U.S. Senate
Committee on Health, Education, Labor, and Pensions

Full Committee Hearing:
Fixing No Child Left Behind: Testing and Accountability

January 21, 2015

Written Statement of:

Dr. Martin R. West
Associate Professor of Education, Harvard Graduate School of Education
Deputy Director, Program on Education Policy and Governance, Harvard Kennedy School
Non-Resident Senior Fellow, Brookings Institution

Chairman Alexander, Senator Murray, Members of the Committee:

Thank you for the opportunity to appear before you today. My name is Martin West. I am an associate professor of education at the Harvard Graduate School of Education and a non-resident senior fellow at the Brookings Institution. Over the past fifteen years, I have conducted my own research on test-based accountability systems, reviewed the research of others working in this area, and consulted with state and federal policymakers on the design of accountability policies.

I would like to begin by congratulating the committee on its decision to put the reauthorization of the Elementary and Secondary Education Act at the top of its legislative agenda for the 114th Congress. Nothing is more important to our nation's future than ensuring that all American children have the opportunity to reach their full academic potential. Congress cannot do that on its own, but it can help by addressing the very real shortcomings of the most recent reauthorization, No Child Left Behind (NCLB), and restoring the predictability with respect to federal policy that state and local officials need to carry out their work.

My testimony aims to inform this effort by providing information on:

1. the validity of test scores as measures of student learning;
2. the effects of NCLB's testing and accountability requirements, both overall and in schools identified as in need of improvement; and
3. the implications of eliminating the law's annual testing requirements.

I conclude with recommendations on how to address NCLB's most serious flaws while building on its most important contribution: the provision of far greater transparency about the academic achievement of American students. The law's requirement that students be tested annually in math and reading in grades 3-8 and once in high school, and that the results be reported by school and disaggregated by student subgroup, has provided parents, teachers, and other citizens with essential information about students' performance in these foundational subjects – and therefore the extent to which they have mastered skills that are prerequisites for other educational goals. This information has called attention to gaps in achievement along lines of race, ethnicity, and class across entire states and within specific schools; it has ushered in a new era in education

research; and it has made it possible to develop new indicators of schools' performance based on their contribution to student learning.

My principal recommendation is therefore to maintain the law's current requirement that states test students annually in math and reading in grades 3-8 and at least once in high school, while restoring to states virtually all decisions about the design of their accountability systems, including how schools and teachers are identified as under-performing and what should be done to improve their performance.[1]

The validity of test scores as measures of student learning

Test-based accountability is premised on the notion that student test scores in core academic subjects are valid indicators of student learning that matters for valued long-term outcomes. That is, policymakers are generally not interested in boosting math and reading test scores per se, but only insofar as those test scores predict outcomes such as post-secondary success and adult earnings. It has long been known that student performance on low-stakes tests are strong predictors of individual labor-market success and, in the aggregate, of national economic growth rates (see, e.g. Johnson and Neal 1996; Hanushek and Woessmann 2008). Only recently, however, have researchers been able to examine the predictive power of the kinds of tests administered by states and school districts in low- and high-stakes settings. Two new studies in this area are particularly instructive.

Chetty, Friedman, and Rockoff (2014) link math and reading test scores from New York City students in grades 3-8 to Internal Revenue Service records for the same students as they became adults. The tests were administered on an annual basis to all New York City students but were not at the time used to hold teachers or schools accountable for their performance. The study shows that being assigned to a more effective teacher, as defined by her past impact on other students' test scores, has a positive impact on the likelihood that students will attend college and on their earnings at age 28, as well as on other variables such as the likelihood of avoiding teenage pregnancy (for girls) and whether the student is saving for retirement. The magnitude of these relationships is impressive. For example, being assigned to a teacher in the top-5% in terms of her success in raising student test scores, as opposed to an average teacher, increases a child's lifetime income by roughly $80,000. These results clearly highlight the importance of teacher effectiveness in shaping student outcomes. As important, they confirm that a teacher's impact on test scores accurately predicts her impact on more distant outcomes, at least when those tests are low-stakes.

Deming et al (2014), in turn, use data from Texas to examine the predictive validity of gains in test scores induced by the state's high-stakes school accountability system. They find that high schools responded to the threat of being assigned a very low rating by increasing their students' achievement on high-stakes tests. At age 25, these same students were more likely to have completed a four-year degree and have higher earnings. Deming et al also find that schools pressured to earn a higher rating responded not by improving achievement, but by classifying more low-scoring students as students with disabilities in order to exempt them from the

[1] My testimony draws in part on research conducted jointly with Matthew Chingos, Mark Dynarski, and Russ Whitehurst of the Brookings Institution and published in Chingos and West (2015) and Whitehurst et al (2015).

accountability system; these same students suffered large declines in their post-secondary attainment and earnings. Overall, their results therefore illustrate not only the predictive validity of accountability-induced gains in student test scores, but also the critical importance of designing accountability systems carefully in order to avoid unintended consequences.

