Aiming for Efficiency
Rather than Proficiency

Derek Neal
University of Chicago and NBER

Aiming for Efficiency Rather than Proficiency

Derek Neal

February 3, 2010

University of Chicago and NBER

Introduction

The No Child Left Behind Act (NLCB) of 2001 required states to develop accountability systems that monitor the progress of all schools toward the goal of all children being proficient in reading and math by 2014. The act requires that states develop ways of measuring whether or not schools are making Adequate Yearly Progress (AYP) toward this goal, and it also requires that schools face sanctions and possible re-organization if they persistently fail to meet AYP targets.

NCLB is conceptually flawed for at least three reasons that are all linked to NCLB's focus on proficiency standards. First, the requirement that all children be proficient is rather vague, and thus reports concerning the proficiency of students are obvious targets for political manipulation. These concerns are heightened by the fact that NCLB delegates to the states the tasks of defining proficiency standards and measuring student performance relative to these standards.

Second, it is likely not socially efficient to require that all children reach any single proficiency standard. Most social welfare functions used for policy analysis imply that, even if teachers devote efficient effort levels to all students, expected achievement levels will still vary considerably among students because of differences among students in their current and past home environments. Thus, it is hard to imagine why the goal of any education policy should be to induce schools to make sure that all students simply pass a common threshold. Further, existing empirical work suggests that using proficiency counts to measure school performance has the practical effect of causing schools to divert resources away from students who are currently far below existing proficiency standards, and it seems doubtful that this is a socially desirable response to the introduction of an accountability system.

Third, given any reasonable definition of proficiency, NLCB's threats of severe sanctions for failure to reach 100% proficiency by 2014 are not credible, and therefore, the text of NCLB creates confusion and uncertainty concerning how the law will actually be enforced in the future. This uncertainty may weaken the performance incentives contained in NCLB as we move closer to 2014.

In sum, NCLB offers an accountability system that pairs vague, manipulable, and likely inefficient standards with threatened sanctions for failure to meet these standards that are not credible, and in one way or another, all of these problems arise from NCLB's flawed focus on proficiency. The federal government provides a small portion of the funding for public schools and has limited control of the resources that states and local districts allocate to education or the decisions of parents concerning how much parents invest in the cognitive and emotional development of their children. Thus, even if it were socially desirable for the federal government to guarantee that all children meet demanding proficiency standards in math and reading, the federal government does not have the capacity to achieve this goal.

The most the federal government can hope to do is to design accountability systems that require schools receiving federal funds to demonstrate that they

are using the federal, state, and local funds they receive efficiently. I argue that the goal of accountability systems should be to organize competitions among schools that serve similar students and to let the outcomes of these contests dictate which schools receive government funding and to what extent.

If education officials pursue the paradigm of requiring public schools to compete with each other for the privilege of receiving public support, they must develop relative performance measures that assess the outcomes of these contests. Here, I discuss two different types of relative performance measures. Value-Added Models (VAM) produce measures of the contribution of schools or teachers to the achievement growth of their students. Most applications of VAM methods produce universal rankings of educator quality because these models typically assume that quality is a fixed characteristic of the educators serving in a particular school.

However, relative performance measures in education should take seriously the possibility that some teachers may be particularly effective with some types of students but not with all students. Below, I discuss one method for deriving context specific measures of school performance. A Percentile Performance Index (PPI) tells public officials how often the students in a particular school or classroom perform better than students in other schools who began the year in similar circumstances with respect to their prior achievements, the compositions of their classmates, and their family backgrounds.[1] This index of relative performance employs only the ordinal content of assessment results but still provides information policy makers need in order to have any hope of deciding whether or not it makes sense to re-organize a given school and give a new staff the opportunity to prove they can do better.

