability systems arouse as much enthusiasm as an opportunity to watch plastic flowers grow. But as with all things, the details are key to understanding New York City's curious grade vacillations, as well as the promise and perils of "growth models" — systems designed to estimate the progress individual students make over time — for school accountability.


But there is no panacea here. While growth models improve on accountability systems in many ways, readers must also understand that a poorly designed growth model is no better than the poorly designed proficiency model we have now.

WHAT ARE GROWTH MODELS?

By now, you've probably heard terms like "growth model" and "value-added" tossed around in faculty meetings or in education policy debates. In practice, these terms are used interchangeably, but it's worth distinguishing between the two.

Growth models represent a wide range of approaches to assessing schools, all of which share one feature: a focus on student progress over time. Unlike NCLB's accountability model, which focuses on performance on a single test, growth models require repeated measures of performance for the same students over time. A basic growth model might simply ask: "By how much, on average, did math scores of the students attending this school improve over the course of a year?" Some of this growth is due to the work of the school and its teachers, while some is likely attributable to out-of-school factors.

Value-added models, on the other hand, are a type of growth model that goes further and attempts to isolate an individual school or teacher's impact on student progress. They accomplish this by first "predicting" student performance based on prior performance and a set of student characteristics — such as poverty — that are beyond the school's control. Schools (or teachers) with achievement that is higher than predicted are deemed to have high "value added." Schools that do worse than predicted have low "value added." These models are often quite technically so-

 PDK members can comment on this article at PDKConnect, the organization's new online community. Log in at www.pdkintl.org to join the conversation.

phisticated, but the idea is basically the same: accounting for student progress over time.

In this article, we use the broader definition of growth models. These models may or may not have a value-added component. There are advantages and disadvantages to the value-added approach. Though there is widespread discussion about using value-added models to evaluate individual teachers, we focus here on using growth models for school accountability.

IMPROVE ON NCLB ACCOUNTABILITY

What problems do growth models provide an answer to? The flaw of No Child Left Behind-style accountability — with its focus on passing rates — is that it does not attempt to separate school influences on achievement from out-of-school influences. We know that students' lives outside of school have powerful effects on how well they do in school. Using data from the Early Child Longitudinal Study, Kindergarten Cohort, we calculate that the average poor child in the United States arrives at kindergarten reading at lower levels than 72% of nonpoor children (see also Lee and Burkham 2002). And disadvantaged students aren't spread evenly across schools. Some schools disproportionately serve kids confronting the obstacles of poverty and low parental education, while others do not.

Under NCLB, schools serving advantaged kids got a leg up. Two equally effective schools — one serving an advantaged population, the other serving a disadvantaged population — get very different appraisals. But it makes no sense to label some schools as “in need of improvement” simply because their students were lower performing to begin with. Nor does it make sense to give other schools a free pass because their students were proficient when they walked through the door.

A second set of problems is created when schools are evaluated, not by student progress, but by their success in pushing students over the proficiency bar. States can and do game the system by setting a low bar for proficiency, a phenomenon that becomes apparent when we line up results from state tests and the National Assessment of Educational Progress (Ho 2007). We also know that some schools perform educational triage on their students, reallocating resources to students most likely to become proficient in the very short-term (Booher-Jennings 2005; Neal and Schwanzenbach 2007). Finally, under a proficiency-based system, the public gets a very limited view of how students are doing. We don't know if kids below or above the proficiency bar are making

progress; we only know what fraction jumped over the bar. We also have very little information about how far above or below the proficiency bar students really are. Moreover, policy makers are able to make misleading claims about declining racial achievement gaps based on proficiency rates, even as these gaps are unchanged or growing when we consider these groups' average scores.

In 2006, the U.S. Department of Education introduced a modified version of its NCLB accountability model, dubbed the “growth model pilot.” Unfortunately, this program offers little improvement over proficiency-based models (Weiss 2008). That's because the growth model pilot allows only for a growth-to-proficiency model, where all students are required to move quickly toward proficiency regardless of initial achievement. The growth model pilot doesn't permit true growth or value-added models because they conflict with NCLB's goal of 100% proficiency. As a consequence, few schools that fail to make AYP based on their proficiency rates make AYP based on the growth provision. This is not because low-proficiency, high-growth schools don't exist — many researchers have found that they do — but rather because of an artifact of the pilot model's limited definition of growth (Weiss 2008; Downey, von Hippel, and Hughes 2008).

GROWTH MODELS AREN'T A PANACEA

To be sure, growth models offer an improvement over our existing proficiency-based system. But they aren't without their own challenges and pitfalls. Consider some of the technical and political challenges that must be addressed to successfully use growth models in a school accountability system.