In sum, evidence confirms that the scores that students receive on standardized tests administered in schools are strongly predictive of later life outcomes that are of great value to those students and the nation. Moreover, gains in test scores that result from interventions such as being assigned to a particularly effective teacher or attending a school facing accountability pressure also predict improvements in adult outcomes. Of course, teachers and schools also contribute to student outcomes in ways that are not captured by test scores and therefore harder to measure (see, e.g. Jackson 2012). However, information on school performance that does not include data on student learning as measured by tests that are comparable statewide would be badly compromised.

The effects of No Child Left Behind's testing and accountability requirements

With the No Child Left Behind Act of 2002 (NCLB), Congress for the first time required states accepting federal funds through Title I to put into place consequential test-based accountability systems. In particular, states were required to adopt challenging content standards in math, reading, and science; test students annually in math and reading in grades 3-8 and once in high school; report the share of students performing at proficient levels in each subject (disaggregated by student subgroup); and intervene in schools where students overall or within a specific subgroup failed to exceed statewide performance targets. States had to raise these targets over time in line with the goal of having all students achieving at proficient levels in core academic subjects by 2014.

Evaluating the impact of NCLB's testing and accountability requirements is difficult, as the law required all states to implement the same basic policies. Although student achievement in grades 4 and 8 as measured by the National Assessment of Educational Progress (NAEP) has risen since the law's enactment, this trend could be driven by other factors. NCLB did not come from nowhere, however: Roughly half of states had established consequential test-based accountability systems in the 1990s, and many NCLB requirements were based on elements of those systems. Insight into the law's effects can therefore be gleaned by comparing the states required by NCLB to implement test-based accountability systems for the first time to those that already had such systems in place and were thus less affected by the law's requirements.

Taking this approach, two independent teams of scholars (Dee and Jacob 2010; Wong, Cook, and Steiner 2011) find that NCLB's testing and accountability provisions have generated modest improvements in student achievement as measured by the NAEP in states required to implement test-based accountability systems for the first time. The gains in achievement have been concentrated in mathematics, as opposed to reading, and among the low-performing students that were most directly affected by the law's accountability system. These findings are consistent with earlier research indicating that states adopting consequential test-based accountability systems in the 1990s improved more on the NAEP than did other states (Carnoy and Loeb 2002; Hanushek and Raymond 2005). Both sets of studies are noteworthy in that they document gains

on the low-stakes NAEP and therefore should not be influenced by schools facing pressure to improve students' performance on a specific test.

A second approach to examining NCLB's effects has been to study the law's effects on schools at risk of being identified by states as in need of improvement and on schools so identified and subjected to various interventions. In the only nationally representative study of this kind, Reback, Rockoff, and Schwartz (2014) find that attending a school at risk of being identified as in need of improvement had positive or neutral effects on students' achievement on low-stakes reading tests, no clear effects on their achievement on low-stakes math and science tests, and positive effects on their enjoyment of learning in those subjects. Several studies conducted in specific states or school districts have also found that students enrolled in schools not making Adequate Yearly Progress (and therefore placed at risk of sanction) made greater than expected gains on their state test (see, e.g., Springer 2008; Krieg 2008; Ladd and Lauen 2010; Neal and Schanzenbach 2010; Hemelt 2011). Neal and Schanzenbach (2010) and Krieg (2008) find that these improvements were concentrated among students on the margin of proficiency – so called "bubble kids" (Booher-Jennings 2005) – suggesting that schools may have shifted their instructional energies away from students performing at much higher or lower levels. Studies of the effects of actual sanctions for under-performing schools required under NCLB provide a more mixed picture. Anh and Vigdor (2014), however, find positive effects on student achievement in schools forced into restructuring with leadership or management changes.

In sum, the best available evidence indicates that NCLB has generated improvements in student learning, concentrated in math, among the nation's lowest-performing students – precisely those on whom the law was focused. These gains have been relatively modest in size, however, far short of the rate of improvement required to bring all students to a reasonable definition of proficiency by 2014.

