Intergovernmental Approaches for Strengthening K-12 Accountability Systems

Edited Transcript

**Of a Symposium convened October 29, 2007, in Chicago, Illinois,
by The Nelson A. Rockefeller Institute of Government
with the support of the
Spencer Foundation and the Joyce Foundation**

and

A Framework Paper Circulated in Preparation for the Symposium

Intergovernmental Approaches for Strengthening K-12 Accountability Systems

Edited Transcript

Of a Symposium convened October 29, 2007, in Chicago, Illinois,
by The Nelson A. Rockefeller Institute of Government
with the support of the
Spencer Foundation and the Joyce Foundation

Edited by
Allison Armour-Garb

and

A Framework Paper Circulated in Preparation for the Symposium

The Nelson A. Rockefeller
Institute of Government

The research reported in this report was made possible in part by a grant from the Spencer Foundation and the Joyce Foundation. The views expressed are those of the authors and do not necessaily reflect the views of the funders.

Photos by Barbara Stubblebine

The Nelson A. Rockefeller Institute of Government
411 State Street
Albany, New York 12203-1003
www.rockinst.org

Table of Contents

About the Symposium

Six years into the implementation of the No Child Left Behind law (NCLB), many state governments lack access to, and the resources to pay for, expertise needed to implement K-12 educational accountability systems. Moreover, policymakers, educators, and testing companies face incentives to cut corners, lower standards, and game the system, and the public lacks clear ideas about the effectiveness of K-12 education because standards and measures of performance vary from state to state and between states and the National Assessment of Educational Progress.

Some stakeholders and experts argue that the federal government should address these problems by establishing national standards and tests, but state governments and other stakeholders and experts are leery of federal control. In the meantime, a number of states are collaborating to develop and use common standards and improve testing systems.

In October 2007, the Rockefeller Institute of Government convened forty state and federal education officials, testing experts, and educational researchers and education policy experts at a symposium in Chicago on prospects and possibilities for developing new intergovernmental approaches for K-12 standards and test-based accountability. The symposium was supported by the Spencer and Joyce Foundations.

Reactions to the discussion were positive. The morning panel was devoted to possible institutional models for setting K-12 standards. The lead speakers were Michael Cohen, president of Achieve, Inc., and Chester E. Finn, Jr., president of the Thomas B. Fordham Foundation. This section was moderated by *Education Week*'s Lynn Olson. The afternoon panel was devoted to intergovernmental approaches to the oversight of testing. The lead speakers were Robert Linn of the National Center for Research on Evaluation, Standards, & Student Testing, and Thomas Toch of Education Sector. The moderator was John Merrow of Learning Matters.

A framework paper for the symposium by Allison Armour-Garb on intergovernmental collaboration, possible functions for an intergovernmental collaborative entity, and institutional alternatives was circulated in advance of the symposium to participants and is included in this volume. An article about the symposium, entitled "Intergovernmental Approaches to Standards and Assessments," by Lynn Olson, will be published in the *Rockefeller Institute Bulletin*, and Allison Armour-Garb and Richard Nathan are preparing a final report on this project.

In light of the 2007 extension of NCLB without amendments, there is a window of opportunity now for interested parties to consider options, develop coalitions, and work on new intergovernmental approaches to strengthen educational accountability in the post-NCLB era.

List of Participants

David M. Abrams
Assistant Commissioner for Standards, Assessment and Reporting
New York State Education Department

Gordon M. Ambach
International Association for the Evaluation of Educational Achievement
Former Executive Director, CCSSO
Former NYS Commissioner of Education

Allison Armour-Garb
Director, Education Studies
Nelson A. Rockefeller Institute of Government

Eva Baker
Co-Director
National Center for Research on Evaluation, Standards, and Student Testing

John H. Bishop
Associate Professor of Human Resource Studies
Cornell University

Robert L. Brennan
Director, Center for Advanced Studies in Measurement and Assessment (CASMA)
University of Iowa

Wayne Camara
Vice President for Research and Development
College Board

Mitchell D. Chester
Senior Associate Superintendent for Policy and Accountability
Ohio Department of Education

Michael Cohen
President
Achieve

Stephanie Dean
Program Coordinator
James B. Hunt, Jr. Institute for Educational Leadership and Policy

Chester E. Finn, Jr.
President
Thomas B. Fordham Foundation

Anne Fitzpatrick
Senior Researcher, Educational Testing Service, and President, National Council on Measurement in Education

Margaret Goertz
Co-Director
Consortium for Policy Research in Education

Bryon Gordon
Assistant Director for Education, Workforce, and Income Security Issues
U.S. Government Accountability Office

Paul Goren
Vice President
The Spencer Foundation

Laura Hamilton
Senior Behavioral Scientist
RAND Corporation

Lisa Hansel
Managing Editor
American Educator

William G. Harris
Executive Director
Association of Test Publishers

Judith Anderson Koenig
Senior Program Officer, Board on Testing and Assessment
National Research Council

Robert Linn
Distinguished Professor Emeritus of Education
University of Colorado

Bethany Little
Vice President, Policy and Federal Advocacy
Alliance for Excellent Education

John Luczak
Education Program Officer
The Joyce Foundation

Scott Marion
Associate Director
National Center for the Improvement of Educational Assessment

Lorraine McDonnell
Professor of Political Science
University of California at Santa Barbara

Michael McPherson
President
The Spencer Foundation

John Merrow
Executive Producer/Host and President
Learning Matters, Inc.