Technical Challenges

Measurement error. What exactly is “measurement error”? Bear with us for a moment, because the concept is critical to understanding the central challenge of growth models. As teachers well know, a test score is just a proxy for students' underlying skills and competencies. A student's test score represents a combination of true skill and measurement error. This error may be a function of such idiosyncratic factors as skipping breakfast (which might hurt your score), the good fortune of having studied the material that appears on the test (which might increase your score over your true level of skill), or a dog barking during the test (which might decrease scores of all students in a classroom). When measuring growth as the change from one test score to the next, *both* scores will

>> READ MORE ABOUT GROWTH MODELS

Ballou, Dale. "Sizing Up Test Scores." *Education Next* 2 (Summer 2002): 10-15. www.hoover.org/publications/ednext/3365706.html.

Koretz, Daniel. "A Measured Approach." *American Educator* (Fall 2008): 18-39. www.aft.org/pubs-reports/american_educator/issues/fall2008/index.htm.

Welner, Kevin G. "The Overselling of Growth Modeling." *The School Administrator* 65 (June 2008). www.aasa.org/publications/saarticledetail.cfm?ItemNumber=10512&snItemNumber=950.

be measured with error, which can muddy our view of student progress.

If measurement error were constant across tests, then it would just cancel out when we calculate the difference between the two scores. But we know that measurement error is more likely to be random. In short, the more measurement error (or "noise") in the results, the harder it is to detect the "signal" that represents a school's actual contribution to growth in student learning.

Measurement error is the culprit that produced New York City's wild year-to-year swings in school grades. New York City relies on only one year of test data, in contrast to other growth models that incorporate multiple years of annual growth to smooth out the measurement error that creates the appearance of a particularly good or bad year. In this way, New York City's foibles offer an important lesson to other states and districts that are considering growth models: Ignore measurement error at your own risk.

Scaling. The scaling of standardized tests is an arcane enough science that few of us have thought very much about it. But like measurement error, getting the scaling right is key to a valid, accurate, and reliable growth model. Two types of scaling issues are important here. First, tests must be designed for the specific purpose of comparing student progress across grades. "Otherwise," as Dan Koretz wrote about the case of New York City, which had applied a growth model to a test not designed for this purpose, "one has no way of knowing whether, for example, a student who gets the same score in grades 4 and 5 improved, lost ground, or treaded water" (2008a). Yet many states and districts are ignoring the advice of psychometricians and estimating growth models with tests never designed for that purpose.

A second scaling issue is whether a certain amount of growth — say, 50 points — for a low-achieving student means the same thing as 50 points of growth for a high-achieving student (i.e., whether a math SAT gain from 400 to 450 means the same thing as a gain from 750 to 800). Growth models make the assumption that this is the case. A related test-design issue is that of "ceiling effects" — the idea that tests measure achievement only up to a certain point. A student with a 790 on the math section of the SAT can gain only 10 points before achieving the highest possible score, whereas a student with a 400 has much more room to grow.

And a final complication is not a concern of scaling, but of the existence of different growth trajectories for different types of kids. In faculty lounges and over lunch, teachers have long discussed students' varying rates of growth and have questioned whether it is easier to move struggling kids up or to push advanced students forward even further. Many teachers have also noticed that growth is not a linear affair. Students may make great strides in one grade, plateau in the next year, and then take off again in the next year for reasons that have little to do with teachers or schools. This issue can become even more pronounced with students who are English language learners and students with disabilities, and growth model architects will need to take these issues into account.

Political Challenges

Accuracy versus transparency. Sophisticated growth or value-added models may be the most accurate, but they are also the most difficult for parents and educators to decode.

From our experience observing New York City's use of growth models, we worry that the technical details can get so heady that few understand how the system works. One New York principal put it bluntly, "The formula for establishing peer groups is so confusing that I can honestly make no sense of it" (Medina 2008). Those designing growth models, then, must consider not only the accuracy of these models, but also the difficulties of translating them for public consumption.

Reorienting our views of what it means for a school to be "good." Most Americans currently think of "good schools" as those that have high scores, irrespective of how far forward the school moved its students. Growth models will provide a very different portrait of which schools are effective and, in doing so, disrupt the conventional wisdom about how local schools are performing. Some parents will be surprised

to find out that schools with low proficiency rates are doing much better than they realized. Other parents will be unhappy to learn that their high-performing school isn't pushing students forward on standardized tests. The key point is that growth models require us to relax longstanding assumptions about the relationship between proficiency rates and school quality.

The tension between more accurate — and arguably more fair — assessments of school quality and uniformly high standards for achievement is not one that can easily be resolved.

How much growth? Many policy makers favor the current version of NCLB because it provides an absolute, clearly defined proficiency target. Some argue that growth models remove a school's incentive to move low-performing students all the way to proficiency. The danger, however, is that setting unrealistically high targets for growth only replicates the NCLB model, which has required more rapid progress than has ever been recorded in American public schools. This tension between more accurate — and arguably more fair — assessments of school quality and uniformly high standards for achievement is not one that can easily be resolved. Ultimately, this will be a decision for the public and its policy makers to make.