As the law's deadline for universal proficiency approached, the NCLB accountability system therefore became unworkable, with a majority of schools in some states identified as under-performing. In response, the Obama administration, through its ESEA Flexibility Program, offered states limited flexibility with respect to the design of their accountability system in exchange for complying with new requirements in areas such as teacher evaluation and school turnaround models. While the appropriateness and aspects of the design of this state waiver program are hotly debated, the acute need to address the shortcomings of NCLB's accountability model is not in dispute.

It is also important to acknowledge evidence of the unintended consequences of the NCLB accountability system. For example, research has clearly shown that test-based accountability can result in a narrowing of the curriculum to focus on tested subjects at the expense of those for which schools are not held accountable. Consistent with this, the initial implementation of NCLB was associated with large increases in the amount of instructional time elementary school teachers reported spending on reading and declines in the coverage of history and science (West 2007). Harder to track systematically is the law's effects on other aspects of classroom practice. Yet some evidence suggests that heavy handed test-based accountability policies can promote rote, teacher-directed instruction and encourage schools to focus narrowly on test-preparation skills rather than ensuring that students are exposed to a curriculum rich in academic content.

These tendencies may be strongest in schools with high minority and low-income populations, which typically face the strongest pressure to improve (Diamond and Spillane 2005).

Important concerns have also been raised about the amount of time students now spend taking standardized tests. Unfortunately we lack systematic data on the amount of time students nationwide spend taking standardized tests and how this changed with the implementation of NCLB and related federal policies. Nor do we know the amount of test-taking time that would be optimal. A handful of recent district- and state-level analyses suggest that students are scheduled to spend 1-3% of the school year taking standardized tests, depending on the grade level, a figure that sounds appropriate given the value of the information they provide and evidence that taking tests can support learning (Lazarin 2014; Teoh et al 2014; Nelson 2013; Ohio Department of Education 2015). That said, we also know that these official figures likely understate the true amount of instructional time teachers lose as a result of testing, that schools in some districts test much more than these averages, and that far too many schools devote excessive time to narrow test-preparation activities in an attempt to avoid federally mandated sanctions. The concerns now being voiced by parents and educators in these situations are legitimate.

Implications of eliminating annual testing requirements

Eliminating annual testing requirements is not necessary to reduce over-testing where it exists, however. Indeed, doing so would only make it harder for states to address the flaws of the NCLB accountability system and develop new ones that provide good information on schools' contribution to student learning and set realistic targets for improvement. It would also have other important negative consequences.

Eliminating annual testing is unnecessary because the annual tests in math and reading (and grade-span testing in science) currently required under NCLB typically account for less than half of the total amount of time students spend taking standardized tests. For example, a recent testing audit conducted by the Ohio Department of Education (2015) found that NCLB-mandated tests are responsible for 32 percent of testing time in that state. Another 26 percent of testing time is devoted to new assessments developed to implement a teacher evaluation system the state adopted as a condition of receiving a waiver through the Obama administration's ESEA Flexibility Program. The remaining 42 percent of testing time is devoted to tests required not by the federal government, but by the state or local school districts.

The most important flaw of the accountability system states are required to use under No Child Left Behind is its exclusive reliance on student performance levels as a measure of school performance. Under that system, whether a school makes Adequate Yearly Progress is determined primarily based on the share of students who are proficient in math and reading in a given year—a level-based measure of student achievement. Yet the level at which students perform at a given point of time is a poor indicator of school quality, as student achievement is heavily influenced by factors outside of a school's control. Measures based on the amount students learn from one year to the next can provide a more accurate gauge of schools' contribution to student learning (Deming 2014). These kinds of measures are only possible, however, when students are tested in adjacent grades.

In a recent analysis (Chingos and West 2015), Matthew Chingos of the Brookings Institution and I used roughly a decade of student test scores from all public elementary schools in North Carolina and Florida to compare how schools would look if they were judged based only on their average test scores in a single grade—as might be the case under a grade-span testing regime—to how they can be judged using measures based on year-to-year growth in student test scores. The analysis yielded two important conclusions.

First, growth measures do a far better job of identifying the schools that contribute the least to student learning. For example, North Carolina students in the bottom-15% schools in terms of average scores learn only about a third of a year less in math than the statewide average, whereas the difference for students in the bottom-15% of schools in terms of growth is more than half a year of learning.

Second, judging schools based on test score levels has a punishing effect on schools serving disadvantaged students, which are often identified as underperforming even when their students are learning more than students elsewhere. For example, 56 percent of North Carolina schools serving predominantly low-income students would be classified as bottom-15% based on their average scores, whereas only 16 percent would be labelled as such based on their growth. Accountability based on grade-span testing judges schools based on the students they serve, not how well they serve them.

