THE HAMILTON PROJECT

Advancing Opportunity, Prosperity, and Growth

POLICY BRIEF 2011-09

New Assessments for Improved Accountability

SEPTEMBER 2011

WWW.HAMILTONPROJECT.ORG

The Hamilton Project seeks to advance America's promise of opportunity, prosperity, and growth.

We believe that today's increasingly competitive global economy demands public policy ideas commensurate with the challenges of the 21st Century. The Project's economic strategy reflects a judgment that long-term prosperity is best achieved by fostering economic growth and broad participation in that growth, by enhancing individual economic security, and by embracing a role for effective government in making needed public investments.

Our strategy calls for combining public investment, a secure social safety net, and fiscal discipline. In that framework, the Project puts forward innovative proposals from leading economic thinkers — based on credible evidence and experience, not ideology or doctrine — to introduce new and effective policy options into the national debate.

The Project is named after Alexander Hamilton, the nation's first Treasury Secretary, who laid the foundation for the modern American economy. Hamilton stood for sound fiscal policy, believed that broad-based opportunity for advancement would drive American economic growth, and recognized that "prudent aids and encouragements on the part of government" are necessary to enhance and guide market forces. The guiding principles of the Project remain consistent with these views.

The Hamilton Project Update

A periodic newsletter from The Hamilton Project is available for e-mail delivery.
Subscribe at www.hamiltonproject.org.

ADVISORY COUNCIL

New Assessments for Improved Accountability

Over the past decade, educational reforms have increased efforts to hold teachers and schools accountable for student test scores. Schools without significant progress on test scores have been subject to reductions in funding and even replacement of school leadership. The purpose of these actions is to increase student achievement by raising teacher effectiveness and bringing up the performance of low-performing schools. Yet critics of these accountability systems have argued that they will not lead to meaningful increases in student learning because of incentives to "teach to the test" at the expense of more valuable classroom activities, leading students to have deficits in critical thinking skills.

Based on work he has done for The Hamilton Project, Derek Neal of the University of Chicago outlines a plan to create better assessments and accountability systems to avoid these perverse incentives. The new assessment system would use two different styles of examinations: one traditional test to evaluate student achievement, and a new examination to evaluate teacher performance. Neal provides guidelines for the development of this innovative approach to assessment and details how teacher performance can be measured using a relative scale. An ideal accountability system would combine these new assessments with non-test metrics such as classroom observations, school inspections, and parental input in order to also capture students' social and emotional development.

The Challenge

Since No Child Left Behind (NCLB) was implemented in 2002, test-based accountability has become the norm for schools at the national level. Test-based accountability for individual teachers is also becoming increasingly popular as a result of the U.S. Department of Education's Race to the Top program. Race to the Top, launched in 2009, offers grants for states to develop ways to assess teachers based on growth in their students' test scores.

This increase in accountability has brought testing to the forefront of education policy. Incentives based on accountability systems have the potential to better focus schools and teachers on student achievement. Indeed, accountability systems may help educators identify the activities that contribute most to student achievement which is, after all, the very goal of these systems. For example, in a 2006 Hamilton Project discussion paper, Thomas Kane, Robert Gordon, and Douglas Staiger provided guidelines for improving teacher quality and student outcomes by using measures of teacher effectiveness in decisions about tenure and pay.[1]

However, using a flawed assessment system to collect data on educator performance can be counterproductive. In particular, Derek Neal highlights that some tests provide opportunities for coaching that does not contribute to learning. Teachers can coach their students or "teach to the test" by helping their students learn certain test formats or drilling them on questions from old tests. When this happens, students may disproportionately learn test-taking skills rather than learn critical thinking skills and concepts. An illustrative example of format-specific learning occurred when students in New Jersey were taught to answer math questions posed in the vertical format (so that the two numbers are stacked on top of each other) then performed poorly on another test that presented similar questions, but in a horizontal format (52 + 29 = ??).[2]

Neal argues that coaching is easy on the current high-stakes assessments because these tests are constructed to facilitate the consistent measurement of student achievement over time—tests are developed to answer the question of how much this year's sixth graders know relative to previous years of sixth graders. In attempts to answer such questions, test developers not only maintain constant formats for assessments, but also repeat questions from previous tests. By repeating questions, test developers create links between different tests that allow them, in theory, to scale scores consistently across years.