Doug Mesecar
Acting Assistant Secretary
Office of Planning, Evaluation and Policy Development
US Department of Education

Richard P. Nathan
Co-Director
Nelson A. Rockefeller Institute of Government

Lynn Olson
Editor
Education Week

Cynthia Schmeiser
President and Chief Operating Officer
Education Division, ACT

David F. Shaffer
President, Public Policy Institute
NYS Business Council

Theresa Siskind
Deputy Superintendent, Division of Accountability
South Carolina Department of Education

Thomas Toch
Co-Director
Education Sector

Susan Traiman
Director of Education & Workforce Policy
Business Roundtable

Robert Ward
Deputy Director
Nelson A. Rockefeller Institute of Government

Phoebe Winter
Independent Consultant

Lauress Wise
President
Human Resources Research Organization

List of Acronyms and Abbreviations

ACT	American College Testing Program
ADP	American Diploma Project
AP	Advanced Placement
AYP	Adequate Yearly Progress
CCSSO	Council of Chief State School Officers
CRESST	Center for Research on Evaluation, Standards & Student Testing
DMV	Department of Motor Vehicles
ECS	Education Commission of the States
ED in '08	Strong American Schools campaign
ETS	Educational Testing Service
FDA	Food and Drug Administration
GED	General Educational Development Tests
IASA	Improving America's Schools Act of 1994
IB	International Baccalaureate
IES	Institute of Education Sciences
IRT	Item Response Theory
MDRC	Manpower Demonstration Research Corporation
NAEP	National Assessment of Educational Progress
NAGB	National Assessment Governing Board
NCLB	No Child Left Behind Act
NCME	National Council on Measurement in Education
NCTM	National Council of Teachers of Mathematics
NECAP	New England Common Assessment Program
NESIC	National Education Standards and Improvement Council
NIST	National Institute of Standards and Technology
PISA	Program for International Student Assessment
RFP	Request for Proposals
SAT	Scholastic Assessment Test
TAC	Technical Advisory Committees
Test Standards	*Standards for Educations and Psychological Testing*
TIMSS	Trends in International Mathematics and Sciences Study

Intergovernmental Approaches for Strengthening K-12 Accountability Systems

Edited Transcript

Of a Symposium convened October 29, 2007, in Chicago, Illinois,
by The Nelson A. Rockefeller Institute of Government
with the support of the
Spencer Foundation and the Joyce Foundation

Edited by
Allison Armour-Garb

Transcript Table of Contents

Introductions

Richard Nathan, Rockefeller Institute of Government

Good morning. My name is Dick Nathan and I am the co-director of the Rockefeller Institute of Government.

Just a couple of introductions: Allison Armour-Garb is the Rockefeller Institute's lead person for education and wrote the framework paper that we sent out in advance. She did all the work to gather this group. The idea of her paper (and she'll say a little bit about it) was not to take a position, but to suggest ways that this conversation might go and to lay out some ideas and precedents that would be useful to consider.

Richard P. Nathan

I'm going to say a word about the Rockefeller Institute and introduce some of our people, then talk about the purpose and the procedure for today's conversations.

About the Rockefeller Institute: We are the public policy research arm — call us a think tank — of the New York State University system. We report to the trustees of the 64 campuses in the New York State system. We have our own buildings and our own campus. We do not offer courses. We work with many faculty members and many graduate students. Our notion of the world is to educate, not advocate. We've got enough politicians in the world.

We particularly specialize in federalism, state government, intergovernmental relations, public finance, and public management. If you want one word to think about what we see as our niche, it's "states." States are very important players in American domestic government. That is my longstanding, strong personal interest.

Now to say a word about the people here: As I said, Allison is our director of education studies. We want to do more in this field, because there is so much related to federalism and states. Bob Ward is our deputy director. David Shaffer, who is the president of the Public Policy Institute of the Business Council of New York State, is helping us. Barbara Stubblebine is the person who organized this meeting, working with Allison.

The format is conversational. The purpose is to come from this with something useful that will be valuable to you, and we'll see what it means for us. We have a strong interest and desire to learn.