Policy makers' first concern should not be whether or not students are reaching specific proficiency targets. The first order of business is determining whether or not students are learning as much as possible given the resources allocated to schools.[2] The performance frontier that corresponds to "as much as possible" is not a static concept, and it is not divorced from the family and community environments of students. Rather, relevant performance frontiers for schools are specific to the circumstances in which schools operate, and they evolve over time with developments in pedagogy and school management practices. Contests among schools that serve similar students provide information concerning which schools are operating at these frontiers and which schools are lagging behind. We will know that policy makers have chosen a path toward

---

[1] Barlevy and Neal (2009) propose this index as part of an incentive scheme. We consider a setting in which teacher quality is known by a planner, and teachers have been assigned to classrooms optimally, but teacher effort cannot be observed. We then derive an incentive system called Pay for Percentile that induces all teachers to provide efficient levels of instruction to all of their students. Although we developed PPI for use in a specific incentive scheme designed for use in a setting with equally talented teachers, it may also prove useful as a device for gathering information about differences in teacher talent. Briggs and Betebenner (2009) propose specific econometric techniques for estimating performance indices that are quite similar to our PPI measure.

[2] Optimal policy also requires that authorities allocate efficient levels of spending for all students, but given any level of spending, efficiency requires that resources not be wasted.

meaningful accountability systems when they abandon their focus on proficiency and take up the task of designing systems that force schools to compete for the privilege of receiving public funds. Further, by focusing on the design of competitions rather than the specification of standards, policy makers may realize that expansions of parental choice through charter schools and vouchers are complements to rather than substitutes for assessment based accountability.

Problems with Proficiency

On its surface, NCLB links the receipt of federal funds to the performance of students. Thus, it appears to be an effort on the part of the federal government to provide incentives for states and local districts to improve their performance. This view of NCLB implicitly claims that the proponents of NCLB are motivated by the belief that state education agencies and local school districts do not always monitor school performance in an effective manner. Yet, given this motivation, why does NCLB delegate to the states the tasks of defining performance standards and measuring performance relative to those standards? My wife and I periodically require our children to clean their rooms, but we do not delegate to them the tasks of defining standards for cleanliness or conducting the inspections of their rooms that determine whether or not these standards have been met.

If the education officials in a state believe that it is in their political interest to improve schools, they can adopt and implement their own state level accountability system without intervention from the federal government. If these officials do not believe it is in their interest to hold local schools accountable, is NCLB really an important constraint on their behavior? Any state can render NCLB approximately irrelevant simply by setting low proficiency standards that almost all students will meet whether schools perform well or poorly, and a recent literature charges that some states are pursuing this approach.[3]

There are many ways to compromise the meaning of a proficiency standard. The most obvious way is to simply set a less than demanding cutoff score given the item content of assessments and the scales used to score assessment results. However, states can also create artificial growth in their proficiency counts by making assessments easier over time without rescaling them or by introducing new assessments and falsely equating the scale of the new assessment to the scale of the prior assessment.

For those who think that monitoring the integrity of proficiency standards is a trivial task and that state education officials could never inflate proficiency rates by falsely equating successive assessments, I offer Table 1. The numbers in Table 1 come directly from the 2003, 2006, and 2009 *Illinois State Report Cards*.[4] They describe state-wide proficiency rates in math for 3rd, 5th, and 8th graders in Illinois for four years, 2002, 2005, 2006, and 2009. In 2002 through 2005, Illinois tested only 3rd, 5th, and 8th graders under NCLB. In 2006, when Illinois

---

[3] See Banderia et al (2009), Cronin et al (2007). Dee and Jacob (2009).

[4] See http://www.isbe.state.il.us/.

began testing 4th, 6th, and 7th grades as well, they introduced a new series of assessments for 3rd, 5th, and 8th grades. The state then conducted equating studies that set the cutoff scores for proficiency on the new assessment scale at levels that were allegedly equivalent to the original proficiency standards.