When schools or teachers are held accountable for scores on these types of tests, they will strive to do what they can to raise test scores. Because of the common items and repeated formats in current assessments, teachers can use the previous year's tests as a guide for the next year. In doing so they may not be teaching the content in a way that causes students to understand it in multiple contexts.

Coaching on test formats or drilling students on questions from old tests is an effective strategy for raising test scores, but it can also take time away from more useful teaching that helps students develop academically, emotionally, and socially. To be clear, this coaching is not necessarily ill-intentioned. Teachers may think that teaching to the test is the best way to improve student learning.

The result is that students often perform better on the high-stakes tests that are used in accountability systems but fail to show equal improvement in other measures. Many studies find that when assessment-based accountability systems are introduced, students perform increasingly well on the high-stakes tests, but show little or no improvement on similar assessments that are not part of the accountability system. This divergence suggests that accountability systems encourage educators to teach test-specific skills rather than helping students gain a critical understanding of the material.[3]

But not all the evidence is negative. Though improvement on high-stakes tests is not matched by gains on similar tests, there is evidence that schools make positive policy changes such as supplemental instruction for struggling kids, longer school days, and more time for teacher collaboration and planning.[4]

Still, the divergence in test scores supports Neal's argument that assessments that use repeated or predictable features are undesirable as the basis for high-stakes accountability systems.

A New Approach

Just as any private-sector employee is held accountable for the work that she performs, teachers should be held accountable for the quality of their teaching. Because teacher effectiveness affects the rate of student learning, one way to gauge the quality of teaching is to measure the progress of students. However, it is imperative to design a system that requires students to master the curriculum in ways that promote abilities to understand, apply, and communicate important ideas in multiple contexts.

Thus, Derek Neal proposes that policy-makers begin work on new assessment systems that are designed ex ante to be impervious to coaching. The goal is to develop assessments that measure true subject mastery, not test-specific skills, so that teachers who want to improve tests scores will teach the academic material well rather than teaching test formats or old questions.

New Assessments

The ideal assessment system envisioned by Neal would be predictable in content, but not in other ways. The content should reflect the curriculum that schools and states want students to learn. Thus, the first step in developing the assessment is to clearly define the set of skills and knowledge that teachers need to cultivate among students.

The next step is to design a test that is more immune to coaching. Neal suggests a few guidelines for such a test:

- **Do not repeat questions.** When questions are repeated from one year to the next teachers face strong incentives to drill students repeatedly on the exact set of questions used on previous tests.
- **Vary the formats.** As with the horizontal and vertical addition example mentioned above, there is clear evidence that when students learn material in one format, their command of the material can be limited to the format in question.
- **Avoid or limit multiple-choice questions.** Any scoring rule for multiple-choice questions includes penalties for incorrect answers. Then there is an optimal test-taking strategy that specifies when students should guess and when students should leave questions blank, which could lead teachers to spend class time on these strategies.

Linking Assessments to Accountability

Derek Neal has also proposed that the tests be used in a particular way. He does not advocate linking the gains of students in one class to individual teachers because it could be counterproductive to have teachers within the same school competing against each other rather than working together. Instead of holding a particular math teacher accountable for test scores in her class, Neal would hold all the math teachers in one grade accountable for improvement in the entire grade.