CONCLUSION

Warren Buffett, commenting on the mathematical models that played no small role in recently bringing the world economy to its knees, recently advised, "Constructed by a nerdy-sounding priesthood using esoteric terms such as beta, gamma, sigma, and the like, these models tend to look impressive. Too often, though, investors forget to examine the assumptions behind the symbols. Our advice: Beware of geeks bearing formulas." (Segal 2009, p. A19)

Our advice to educators on growth models is the same. Under the right circumstances, growth models offer a marked improvement on our current accountability system and provide an important lens into how our schools are doing. But they are no holy grail. Like all models, they are only as good as the assumptions on which they are built. And we should never assume that any one piece of information — whether it comes from a proficiency model or a growth model — is sufficient to make summary judgments about school quality. ■

REFERENCES

- Booher-Jennings, Jennifer. "Below the Bubble: Educational Triage and the Texas Accountability System." *American Educational Research Journal* 42, no. 2 (2005): 231-268.
- Downey, Douglas B., Paul T. von Hippel, and Melanie Hughes. "Are Failing Schools Really Failing? Using Seasonal Comparisons to Evaluate School Effectiveness." *Sociology of Education* 81 (2008): 242-270.
- Gootman, Elissa. "In Brooklyn, Low Grade for School of Successes." *New York Times*, September 11, 2008. www.nytimes.com/2008/09/12/education/12school.html.
- Ho, Andrew. "Discrepancies Between Score Trends from NAEP and State Tests: A Scale-Invariant Perspective." *Educational Measurement: Issues and Practice* 26, no. 4 (2007): 11-20.
- Koretz, Daniel. "A Measured Approach." *American Educator* (Fall 2008): 18-39. a
- Koretz, Daniel. "Guest Blogger Daniel Koretz on New York City's Progress Reports." *Education Week*, eduwonkette blog, September 17, 2008. http://blogs.edweek.org/edweek/eduwonkette/2008/09/guest_blogger_daniel_koretz_on_1.html. b
- Lee, Valerie E., and David T. Burkham. "Inequality at the Starting Gate." Washington, D.C.: Economic Policy Institute, 2002.
- Medina, Jennifer. "P.S. 8 Principal Explains F on Report Card." *New York Times*, September 15, 2008. <http://cityroom.blogs.nytimes.com/2008/09/15/ps-8-principal-explains-f-on-report-card>.
- Medina, Jennifer, and Robert Gebeloff. "More New York Schools Get A's." *New York Times*, September 16, 2008. www.nytimes.com/2008/09/17/nyregion/17grades.html.
- Medina, Jennifer, and Elissa Gootman. "Schools Brace to Be Scored, A to F." *New York Times*, November 4, 2007. www.nytimes.com/2007/11/04/education/04reportcard.html.
- Neal, Derek, and Diane Whitmore Schwanzenbach. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." NBER Working Paper w13293. Cambridge, Mass.: National Bureau of Economic Research, 2007.
- Segal, David. "In Letter, Buffett Concedes a Tough Year." *New York Times*, February 28, 2009.
- Weiss, Michael J. "The Growth Model Pilot Isn't What You Think It Is." *Education Week*, June 18, 2008, pp. 28-29.

File Name and Bibliographic Information

k0905jen.pdf

Jennifer L. Jennings and Sean P. Corcoran, "Beware of Geeks Bearing Formulas": Reflections on Growth Models for School Accountability, Phi Delta Kappan, Vol. 90, No. 09, May 2009, pp. 635-639.

Copyright Notice

Phi Delta Kappa International, Inc., holds copyright to this article, which may be reproduced or otherwise used only in accordance with U.S. law governing fair use. MULTIPLE copies, in print and electronic formats, may not be made or distributed without express permission from Phi Delta Kappa International, Inc. All rights reserved.

Note that photographs, artwork, advertising, and other elements to which Phi Delta Kappa does not hold copyright may have been removed from these pages.

All images included with this document are used with permission and may not be separated from this editorial content or used for any other purpose without the express written permission of the copyright holder.

Please fax permission requests to the attention of KAPPAN Permissions Editor at 812/339-0018 or e-mail permission requests to kappan@pdkintl.org.

For further information, contact:

Phi Delta Kappa International, Inc.
408 N. Union St.
P.O. Box 789
Bloomington, Indiana 47402-0789
812/339-1156 Phone
800/766-1156 Tollfree
812/339-0018 Fax

<http://www.pdkintl.org>

Find more articles using PDK's Publication Archives Search at
<http://www.pdkintl.org/search.htm>.