We have two moderators: Lynn Olson and John Merrow. Two panelists will sit up here and start us out. Allison came in one day and she said, "I'm going to say no Power- Points." That's a good idea. No talking heads and no PowerPoints. There will be a hundred good ideas that we'll try to gather and extract from what you have to say.

I've been talking to Mike McPherson about what we might bring to this and (particularly with work in the field, in the states) what it might be useful for us to do. For quite a long time he was our neighbor at Williams College. We worked together on some committees, particularly the National Academy of Science. We came up with this idea of how to start out, how we could be helpful and also learn a lot. I think it is a good thing. Mike is going to say a few words to the group and I'm going to ask Allison to say a little bit about her paper. If there are any questions about how we're doing this, you can ask those questions of her. Then we're going to take a break and bring up Lynn and Mike Cohen and Checker Finn, who will be talking to us and leading the discussion.

At the Rockefeller Institute, I sometimes make the remark that someday we're going to have a conference that is all coffee breaks. People who like to drink coffee have a lot to say to each other. Lynn and John Merrow will wink and nod and see when it would be a good time to have a chance to break and talk amongst yourselves. We see this as your day, so don't hesitate to say to one of the panel leaders or to me or Allison what you think about where the conversation should go.

At the end, we've reserved an hour for a discussion of the implications of what happens today. I think of our role as institution building and institutional learning. What is it that we should talk about under that heading? I haven't even decided for myself how that should go, whether I should call on people or somebody else should call on people. But we have reserved time at the end for "Well, what do you think?"

I'm excited that you're here. I'm so pleased that we got such a good reaction. Allison, you have a magic way of reaching out to people. I don't have anything more to say about logistics. Mike McPherson?

Michael McPherson, Spencer Foundation

I just want to take a moment to welcome you all to Chicago. If Dick does logistics, I realized that what I do is write checks. There are really very few things more satisfying with this foundation gig than getting to write a check that helps support an activity that you then want to be part of. I feel very happy, and I know that my colleague and vice president of the Spencer Foundation, Paul Goren, also feels lucky to be part of this and lucky to be here today. These are very important questions, and a very able group of folks around the room.

Michael McPherson

I'd like to mention our funding and supportive partner, the Joyce Foundation, and my colleague John Luczak, from the Joyce Foundation, who's here.

I spent a long time as an academic, and I do think that in the academy we're often much better at sweeping policy

statements than at questions of implementation. If you decided you wanted to do something, how do you actually do it? And often, God is in those particular details. I think focusing on the question in the way that it has been posed in Allison's terrific paper and in the plans for today, keeping away from the long speeches and moving toward real conversation — I'm very happy about that and very grateful to you all for joining in. I'm very grateful to the Rockefeller Institute and to Dick and his colleagues. So, welcome and thanks.

Allison Armour-Garb, Rockefeller Institute of Government

Good morning. I also want to thank you all very much for coming. When Dick and I first began developing the idea for this symposium, we couldn't predict how the No Child Left Behind reauthorization process was going to play out. But we did know that, no matter what happened, there would be a need to bring people together to talk about alternatives to federal dominance of educational accountability, on the one hand, and a completely devolved system, on the other. Fortunately, the Spencer Foundation and the Joyce Foundation agreed. I want to thank them for their generous financial support, which made this project a reality.

Allison Armour-Garb

The Rockefeller Institute, as Dick has mentioned, has done work on federalism for many years. With this paper and symposium we wanted to explore the problems of implementing accountability at the state level and to look at the different ways in which states can share power and work with the federal government and with each other to solve those problems. I know many of you have read my paper. (See text of the draft paper on page 109.) Basically, I tried to summarize some of the problems of implementing accountability systems that many of you around the room have actually identified in your writings on the subject, such as: the shortage of expertise in testing and accountability; the perverse incentives that exert downward pressure on state proficiency standards; the lack of transparency, resulting from the variation in state standards and tests; and the need for more research on the validity of state assessments and on the validity of accountability systems.

Sections 3 and 4 of the paper separate the questions of what functions an intergovernmental entity might perform and what institutional arrangements could be envisioned to perform those functions.

The possible functions I mentioned in the paper include: developing standards and tests, auditing or accrediting state and district accountability systems, and conducting validity studies. I'm sure there are a wide variety of ideas in this room about what functions would and would not be desirable. Similarly, you all have different ideas and arguments for what types of entities and institutions could best take on standard-setting and testing oversight functions, and for what could be politically feasible.

So, as Dick has said, we really want this to be a conversation among everybody here. It's a great chance for you share your ideas with a diverse and smart group of practitioners from state education departments, federal officials, policy advocates, researchers, testing experts, all of whom are interested in grappling seriously with these issues. I want to encourage you to participate.

In the first session this morning, we're going to focus on models for setting standards. After lunch we'll shift to focus on intergovernmental approaches to oversight of testing, which is also a two-hour session. Then we'll have a coffee break and have that final one-hour wrap-up session.