TABLE 1			
NCLB Math Proficiency Rates for Illinois			
	3rd grade	5th grade	8th grade
2002	74.2	62.8	52.5
2005	79.2	73.1	54.3
2006	85.6	78.6	78.2
2009	85.2	82.4	81.7
(06-05)/(09-02)	0.58	0.28	0.82

The last row of Table 1 presents the fraction of the total increases in Illinois 3rd, 5th, and 8th grade proficiency rates under NCLB that occurred between 2005 and 2006. These fractions are .58, .28, and .82 respectively. I know of no change in state policies governing teacher hiring, training, or classroom practice that could account for this result, and similar jumps in proficiency rates do not occur in any other year.

Do the results in Table 1 prove that Illinois officials manipulated proficiency rates by manipulating the 2006 equating studies? No, and that is not the point. The point is that NCLB's reliance on state-specific proficiency standards offers states opportunities to manipulate these standards in ways that are costly for federal officials to detect. The methods used to equate psychometric scales are quite complex, and the application of these methods requires elements of both art and science. Thus, any plan that sought to monitor the integrity of proficiency standards in all states, subjects, and grades would greatly increase the enormous costs that taxpayers already pay to fund NCLB compliance activities.

Aiming at the Wrong Target

Many advocates of NCLB may respond that NCLB should simply be amended to include national standards and a national system of assessments that measures proficiency relative to these national standards. However, before heading down this road, one must consider whether or not it makes sense to build accountability systems around the concept of proficiency in the first place.

Let us consider the goals of education policy through the lens of economics rather than the lens of "standards-based" reform. At the beginning of a given school year, assume that a particular student possesses a set of skills that can be summarized by a skill index, X. At the end of the school year, she will possess a new set of skills indexed by X'. For ease of exposition, normalize the units of our skill index so that $V = X' - X$ is the value to society of the progress our student makes during the school year. V may reflect improvements in her expected life earnings, her health as an adult, her efficacy as a parent in the future, as well as benefits that others derive from the enhanced skills that she brings to her interactions with family members, friends, and neighbors.

Let e denote the resources that this student's school devoted to her education. Imagine that e equals units of effective instruction directed specifically to our student, and choose the length of this time unit so that the cost of providing one unit of effective instruction in a safe, well supplied classroom is one dollar.

In this framework, what is the right policy target for X'? The proficiency framework that dominates education policy suggests that there is some proficiency target X'_p and that our student's school has performed well if $X' > X'_p$. However, beginning economics students will instead focus on a target for hours of instruction, e^*, such that when $e = e^*$, the following efficiency condition holds:

$$E\left[\frac{\partial V}{\partial e}\right] = 1$$

This condition states that when schools devote the efficient level of resources, e^*, to a student, the expected gain to society at the margin equals the marginal resource cost.[5] It is an expected gain because teachers cannot force any student to learn. There are forces beyond a teacher's control that often arise from shocks to a child's home environment or health or both that affect how fast a given child learns. Still, we expect that this condition will characterize optimal ex ante effort. It seems reasonable to assume that, for all students, the expected gains from instruction exceed the costs at low levels of e. However, the marginal gains from additional effective instruction should diminish at some point because the attention and energy of any given child are finite. Going from four hours on task per day to five should yield greater benefit on average than going from 14 to 15 hours on task per day.

Thus, for each student, there is a particular e^* that describes the efficient level of investment that she should receive from her school, and for each student, there will be a particular level of skill X'^* which equals her expected skill at the end of the year assuming that her school provides e^*.

Within this framework, it is difficult to imagine how one can create a rational justification for the use of universal proficiency standards in accountability programs. Given any realistic characterization of the differences in baseline skills that students bring to a particular grade, it is socially wasteful to devote the resources required to bring the expected achievement of *all* students up to any reasonable proxy for proficiency. For some students, the resource costs required to reach proficiency likely exceed the relevant values of e^* and thus, at the margin, exceed the expected value to society of bringing these students up to proficient skill levels. In addition, other students will be deemed proficient even if their schools choose to ignore them, $e = 0$.