Given a set of assessments that lack repeated questions, it would be impossible to compare test results from different years or to set absolute goals or proficiency standards such as those that exist under NCLB. However, according to Neal, policy-makers can still build accountability systems. If policy-makers can reliably rank students according to year-to-year test score gains in particular subjects, they can use these ranks to create useful performance metrics without setting an absolute scale for the test scores. In other work, Gadi Barlevy and Derek Neal describe a performance metric for educators based solely on the ranks of their students on particular assessments.[5] Their metric is as follows:

- Consider all students in a large school district or state who are taking the same class, e.g., fifth-grade math. Group each student with other students who are similar in terms of their past performance, demographic characteristics, and the characteristics of their classmates and schoolmates.
- At the end of each year, rank all students in each group based on their end-of-year scores, and assign each student a percentile score based on the fraction of students in their group who performed the same or worse.
- Take the average of these percentile scores for a particular subject over all the students in one grade at one school. This average is the score for the teachers who teach a given subject in a particular grade in a specific school, and represents how often students in a given course in a given school perform as well or better than comparable students elsewhere.

This system differs from the vast majority of existing schemes in at least two important ways. First, this system does not require scaled assessments, and thus can be implemented based on a separate set of assessments designed to avoid coaching and teaching to particular test formats. Second, this system uses direct competition to create performance metrics. Importantly, teachers are only competing against other teachers who are in similar situations, so will not be unfairly blamed for circumstances that are beyond their control.

Evaluating teachers based on relative performance would also put the gains of all students on equal footing. The current cutoffs defined by NCLB encourage educators to focus attention on students who are near the proficiency threshold.

Non-Test Measures for Teacher Accountability

Although many conversations about teacher accountability focus on test scores, assessments will never be able to provide a full picture of teacher performance. Teaching academic knowledge, as measured by even the best tests, is only one part of what schools are expected to do. Schools also should be given credit for their contributions to the emotional and social development of their students as well as for their health and safety. To this end, classroom evaluation methods are being explored and developed by education experts in New York City, Washington DC, and other large cities.

To some extent, classroom evaluation systems seek to accomplish the same objectives as assessment-based accountability systems: evaluation systems often focus on the academic quality of classroom instruction. However, these two approaches differ conceptually in that evaluation systems also provide information about the attention that individual teachers are devoting to the noncognitive development of their students. In this way, classroom evaluations can complement information from assessments.

The persons who possess the best information about how teachers are performing in terms of promoting the social and emotional well-being of their students often are parents or guardians. Thus, a key design task for policy-makers is to figure out how to elicit accurate reports of the information that parents possess.

Challenges

Developing and maintaining new assessments would require a large upfront investment, and it is not obvious where the responsibility of creating these tests would fall. A couple of recent examples from Race to the Top provide some guidance. As a part of Race to the Top, the Department of Education awarded grants to two assessment consortia to develop new tests that will better measure critical thinking skills.[6] These two consortia demonstrate that the cost of developing and maintaining innovative assessments need not be large if it can be spread between states.

Roadmap

- Assessments that are used in accountability systems should be carefully constructed to avoid giving teachers the incentive to teach to the test. Tests should avoid repeated questions and should vary question formats so that teachers can raise scores only by teaching content.
- Accountability systems based on absolute standards or proficiency levels cannot be built on the new test types. Instead, teachers should be compared to similar teachers and rewarded or penalized based on their relative performance.
- Tests provide only a snapshot of one aspect of what teachers are expected to do. In measuring teacher performance, tests should be supplemented by classroom observation, parent surveys, and other non-test measures.
- The cost of developing and maintaining new assessments can be relatively low if it is divided among many states. Race to the Top provides a model for implementing the development of new tests.
- The new assessments would complement but not replace current standardized tests, which are still necessary to measure change in student achievement over time.

Learn More About This Proposal

This policy brief is based work done by:

DEREK NEAL
University of Chicago

Additional Hamilton Project Proposals

The Power and Pitfalls of Educational Incentives

There is widespread agreement that America's school system is in desperate need of reform, but many educational interventions are ineffective, expensive, or difficult to implement. Recent incentive programs, however, demonstrate that well-designed rewards to students can improve achievement at relatively low costs. This paper draws on school-based field experiments with student, teacher, and parent incentives to offer guidelines for designing successful education incentive programs. Incentives for inputs, such as doing homework or reading books, produced modest gains and might have positive returns on investment, and thus provide a promising direction for future programs. Additionally, this paper proposes recommendations for future incentive programs and concludes with guidelines for educators and policymakers to implement incentive programs based on the experiments' research findings and best practices.