I also just want to thank Barbara Stubblebine who coordinated all the logistics for this event.

Again, thank you all so much for coming. I'm really looking forward to the discussion.

First Panel: Intergovernmental Models for Setting Academic Standards

Lynn Olson:

I'm going to start with a few questions and then quickly throw it open to the audience to weigh in.

So, I thought we'd start by looking down the road. Jumping over the current reauthorization timeline, say we're five years down the road. Where are we on the effort to develop common rigorous standards, and how did we get there? Checker?

A Cautionary Tale, and a Vision for the Future

Chester E. Finn, Jr.:

Well, it mostly depends on Mike whether we get anywhere at all. Where we need to get is a whole lot easier to say than whether we'll get there in five years. We've got a well-documented problem, which is that while standards-based reform is not going away, the current structure for doing standards isn't working well at all. What we lack as a country is an obvious mechanism for solving this problem. We lack confidence in the federal government in this area, maybe as much as in any area I can think of.

While standards-based reform is not going away, the current structure for doing standards isn't working well at all. What we lack as a country is an obvious mechanism for solving this problem.

We had some bad experiences in the 1990s that are cautionary tales for moving forward. You may remember the National Education Standards and Improvement Council (NESIC), which was actually enacted as part of Goals 2000. It was a Rube Goldberg, Noah's Ark structure for monitoring, moderating, supervising, and approving state standards and tests. And it was so unwieldy that it never got appointed, and then Congress abolished it. It was indeed a cautionary tale: Be wary of complex Noah's Ark style structural arrangements devised by Congress.

Where I come out is that we do need an interstate or multistate arrangement that we don't have today. It needs to be voluntary for states, not compulsory, because I don't think we will ever do compulsory for 50 states in this country, though it can be made appealing for more states to want to join, and it should have as little as possible to do with the federal government. If there were life in Education Commission of the States (ECS), it might be ECS. It could conceivably be the state education chiefs, who are edging this direction all by themselves. It could conceivably be the National Governors Association.

My personal candidate is Achieve and the American Diploma Project (ADP) as the initiator. There could be more than one of these, incidentally, and maybe there'll need to be, but I could start with ADP. I've told Mike this. He knows what I think he should do. He should go off for five years and build backwards from the high school exit standards that they've already got, all the way down through the grades, and then build an assessment to go with them and let states that want to, use it.

And keep the National Assessment of Educational Progress (NAEP) separate and apart as a kind of independent audit for both the states that do join this venture and those that don't, so that there is some external metric against which they're all then monitored. But don't tangle this up with NAEP as the testing entity.

From left: Michael Cohen, Chester E. Finn, Jr., Lynn Olson

Lynn Olson:

So, Mike, Checker's now given you your orders for the next five years. Five years from now, where are you and how did you get there?

Michael Cohen:

Before I do that, I want to acknowledge that among the 1990s disasters that Checker pointed to, Goals 2000 was something that I was intimately involved in. I'm pleased that even after that, Checker is still willing to let me play a role in the rescuing.

Simply telling states to align their high school standards with the demands of college and work wouldn't happen because they didn't know how to do it.

Chester E. Finn, Jr.

Redemption is always possible.

The American Diploma Project

Michael Cohen:

Let me say a word or two about what the American Diploma Project is around standards and testing, for those of you who are not familiar with it, and then answer the question of where we might be in five years from now.

For those of you not familiar, the American Diploma Project is now a network of 30 states. It originated in some research that Achieve did along with Fordham and the Education Trust, and for a while it was around the National Alliance of Business. It tried to define essential skills in math and English language arts that high school graduates need to have in order to be prepared to enter and succeed in credit-bearing courses in postsecondary institutions and have access to jobs that pay well and have some growth potential.

What we found in that research was what we believe is a common set of essential skills that students need for college and for work, to use a shorthand. We also found at the time we did that research that almost no state required students to either demonstrate they learned those skills, or are even taking courses that arguably, by title alone, would suggest they might have learned those skills. So we found this big gap between what students need to know when they leave high school and what they were actually required to demonstrate.

We challenged states to close that gap by aligning their standards for the end of high school, the assessments that they use to measure student performance, the curriculum and graduation requirements, the courses they required students to take, and the way in which they hold high schools accountable — to align all that with the knowledge and skills students need after high school. We found only about five states that were willing to take on that challenge. We now have 30, and signs are that the network will grow.

One of the first things we realized since this network came together was that simply telling states to align their high school standards with the demands of college and work wouldn't happen because they didn't know how to do it. So we organized a variety of training programs, supports, tools, etc., to help states do this. As we've worked now with close to 20 states, about 15 of which have actually completed this process, one of the things we are learning is that the standards that states set — their academic, college-, and work-readiness standards — are a lot more consistent across the states than the standards that were being replaced.