Some may counter that, by linking sanctions and proficiency rates, NCLB compels schools to invest heavily in their most disadvanatged students and therefore minimize the number of students who reach adulthood without the

---

[5]Policy makers may also be inerested in simply maximizing the value of school output for a fixed level of school spending. The solution to this problem involves a similar condition, but in this case, di erent levels of school spending imply di erent shadow prices for school resources.

basic skills required for success in modern economies. However, the evidence suggests that this is not true.

In Neal and Schanzenbach (2009), Diane Schanzenbach and I examine test scores in Chicago following the introduction of NCLB and also the introduction of a city wide accountability system in 1996. Our study is unique because the tests used in the first years of both accountability systems had been in place for some time before these systems were announced. Thus, we are able to estimate how achievement evolves on these assessments when they are taken under low stakes in two different grades. Then, by examining the outcomes of students who took assessments both in the first years of an accountability program and several years earlier when the assessment involved no stakes, we can measure the change in test scores induced by the accountability systems at different points in the baseline achievement distribution. We consistently find that proficiency based accountability systems offer no benefits for students who have no reasonable chance of becoming proficient in the near term, and students at the very bottom of the baseline skill distribution may well have been harmed by the introduction of NCLB.

A significant ethnographic literature on "education triage" details the strategies that educators employ in response to high stakes testing systems built around cutoff scores. Systems that do not reward educators for helping students make progress that does not change their proficiency status shift resources away from those who are far below existing proficiency standards.[6]

Federal education officials have recently approved several waivers that allow states to measures the performance of their schools based on the growth of students toward proficiency rather than proficiency per se. This approach may mitigate the tendency for teachers to respond to NCLB by ignoring their least able students. However, these waivers leave states with systems that are still fundamentally flawed because the growth model plans adopted by states must, "Ensure that all students are proficient by 2014, and set annual goals to ensure that the achievement gap is closing for all groups of students."[7]

Credibility

NCLB's fixation on the goal of 100% proficiency for all students is not only difficult to justify on efficiency grounds, it may also render the act irrelevant in the near future. I will not explore all the gory details of the AYP calculations that determine whether or not schools are making adequate progress toward the goal of 100%. However, I note several themes that all raise concerns about NCLB's future credibility.

---

[6]See Gillbourn and Youdell (2000), Booher-Jennings (2005), and White and Rosenbaum (2007). Reback (2007) also provides evidence using large data sets on individual assessment results from Texas. Since he does not have access to assessments taken under low-stakes, he uses the details of the Texas accountability formulas to isolate children who o er high expected returns from extra attention and tutoring because their progress is given large weight in the calculations of their schools' performance ratings.

[7]See http://www.ed.gov/admins/lead/account/growthmodel/proficiency.pdf.

To begin, the AYP rules in each state must reflect NCLB's insistence that all children be held to the same proficiency standards. Thus, all state plans must contain a transition path toward requiring that all students be proficient by 2014, and along the way, schools are judged according to whether or not all the proficiency rates specific to various subgroups of students exceed the same intermediate targets for proficiency rates. Second, the provisions of NCLB that specify how schools will be sanctioned under NCLB for failing to meet AYP deal with the performance of each school in isolation. The threat of sanctions is in no way influenced by whether or not a particular school is the only school in its state failing to meet AYP or whether it is one of several thousand schools serving similar students that are also failing to meet AYP. Finally, NCLB specifies sanctions for failure to meet AYP that are the same regardless of the distance between a failing school's proficiency rate and its AYP target.[8]

The following hypothetical scenario demonstrates how these features may soon interact in ways that create a credibility problem for NCLB. Consider a hypothetical school in a disadvantaged neighborhood in Chicago in 2017. Assume that 95% of the student population in this school is eligible for free or reduced price lunches and that 90% of the students are proficient in math and reading. Further, assume that this school has failed to make AYP for five straight years. Given these assumptions, it is easy to imagine that, among the dozens of schools in the Chicago area that serve populations of severely disadvantaged students, 90% would be one of the highest proficiency rates reported. If this is the case, would federal officials force the state to close our hypothetical school or reorganize it? Would federal officials even be so bold as to force the state to transfer resources from this school to private agencies that provide supplementary services? And, if federal officials were this bold, would it really make sense to take resources away from a school that performs better with their students than the vast majority of schools that serve similar students?