Organizing Schools to Improve Student Achievement: Start Times, Grade Configurations, and Teacher Assignments

Education reform debates often center on expensive, politically controversial, and dramatic changes in policy. This has obscured an important direction for raising student performance — namely, reforms to school management and organization that make sure the "trains run on time" and improve administrative decisions that affect the instructional process. Such reforms may substantially increase student learning at modest cost. The paper discusses three reforms that evidence suggests have highly favorable benefit-cost ratios: later start times for older students, restructuring the stand-alone middle school, and ensuring teachers are assigned the grades and subjects in which they are most effective.

If one state were to ask assessment developers to take on the cost of developing and maintaining the tests, then the cost per pupil in the state would be large, but if more states participate, then some of the fixed costs can be shared between states. The two consortia requested grants of $150 million each for development, which becomes even more of a value when shared among twenty or more states. The expected cost to maintain and grade the tests is between $15 and $50 per pupil, depending largely on what grading mechanisms the states use.[7]

The other principal challenge is finding classroom time for the new assessments. Educators will still want tests that compare students in one year to students in the next year. That means students will still need to be tested on common items and formats. Many tests, such as the National Assessment of Educational Performance (NAEP), use this format. The assessments that Neal proposes would likely be layered on top of current assessments. Many already complain about the time devoted to testing, but it may be that tests that encourage excellent teaching will also be tests that provide valuable learning experiences for students. Given a commitment to avoid repeated questions and formats as well as a commitment to ask questions that probe subject mastery, both the exercise of taking such tests and participation in future sessions that review the correct answers to each year's test may provide valuable learning experiences.

Conclusion

Assessments are an increasingly important tool in determining the success of teachers and students in the public school system. The current assessments, however, create perverse incentives for teachers to coach their students on a specific test when they are used in accountability systems. We do not yet fully understand the extent of this challenge, but the divergence of scores between high- and low-stakes tests demonstrate that the learning gains on high-stakes tests may not be real improvements in student achievement. It is therefore important for policy-makers to focus on developing tests that more accurately measure student achievement and teacher effectiveness.

To create tests that allow schools to harness the power of accountability while avoiding the problem of teaching to the test, Neal describes a new type of assessment with innovative question formats that are not subject to coaching. Since these tests lack repeated items, accountability systems will not be able to set absolute standards but instead can compare the performance of teachers with similar students. By offering a better way to measure both teacher quality and student performance, Neal provides the important foundation for a new system of accountability.

Questions and Concerns

Will Race to the Top consortia develop ideal assessments for teacher-level accountability?

The consortia seek to develop question formats that provide a more accurate portrait of student critical thinking skills and knowledge. The questions will be largely open response, including essays, performance items that involve research, analysis, oral or written reports, and constructed-response math items. These are the sorts of assignments that many teachers, parents, and even students (if forced to admit it) would say contribute to the learning process. Some questions will be graded by computers, some by professional graders, and some by teachers who will be provided a rubric for guidance. The primary goal of these consortia is to develop assessments that can be scaled across years; thus, they do not completely address the issue of teaching to the test, but the research they are conducting into innovative question types can inform the development of assessments that address Neal's concerns.

Neal also argues that recent experience with the American Institute of CPA's Uniform CPA exam and the market for test preparation classes demonstrate that it is possible to effectively coach students for tests that include open-response questions or performance events if the test repeats questions and follows a fixed format. The Uniform CPA exam includes open-response questions and performance events, but it still created opportunities for test-specific coaching. This coaching may be more useful for students than the test-prep behaviors induced by multiple-choice tests with fixed formats and repeated questions, but the decline in pass rates after changes in test format demonstrates that the exam was not accurately measuring true knowledge. Neal suggests that the development of open-response and performance question types is not sufficient to address his concerns.[8]

Is there a precedent for accountability systems based on relative performance rather than absolute standards?