Using average test scores from a single year to judge school quality is therefore unacceptable from a fairness and equity perspective. One possible alternative to growth-based measures is to use a single year of test data, such as would be available under a grade-span testing regime, but adjust it based on student demographics. In other words, schools serving students who tend to score lower, such as low-income and minority students, would be compared to schools serving similar student bodies rather than all schools in the state. Using demographic adjustments is an unsatisfying alternative for at least two reasons, however. First, it provides less accurate information about schools' contribution to student learning. Second, making demographic adjustments implicitly sets lower expectations for some groups of students than for others.

In addition to preventing the development of better and fairer measures of school performance, eliminating annual testing would have other negative consequences:

First, it would all but eliminate school-level information about the learning of student subgroups, as testing only a single grade within each school often results in sample sizes for groups such as English learners or blacks that are too small to generate reliable information for the school as a whole (Whitehurst and Lindquist 2012).

Second, it would sharply limit the information available to parents making choices about the school their child attends, whether through open-enrollment programs in traditional public schools or under charter school programs. School choice is empty without valid information on school performance, and how much schools contribute to student learning is the most important information parents need to know.

Third, it would prevent policymakers and researchers from evaluating the effectiveness of new education programs when, as is typically the case, the appropriate research design depends on knowledge of students' recent achievement. By hampering our ability to learn about what's working, jettisoning annual testing could slow the overall rate of improvement in student achievement over time.

A key reason Congress in 2002 required that states use a school accountability system based on student achievement levels was that many states were not yet testing students annually and those that did often lacked the capacity to track the performance of individual students over time. That situation has now changed, thanks to No Child Left Behind and related federal investments in state data systems. It would be ironic and, in my view, unfortunate if, in seeking to fix No Child Left Behind, Congress were to recreate the conditions that led to the adoption of an ill-designed accountability system in the first place.

Recommendations

1. Maintain the law's requirement that states test all students annually in math and reading in grades 3-8 and at least once in high school using tests that are comparable statewide.

The federal government has a critical role to play in ensuring that parents and citizens in states accepting federal funds have good information about their local schools' performance, and good information requires the data that come from annual testing using assessments that are comparable statewide. States should continue to be required to gather this information and to report on it disaggregated by student subgroup as a condition of receiving Title I funds.

To ensure that this requirement does not interfere with the ability of states to develop new forms of assessment, including competency-based assessments that are not tied to a specific grade level and are administered at varying times during the school year, Congress may wish to consider developing a pilot program for a small number of states doing innovative work in this area. However, such a pilot should be designed so as to provide rigorous evidence as to how the information it generates compares to that generated under an annual testing regime.

2. Return to states virtually all decisions about the design of their accountability systems, including how schools and teachers are identified as under-performing and what should be done to improve their performance.

The federal government lacks the capacity to design a single accountability system that is appropriate to the needs of each state, and has a poor track record when attempting to dictate the required elements of efforts to improve under-performing schools. States should be required to develop their own systems of school accountability and improvement, provided only that those systems are based in part on student achievement data from tests that are comparable statewide and, in the case of high schools, four-year adjusted cohort graduation rates. Federal accountability requirements, if included, should be limited to schools that fail at basic functions, i.e., elementary and middle schools in which a significant percentage of students do not acquire even basic competencies in reading and math, or high schools where a significant percentage of students do not graduate.

3. Require the publication of timely, accurate school-level spending data.

Consistent with the federal role in increasing transparency about educational performance, Congress should condition the receipt of Title I funds by schools and districts on the timely disclosure of comparable measures of per-pupil spending at the level of the state, district, and school. This recommendation, which is included in the current discussion draft, would build on the school-level expenditure reporting mandated as a one-time requirement under the American Reinvestment and Recovery Act of 2009 and would improve the accuracy and facilitate the broader dissemination of that information. By requiring that school spending reports reflect actual teacher salaries rather than district-wide salary averages (the common practice in district financial reporting), it could serve to highlight within-district disparities in spending and create pressure on school districts to address them. It would also permit the generation of performance measures that provide information on the relative return-on-investment for educational spending across districts and schools.

4. Continue to require that states participate in the National Assessment of Educational Progress (NAEP) exams administered biannually.

With states continuing to select their own academic standards and tests and provided with new flexibility with respect to the design of their accountability systems, the NAEP will continue to serve as an essential audit of the performance of state educational systems, enabling advocacy organizations and ordinary citizens to push for improvement. A "Secretary's Report Card" to Congress and the public on the educational performance of the nation and each state, as proposed in the current discussion draft, is an attractive new mechanism for heightening competition among states to lift all students to high levels of achievement.