The sanctions in NCLB take the form of "perform to this absolute standard or else" with no consideration for how other schools that serve similar students are performing. However, elected officials run the risk of serious political consequences if they close a school that is performing as well or better than almost all schools that serve similar students because, in this circumstance, there is no reason to be confident that students in this school have better options elsewhere. Further, if hundreds of schools in a large city are facing the threat of reorganization at the same time, the leaders of these schools know that it is simply not logistically possible to sanction all of these schools at once.

The goal of accountability systems should be to make sure that students are learning as much as possible given their pre-school preparation, the support they receive from their families, and the resources available to their schools. For some students, "as much as possible" may fall short of proficiency, and for these students, the text of NCLB prescribes a chaotic cycle of school sanctions and reorganizations. Because such chaos will harm both children and educa-

---

[8]The Safe Harbor provisions in NCLB allow some schools to meet AYP by making su cient progress toward their current AYP target, but if the law is enforced as written, the 100% target will become relevant for all schools at some point in the future.

tors, I predict that the most severe sanctions spelled out in NCLB will never be enforced, and I also conjecture that educators understand this. In 2001, the proponents of NCLB may have believed that spelling out severe penalties for all schools that do not achieve 100% proficiency by 2014 was politically valuable. However, as we near 2014, this requirement may make the political costs of enforcing the law unbearable and thus render NCLB impotent as an accountability system.

Focus on Relative Performance

Federal accountability policies must shift their attention from proficiency to efficiency. There is no reason to believe that it would be socially efficient to devote the resources required to make every child proficient based on standard notions of proficiency. Further, the federal government does not have the capacity to compel states and localities to devote the resources required to make all children proficient. However, federal accountability plans based on competition rather than psychometric standards may compel state and local governments to use the resources they devote to education more efficiently.

Policy makers do not possess the information required to set efficient achievement targets for every student that vary appropriately with individual differences in background and prior achievement. The achievement levels we should expect from various types of students are constantly evolving because teaching techniques and our understanding of the ways that different children learn evolve over time. Thus, the best way for policy makers to promote efficiency is to build accountability systems around organized competitions among schools that reveal how achievement frontiers are evolving over time for different types of students. The measures of school performance that determine the winners and losers in such competitions are not measures of performance relative to some set of pyschometric standards but measures of performance relative to other schools that serve similar students.

The Percentile Performance Index described in Barlevy and Neal (2009) provides one tool for gathering information on how schools are doing relative to other schools that serve comparable students. Assume that for each grade*subject pair, policy makers can estimate $F(t|Z)$, the conditional distribution of a student's end of year assessment score, t, given Z, a vector of baseline characteristics that shape the costs of bringing the student up to various levels of achievement. When forming an estimator $\hat{F}(t|Z)$, policy makers may employ measures of past achievement, family resources, past achievement by peers, and other resource measures as elements of Z.[9]

Given $\hat{F}(t|Z)$ and individual results on an end of year assessment, education

---

[9]I will not comment here on the di erent methods that may be employed to produce estimators of $\hat{F}(t|Z)$ other than to note that estimating conditional distributions of this type will require large data sets and may require additional state-wide integration of data systems. Briggs and Betebenner (2009) develops an estimator for $\hat{F}(t|Z)$ by employing a series of quantile regression models and implement the estimator using data from Colorado schools.

officials can assign each student a percentile score $p = F^{\wedge -1}(t|Z)$ that denotes his or her percentile score among all students in the school system that share the same vector of baseline characteristics Z. The average value of p over all the students in a school is the PPI for that school with respect to a given subject*grade combination. One can calculate a school wide PPI by forming a weighted average of the PPI's for each subject*grade assessment. Thus, if a given school has a PPI of X, this implies that, roughly $1 - X$ percent of the time, students in other schools with similar baseline characteristics perform better than students in this school.