As states did this work, they asked the question, "What are the real-world demands students will face and what do they need to know in order to meet them?" The answer turns out to be pretty much the same no matter which state you're in. So there's an increase in consistency among the states in the knowledge and skills that they've identified for students. It doesn't mean they're the same. This is not taking somebody's expectations, having the government pay for Xeroxing, and then putting your state logo on it. But there's a high degree of consistency while there's variation around the core.

"What are the real-world demands students will face and what do they need to know in order to meet them?" The answer turns out to be pretty much the same no matter which state you're in.

The second thing that happened is a number of states basically said to us if we're going to go in this direction and typically require students to take advanced mathematics courses, like Algebra II, we're pretty sure that if all we do is require students to take that, then in half the high schools in our state Algebra II will look suspiciously like Algebra I. We need a tool, a strategy, a policy lever, that can help create a greater level of consistency. A number of chief state school officers said, "It looks like an end-of-course exam is the right way to go on that." But they also said, "We don't really each need our own separate Algebra II end-of-course exam." For crying out loud, why would you want Algebra II to differ from state to state? So out of that grew an effort, initially with 9 states, now with 14, that are creating a common Algebra II test that will be administered for the first time this spring.

So you can begin to see how the work that we've started to do state by state is beginning to look like a strategy for creating a high degree of consistency in standards and assessments across the states. So within five years, if we're lucky, we would have a set of state standards that are anchored in real-world demands — that means they would reflect what students need to know for post-secondary education, to have access to jobs, and, I would argue, that they would reflect as best as we can the knowledge and skills necessary for civic participation. Related to that, increasingly we will need an international benchmark that would reflect the expectations in other countries as well. These are research driven, data driven. It's not the product of a consensus among experts; it's a product of an analysis of the demands that students need to face.

So, anchored in the real world, state not federal, and (I think I would agree with Checker) voluntary.

Thirdly, these standards wouldn't just be in a document that would sit on a shelf in the state. They would be quite deliberately and carefully translated into curriculum, assessments, graduation requirements and ability systems. And not just the K-12 level; they would have currency in postsecondary education as well, so postsecondary placement policies would reflect what these college-ready standards actually define.

Finally, I'd say they'd be dynamic. That is, if standards are going to reflect the real world, then they have to change over time. The real world is changing pretty rapidly. The demands are exploding. There needs to be a mechanism for updating a set of standards. You'd expect that, while states would have a common set of expectations, the states that have done this work most recently might be a little further ahead, might have innovated a little bit, might have learned some things that the first set of states that did this work didn't know at the time. So you'd expect to see some innovation; you'd see a dynamic process. So that's what I think you'd see in five years. How many states will have done that is an open question.

Lynn Olson:

If standards are going to reflect the real world, then they have to change over time.

From left: Michael Cohen, Chester E. Finn, Jr., Lynn Olson

Mike, I want to push back a little bit on the 30 states, which makes it sound like we're more than halfway there already. So what's the big concern? Why are people agonized about this? Of those 30, give me a more realistic appraisal of how many you think are seriously doing this work in a way that, down the road in five years, they will have all those pieces you were talking about: real tests, linked to postsecondary … that you don't have 30 states that have signed on because they think this is the current thing that they better join on to.

Chester E. Finn, Jr.:

While you're pushing back, having common, agreed-upon standards is not the same as having a common, agreed-upon test with common, agreed-upon cut scores. If we don't get to that, too, I don't think we're where we need to get.

Michael Cohen:

First of all, I agree with Checker.

So here's where the progress is, and here's what hasn't happened yet as we work with states. First of all, with respect to standards themselves, here's a document that says what kids need to learn. Of the 30 states that have signed up to participate in this network, com-

mitting to do that work, we have worked directly with, at the moment, 22 or 23 of them. That work typically involves — actually in almost every case — a technical review of their standards at the front end, comparing what they currently have for high school-level standards with the benchmark expectations that we've discovered through our research and showing them what the gaps are; helping them figure out our process of engaging higher education, employers, and K-12 in any revisions of the standards; and then, depending on what process they use, two or three times afterwards reviewing their revised standards against the ADP benchmarks so that we know, with some degree of certainty, just what the consistency is and what the gaps are. So that's with about 22 of the 30 states. Of the remaining eight, there are probably two to three that look like they're sort of sitting there not doing much of anything.

With regard to curriculum and graduation requirements, when we started this work, there were only two states that actually required students to take the advanced math that we thought was necessary to prepare them for what they'll face after high school. There are now 15 or 16 states that have adopted new graduation requirements, so that's moving.