A key feature of Neal's system for performance accountability is that teams of teachers should be held accountable for gains relative to the gains achieved by teachers at comparable schools. Colorado, Indiana, and Massachusetts currently create metrics that compare student test score gains to their peers' gains, but these metrics were not designed for accountability systems. In 2009, eleven states had adopted these relative measures, called Student Growth Percentiles (SGP).[9] However, these metrics are not designed to address the issue of teaching to the test.

Endnotes

1 Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger , "Identifying Effective Teachers Using Performance on the Job," Hamilton Project Discussion Paper 2006-01 (2006, April), Washington, DC.

2 Lorrie Shepard, "Should Instruction be Measurement Driven? A Debate" (1988), http://nepc.colorado.edu/files/Shepard_ShouldInstructionBeMeasurement-Driven.pdf.

3 Daniel M. Koretz, "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity" *Journal of Human Resources* 27, no. 4 (2002): 752–777. For a full literature review on the effect of accountability systems, see Neal's "Providing Incentives for Educations," *Handbook of Economics of Education*, Vol. 4 (San Diego, CA: North Holland, 2011).

4 Hanley Chiang, "How Accountability Pressure on Failing Schools Affects Student Achievement, " *Journal of Public Economics* 93 no. 9–10 (2009): 1045–1057; Cecilia Elena Rouse, Jane Hannaway, Dan Goldhaber, and David Figlio, "Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure," Working Paper No. 13681, National Bureau of Economic Research, Cambridge, MA (2007).

5 Gadi Barlevy and Derek Neal, "Pay for Percentile," *American Economic Review* (forthcoming).

6 For more information on the two consortia, see their websites: The SMARTER Balanced Assessment Consortium (http://www.k12.wa.us/smarter/) and The Partnership for Assessment of Readiness for College and Careers (http://www.parcconline.org).

7 Numbers taken from the SMARTER Balanced Consortium RTT application (http://www.k12.wa.us/SMARTER/pubdocs/SBAC_Narrative.pdf) and the Partnership for Assessment of Readiness for College and Careers RTT application (http://www.parcconline.org/sites/parcc/files/PARCC%20Application%20-%20FINAL.pdf).

8 For more information on the Uniform CPA test, see http://www.aicpa.org.

9 D. W. Betebenner, "Norm- and Criterion-Referenced Student Growth," *Educational Measurement: Issues and Practice* 28, no. 4 (2009): 42–51.

Highlights

Derek Neal of the University of Chicago provides guidance for the development of a new assessment that can be used in accountability systems without creating incentives to teach to the test.

The Proposal

Assessments without repeated questions or consistent formats. Although the content of a test should be predictable, standardized tests with similar questions encourage teachers to teach to the test by coaching students on test formats or drilling them on questions from past tests. New assessments can avoid this problem by varying test formats and not repeating questions.

Accountability based on a holistic measure of teacher performance. Tests can provide only a snapshot of what teachers are expected to do. Other non-test measures such as classroom observation and parent surveys should also play a role. Accountability measures will be based on relative effectiveness rather than any absolute scores or proficiency standards.

A collaborative approach to test development. The testing consortia formed for Race to the Top demonstrate that the large upfront cost of test development and maintenance can be relatively low on a per pupil basis if it is shared between many states.

Benefits

The current accountability systems use assessments that encourage teaching to the test. In this system, tests do not accurately reflect students' critical thinking skills and knowledge, and classroom time is diverted away from real learning to test specific skills. New assessments would mitigate these two problems and ensure that future accountability measures are built on a solid foundation.

THE HAMILTON PROJECT

1775 Massachusetts Ave., NW
Washington, DC 20036

(202) 797-6279

BROOKINGS