The optimal way to employ a PPI index when making decisions concerning which schools should be declared ineligible for government funding remains a matter for future research. However, if a school has a PPI around .10 for several years, we know that other educators are working with students elsewhere from similar backgrounds and, in the vast majority of cases, getting better results. This suggests that other educators may be both willing and able to take over this school and improve student achievement without requesting additional resources.[10]

Among educational statisticians, VAM models are the most commonly advocated tool for creating relative performance measures. VAM models do not specify the conditional distribution of assessment results but rather the conditional expectation of assessment scores. Let θ be the quality of a student's school. VAM models are variations on the following specification for the conditional expectation of t given Z.

$$E(t|Z) = \theta + g(Z)$$

VAM models produce a vector of relative performance measures, $\hat{\boldsymbol{\theta}}$, and the elements of this vector are estimates of each school's quality relative to some hypothetical school of average quality. Thus, if one is willing to treat the units of t as a social welfare index, VAM methods produce a universal quality ranking over all schools.

The problem is that universal quality rankings may provide answers to irrelevant questions. Assume that school A is a neighborhood school that serves disadvantaged students and that school B is a magnet school with selective admission. Assume that although most students would expect higher achievement in school B than school A, the students in school A would not. In this setting, VAM methods may produce results which imply that school B is better than school A, but why should this matter to the students in school A? School B is not better for them.

In addition, if there is little or no overlap in the baseline achievement distributions for schools A and B, the estimated difference in overall quality implied by VAM or any other method of creating universal rankings is to a large extent the creation of the researchers who developed a particular VAM specification

---

[10]Here, I assume that all schools in a district or state receive comparable resources for classroom activities. If this is not the case, PPI measures should compare schools that have access to similar classroom instruction budgets.

for assessment outcomes. The data provide no direct evidence concerning how these two schools perform with respect to a set of comparable students.

In contrast, the PPI approach described above is silent concerning universal performance rankings because it only reflects comparisons among students who were comparable at the beginning of the academic year. These comparisons alone do not allow policy makers to form a universal ranking of school quality. However, these comparisons may allow policy makers to identify schools that are clearly inefficient. Schools with persistently low PPI scores have students who consistently perform worse than their peers in other schools, and this is prima facie evidence that these schools are not using their resources efficiently.

Finally, because the PPI approach is based on ordinal comparisons, PPI results are invariant to the scales that testing agencies choose for reporting assessment results as long as different possible scales preserve order. In comparison, VAM measures of school performance are scale dependent, and these measures are invalid measures of quality unless one is willing to treat the units of a particular psychometric scale as a social welfare index. VAM models require that test scores or some transformation of test scores be reported in units such that, whenever a student gains one point, this gain is of equal value to society regardless of the student's prior level of skill. However, I know no reason that society should assume that a student whose math score moves from 120 to 122 creates the same gain for society as another student whose score moves from 220 to 222. On the contrary, it seems reasonable to suspect that, over certain regions of a given scale, score gains are both more valuable for students to achieve and more costly for teachers to foster, and in this case, the average test score gains that VAM methods associate with school quality are not valid measures of school performance because VAM methods weight all deviations from predicted score gains equally.

Conclusion

NCLB is flawed for many reasons, but the most important is the fact that NCLB is built around proficiency targets. Proficiency rates are not useful metrics of school performance because the variation in proficiency rates among schools, in large part, reflects inter-school differences in student background characteristics. The designers of accountability systems must move away from systems designed around psychometric standards and begin designing systems that organize and promote competition among schools.

Further, any accountability system that uses only student assessment results to judge outcomes in these competitions may fail to promote important and valuable dimensions of performance. A growing literature in economics stresses the importance of the non-cognitive skills that children acquire before going into the worlds of work and higher education.[11] Schools can and should play a role in helping parents encourage children to learn how to work well with others and to become more patient, persistent, and industrious.