The assessment work is not moving nearly as fast, partly because that typically doesn't happen until after the standards set talk with us. There are lots of other things that are locking state assessments in place: existing contracts, NCLB requirements, etc. What is interesting though is that participation in the Algebra II test is growing. Again, we started with 9, there are 14 states now. Not all of them have not figured out what they're going to do with the test, so it's not necessarily the case that in each of these states every high school kid will take it. The test development is done; we're into the policy adoption phase now, so where that goes remains to be seen.

But that test will have a common cut score, to go to Checker's point. In fact, interestingly, when the chiefstate school officers raised the idea of common test, wholly unscripted in a room like this, sitting around saying, "Well, what are we going to do here? We're just starting this network, what are we supposed to do?" When the idea of a common test came up and there was some interest in it, somebody asked the question, "Well, who'll set the cut score?" This was in 2005, and still having the scars from the Goals 2000 effort, my response was, "Well, you're all independent states. You'll each set your own cut scores. We wouldn't dream of imposing a common cut score." And in unison, the participating states said, "Oh please, don't make us do that. You do it. First, we want a common cut score because we want common comparisons. And secondly, we want that externally because in the state, all the forces will drive the cut score down, when we see this effort as something to raise things up."

Again, what I don't know is how many other states are going to do this, and how many states are developing tests on their own, or in other groups or whatever will wind up with tests that are really rigorous and that will enable us to know when kids are really meeting those standards.

John Merrow:

Just an information question: In the 30 states, what percentage of the kids in the country does that represent?

Michael Cohen:

The 30 states collectively educate approximately 60 percent of the public school students.

John Merrow:

Even though you don't have California?

Michael Cohen:

Right. The big states that we don't have are California, New York, Illinois, and Florida. States that have come to us recently to find out more about joining include California, New York, Illinois, and Florida.

I don't know how many decades it took before they got a standard gauge on the railway system, but it took a long time.

Incentives for State Participation

Susan Traiman:

Five years from now, approximately 2013, the United States will then have been at this whole process of standards, graduation requirements, for about 30 years. So what makes you think that the historians won't look back and say, "The country has spent 30 years fiddling around with standards and assessments and graduation requirements, and no more kids were college- and work-ready?"

Chester E. Finn, Jr.:

I don't know how many decades it took before they got a standard gauge on the railway system, but for much of the nineteenth century, every railroad company had its own size track, and you couldn't move the rail cars or the engines from one company's lines to the next. And then somebody invented standard gauge railroad tracks. Some countries still don't have it, by the way. But it took a long time. They just figured out that having different gauges doesn't work well, and eventually somebody says, "Let's use the same ones."

Lynn Olson:

So, Checker, to follow up on that, you talked about this being a voluntary system that states would enter into, and I think the question is: "How do you get to a high common

From left: Robert Linn, Susan Traiman, Lorraine McDonnell, Richard Nathan, Judith Koenig

gauge in a voluntary system?" You said, "Make it appealing." What will make it appealing at a level where this occurs fast enough and at a rigorous enough level?

Susan Traiman:

To follow up, Lynn, what will it take to make states able to tolerate high percentages of students not getting their diplomas or, in the horrible case of the No Child Left Behind Act (NCLB), high numbers of schools getting identified as actually needing improvement.

Chester E. Finn, Jr.:

Mike Cohen said that it may actually help if it's externally set rather than having to be set within the state. Then you can always blame the common cut score setters. But comparisons on a simple metric have many plusses. Look at all the states that are rushing to join things like the Trends in International Mathematics and Sciences Study (TIMSS) and the Program for International Student Assessment (PISA) all by themselves just because they want to be able to be compared with something beyond themselves. Some states really are led by people who care about things like the economic future of the state and its competitiveness.

Look at all the states that are rushing to join TIMSS and PISA just because they want to be able to compared with something beyond themselves.

People are understandably reluctant to bring the federal government into this, but there are some NCLB-type wrinkles you could give states as incentives: You could give them automatic approval of their standards and tests without having to prove anything further to the Department of Education if they join the common system. You conceivably could give them additional cash to cover their testing costs if they join the common system, maybe sweeten the pot a little bit that way. I'm worried about that, however, because instantly you'll get into a kind of approval process in the federal government that I'd love to avoid.

Test Uses and Stakes

John Bishop:

I'm not sure exactly what Checker means by a common cut score. A common cut score where graduation from high school depended upon it I think would be a disaster. A common cut score to define proficiency is no problem. I think you need a metric that is common, just like the Scholastic Assessment Test (SAT) is the same test taken in all states, or the American College Test (ACT) is the same test, so a 23 on the ACT is equal across the country. You want an exam that does that. But you would have no standards in Iowa and Minnesota if it has to apply in Mississippi and D.C. at the same time. In fact, I doubt that you would be able to retain a lot of the states, or at least you'd have to leave out a whole bunch of

states from the system of states volunteering to participate in the system if they are required to not hand out a diploma to people who fail this test.