References

Anh, T. & Vigdor, J. 2014. The Impact of No Child Left Behind's Accountability Sanctions on School Performance: Regression Discontinuity Evidence from North Carolina. NBER Working Paper No. 20511. Cambridge, MA: National Bureau of Economic Research.

Booher-Jennings, J. 2005. Below the bubble: "Educational triage" and the Texas Accountability System. *American Educational Research Journal*, 42(2): 231-268.

Carnoy, M. & Loeb, S. 2002. Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24(4): 305–331.

Chetty, R., Friedman, J. N. & Rockoff, J. E., 2014. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9): 2633–2679.

Chingos, M. M. & West, M. R. 2014. Why annual statewide testing is critical to judging school quality. The Brown Center Chalkboard, Brookings Institution.

Dee, T. S. & Jacob, B. 2011. The Impact of No Child Left Behind on Student Achievement. *Journal of Policy Analysis and Management*, 30(3): 418–446.

Deming, D., 2014. Using School Choice Lotteries to Test Measures of School Effectiveness. *American Economic Review Papers & Proceedings*, 104(5): 406-411.

Deming, D., Cohodes, S., Jennings, J., & Jencks, C. 2013. School Accountability, Postsecondary Attainment, and Earnings. NBER Working Paper 19444. Cambridge, MA: National Bureau of Economic Research.

Deming, D., Kane, T., Hastings, J. & Staiger, D. 2014. School Choice, School Quality and Postsecondary Attainment. *American Economic Review*, 104(3): 991-1014

Diamond, J., & Spillane, J. 2004. High-stakes accountability in urban elementary schools: challenging or reproducing inequality? *The Teachers College Record*, 106(6): 1145-1176.

Hanushek, E. A. & Raymond, M. E. 2005. Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, 24(2): 297–327.

Hanushek, E. A., & Woessmann, L. 2008. The role of cognitive skills in economic development. *Journal of Economic Literature*, 46(3): 607-668.

Hemelt, S. W. 2011. Performance Effects of Failure to Make Adequate Yearly Progress (AYP): Evidence from a Regression Discontinuity Framework. *Economics of Education Review*, 30(4): 702–23.

Jackson, C. K. 2012. Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina. NBER Working Paper No. 18624. Cambridge, MA: National Bureau of Economic Research.

Krieg, J. M. 2008. Are Students Left Behind? The Distributional Effects of No Child Left Behind. *Education Finance and Policy*, 3(2): 250–81.

Ladd, H. F. & Lauen, D. L., 2010. Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management*, 29(3): 426–450.

Lazarin, M. 2014. Testing Overload in America's Schools. Washington, DC: Center for American Progress.

Neal, D. & Schanzenbach, D. W. 2010. Left Behind by Design: Proficiency Counts and Test-Based Accountability. *Review of Economics and Statistics*, 92(2): 263–283.

Nelson, H. 2013. Testing More, Teaching Less: What America's Obsession with Student Testing Costs in Money and Lost Instructional Time. Washington, DC: American Federation of Teachers.

Ohio Department of Education. 2015. Testing Report and Recommendations.

Reback, R., Rockoff, J. & Schwartz, H. L. 2014. Under pressure: Job security, resource allocation, and productivity in schools under NCLB, *American Economic Journal: Economic Policy*, 6(3): 207-241.

Springer, M. G. 2008. The Influence of an NCLB Accountability Plan on the Distribution of Student Test Score Gains. *Economics of Education Review* 27 (5): 556–63.

Teoh, M., Coggins, C., Guan, C., & Hiler, T. 2014. The Student and the Stopwatch: How much Time to American Students Spend on Testing? Washington, DC: Teach Plus.

West, M. R. 2007. Testing, learning, and teaching: The effects of test-based accountability on student achievement and instructional time in core academic subjects. In Finn Jr., C. E. & Ravitch, D., eds. *Beyond the Basics: Achieving a Liberal Education for All Children*. Washington, DC: Thomas B. Fordham Institute.

Whitehurst, G. & Lindquist K. 2014. Test More, Not Less. Brown Center Chalkboard, Brookings Institution.

Whitehurst, G., West, M. R., Chingos, M. M., & Dynarski, M. 2015. The case for annual testing. The Brown Center Chalkboard, Brookings Institution.

Wong, M., Cook, T. D., & Steiner, P. M. 2009. No child left behind: an interim evaluation of its effects on learning using two interrupted time series each with its own non-equivalent comparison series. Institute for Policy Research Working Paper 09-11. Northwestern University.