---

[11]See Cunha and Heckman (2008).

This observation suggests that voucher systems and test-based accountability programs may be complementary components of a larger accountability system in which the amount of public funding available to a given school is determined by the number of parents who chose the school, the assessments results of students in the school, and results of inspections that assess the safety of the school's physical environment. Parents have access to valuable information about the social and emotional development of their children, and expansions of parental choice through vouchers or the expansion of charter schools allow government to acquire an army of educational performance monitors.

Some may argue that the optimal system involves providing parents with the information produced by student assessments and school inspections and then letting parents direct resources to efficient schools through a voucher system. However, more work remains concerning the design of these assessments and inspections as well as the design of the system used to communicate results to parents. Finally, an optimal scheme that involves vouchers must specifiy rules concerning whether or not schools can practice selective admission, whether or not the generosity of vouchers varies with family income and whether or not schools should be allowed to charge tuition in excess of the voucher level.[12]

To date, economists working in education research have made important contributions by conducting careful and often sophisticated empirical evaluations of existing accountability and incentive systems. However, the economics literature on the optimal design of these systems is much smaller, even though economists, at least relative to other social scientists who work on education policy, should have some expertise as designers of incentive systems. Given the flawed designs of NCLB and many similar state accountability systems, it may be time for economists to change their focus.

---

[12]See Neal (2009) for more on the design of incentive systems for publicly funded schools. See MacLeod and Urquiola (2009) for an analysis of interactions between voucher systems and national assessment systems.

References

[1] Barlevy, Gadi and Derek Neal. "Pay for Percentile," Federal Reserve Bank of Chicago Working Paper, #2009-09, November 2009.

[2] Booher-Jennings, Jennifer. "Below the Bubble: 'Educational Triage' and the Texas Accountability System," *American Educational Research Journal* 42:2 (Summer 2005), pp. 231-268.

[3] Briggs, Derek and Damian Betebenner, "Is Growth in Student Achievement Scale Dependent," mimeo. April 2009.

[4] Cronin, John; Dahlin, Michael; Adkins, Deborah and G. Gage Kingsbury. "The Proficiency Illusion," Thomas B. Fordham Institute, October 2007.

[5] Cunha, Flavio and James Heckman. "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Journal of Human Resources* 43 (Fall 2008): 739-780.

[6] Dee, Thomas S. and Brian Jacob. "The Impact of No Child Left Behind on Student Achievement," NBER Working Paper #15531, November 2009.

[7] Gillborn, David and Deborah Youdell. Rationing Education. Open University Press, Philadelphia, 2000.

[8] Illinois State Board of Education. Illinois State Report Card, 2003, 2006, and 2009.

[9] MacLeod, Bentley and Migual Urquiola. "Anti- Lemons: School Reputation and Educational Quality," NBER Working Paper #15112, June 2009.

[10] Neal, Derek. "The Role of Private Schools in Education Markets," in the *Handbook of Research on School Choice*, edited by Mark Berends, Matthew G. Springer, Dale Ballou, and Herbert J. Walberg, Lawrence Erlbaum Associates/ Taylor & Francis Group, 2009.

[11] Neal, Derek. "Designing Incentive Systems for Educators," forthcoming in *Performance Incentives: Their Growing Impact on American K-12 Education*, edited by Matthew Springer, Brookings. 2009.

[12] Reback, Randall. "Teaching to the Rating: School Accountability and the Distribution of Student Achievement," *Journal of Public Economics* 92:5-6 (May 2008), pp. 1394-1415.

[13] White, Katie Weits and Rosenbaum, James E. "Inside the Blackbox of Accountability: How High Stakes Accountability Alters School Culture and the Classification and Treatment of Students and Teachers," in *No Child Left Behind and the Reduction of the Achievement Gap: Sociological Perspectives on the Federal Education Policy*. Eds. A. Sadvonik, J. O'Day, G. Bohrnstedt, and K. Borman. Routledge, 2007.