Another thing that I think is important for the performance of the kids on the test that Achieve is preparing: There needs to be stakes for the kids — but not necessarily high stakes. Like, it might be one of the grades that count towards the grade in their course, in a system that does not have an end-of-course exam for everybody in the state. Or it might be, as was the case with Regents in the old days, where the end-of-course exam was the final exam, but there were also teacher grades that counted in determining the grade of the student. But if we were to require all students to pass a common end-of-course exam in order to graduate, states would have to lower the cut score in order not to fail too many people.

Michael Cohen:

John, I think your question demonstrates the complexity of this issue. First, let me note, of the states that are participating in this Algebra II test, there is at the moment only one that's even thinking about making passing the test a requirement for graduation. And if they do, my guess is they will need a different cut score than these common ones in order to manage that situation.

A number of the states are thinking about counting the test towards the course grade. Interestingly, once they go down that road, you can ask them, "Well, what else counts towards the course grade? If the test is 25 percent of the grade, 35 percent or whatever, what's going on with the rest of it?" The answer is, "We don't know." It's difficult for them to imagine having a course grading policy statewide. They'll have to think their way through that once they start saying the test is going to count for some portion of your grade. It forces you to worry about the rest.

The other way in which we envision at least this Algebra II test being used is that students who score well enough, proficient presumably, that score ought to at least relieve them of the requirement of taking a placement test to figure out if they're going to be in a remedial versus a credit-bearing course. There are very interesting conversations underway in postsecondary systems in each of these states. "Very interesting" is a euphemism for trying to get higher ed to agree to something like this. That makes the rest of it look pretty easy, but we think we're seeing some progress in some areas on that.

Chester E. Finn, Jr.:

By common cut score, I meant a definition of proficiency, not whether you get a diploma or not. States are going to continue to have their own quirky requirements for getting a high school diploma and I don't think that can or should be standardized. Picture Connecticut kids being told they have to take Texas history before they can graduate from a Connecticut high school. It's not going to happen.

Michael Cohen:

But if you had Texas kids take Connecticut history, that might be a different story.

Margaret Goertz:

I just want to follow up on the sanctions and incentive structure with high school exit exams. The states that have moved for end-of-course exams, for high school exit exams, have been backpedaling in terms of when they're going to actually put that in place. It's complicated by the fact that essentially a high school diploma is a property right. So I'm wondering, putting Algebra II aside, and thinking about Algebra I, Geometry I, and the other areas in which you see there being assessments being developed, how do you manage that tension between that ultimate sanction of the high school exit exam and wanting to ratchet the standards up?

Michael Cohen:

I haven't used the word high school exit exam any place. We haven't said to the states that we're working with, "You have to have exit exams." We've said, "You have to have assessments that are aligned with the standards, and they ought to count in some way." I'm not at all sure that I would march down the high school exit exam route the way it's currently practiced. Or if I did, if I were running a state, it might be only in the ninth and tenth grade courses, not anything much beyond that. Our own studies of what high school graduation exams measure was really eye-opening. We did this study in 2004. In math, to pass those tests, you have to demonstrate that you've learned skills that students in other countries typically learn in seventh and eighth grade. This is not very high.

"You're right. You do have a really good eighth grade math test. In fact, you have two really good eighth grade math tests. You just happen to give one of them in tenth grade."

I'll tell you, when we released our report, I met a governor in one of the states included in that study, one I won't name, who basically was really angry because he said he had gotten a review of his math tests from us a couple years before and we said it was great and now we're saying it's lousy. When we probed that, it turns out it wasn't us, it was one of our colleagues who had done that review. It turned out that we had looked at their tenth grade graduation test; he looked at their eighth grade test. We were able to compare our analyses. To make a long story short, there was a letter that I couldn't quite bring myself to send to the governor that said, "Well, you're right. You do have a really good eighth grade math test. In fact, you have two really good eighth grade math tests. You just happen to give one of them in tenth grade." There was not much of a difference between the two.

So I'm very concerned that those pressures, right now at least, are watering down the standards at a time when it's really important to raise them up and find some other kind of consequences for students that are important enough to motivate them and not so forceful that they drive things down.

European Systems: Multiple Graduation Standards

Lynn Olson:

John, can I ask for a point of clarification about other countries that have national standards? Are they graduation tests, or are they more college entrance tests for a certain portion of the population?

John Bishop:

John Bishop

Well, the systems in France and Britain evolved originally as subject-by-subject graduation tests that simultaneously served as college entrance exams. Then, as a result of democratic pressure in the post-World War II period, they wanted to increase the share of students getting secondary degrees, so they created vocationally oriented examinations that parallel the academic exams that had always existed in their secondary school system.

Any system that is going to graduate a large share of the population has to have an accountability system that essentially hands out different types of credentials reflecting different types of achievement.

So now maybe half of all the baccalaureate qualifications awarded are in professional and technical areas, or maybe it's a third. But anyway, it's a large share of people getting credentials in an occupational category. And although they're nominally the same standard, everyone in France knows that the program in the math/science/hard science area is the most prestigious, and the social science one is the least prestigious in the academic area, and then it goes down in professional. In many countries, there is more than one standard for the math exam you can take at the end of high school, just like we have the Advanced Placement exams and some other exams.

I think any system that is going to graduate a large share of the population has to have an accountability system that essentially hands out different types of credentials reflecting different types of achievement to different kids, because the standard deviation of performance on math tests, or almost any test, is like four grade level equivalents.

Lynn Olson:

So more than one standard?

John Bishop.

So if you're going to have a standard that 90 percent of the population is to achieve, you are not incentivizing a big chunk of the kids.

Lynn Olson:

So more than one standard, and the standard is meant to provide a stepping stone to the next place, higher education or an occupation, rather than just a gatekeeper out of high school?

John Bishop:

Right. One very quick summary phrase is: Everyone can graduate from high school if they stay there long enough and take the exams, but they may graduate with a record that is not particularly great. So it's there on the transcript as to how much you achieved, and you get rewarded for achieving a lot more because you get access to better universities, and you get to be a lawyer or a doctor if you want. Secondly, employers pay a lot of attention to the performance as recorded on your resume or in job applications.

Lynn Olson:

So that is different than pass/fail.

John Bishop:

So it's not a pass/fail. You would typically have four, five, six levels of performance reported. Then, also, the type of exam you take signals competencies, and the number of different exams you take, because you don't necessarily take this, this, and that exam. You have a few required exams; you get to choose the other exams that you take.

So I think in order to get a high graduation rate, you need to have this kind of a flexible system. I think part of the reason for our low graduation rate is the effort to apply the same standard to absolutely everybody in the society.

Michael Cohen:

John, you said a minute or two ago that in European countries, everyone will graduate from high school if they stayed there long enough and ultimately pass.

John Bishop:

Yes. There is a lot of grade repeating in many European countries in the continent.

Michael Cohen:

And that while the transcript provides evidence of longer time, grade repeating, that that clearly reflects negatively.

John Bishop:

Yes.

Michael Cohen:

Even if they've accomplished the same end, that other students did it faster. So both achievement and time turn out to matter, is that clear?

John Bishop:

Yes. So time matters; the number of different courses you prepare and present and take the exam in, because you choose which exams you prepare for; and, of course, the reputation of that exam. So in France, the math exam covers up to Linear Algebra, the end of high school exam for the baccalaureate qualification. And, of course, if you're trying to get into one of the Grandes Écoles, you go even longer and take an even tougher exam.

U.K. and Australia: Choice and Multiple Awarding Bodies

Eva Baker:

I have just a couple of things. Number one, I agree 98 percent with what John said as far as what my understanding of the state of the world is. I just wanted to make a couple of additions. What I think is attractive in a lot of countries is that there's a good deal of student choice available in what subject matters they take. And that choice does not mean that they're necessarily put on an academic go-to-college path or go-get-a-good-job path. There's some crossover on that. For example, in New South Wales, a third of the students opt for an art certification. Wouldn't that be nice? But at any rate, that's one point to make on choice.

From left: Stephanie Dean, Lauress Wise, Eva Baker

Secondly, there are some negative consequences of certifications. That is, in the United Kingdom, kids have to choose early and then they are completely on a narrowed path, so they don't get a chance to say, "Oops, I meant this." Once they're on that path, they're there, and it's very difficult for them to circle back and rethink.

Third, and this is not meant to be, Mike, taken as any kind of implied concern about Achieve as a test development entity or whatever, but I think there are some interesting models out there where there are, like in the U.K., what they call "Multiple Awarding Bodies." That is, groups that provide that the standards are in common. There is even a tighter

syllabus than we tend to have generally out there. And then there is room for modification among different providers, who not only develop and administer the examination, but may also provide professional development material, or activities for teachers, and so on. Then there's some process by which, I'm not so thrilled with, but there's a process by which there are some agreements made that adjust performance on different tests. But very often, for instance, the Political Geography course will be offered by one group, whereas the Physical Geography course will be offered by another group, and the schools and then students make decisions about which ones they take.

So my thought is that we need to hold things together with good, clear standards (not the sort we've been used to). We also need some oversight through a group or entity addressing comparability of standards and performance and the quality of the examination system. In this way, we may infuse this system with a lot more choice for students and I think giving students choice in secondary school matters in their development, maturity, and capturing their interests. I'm for broader choices in certifications or qualifications rather than converging on one, two, three, or four areas in an exit exam.

Chester E. Finn, Jr.:

It might be worth just noting the forgiveness *within* the secondary system that John and Eva were talking about in other countries. We tend to be more forgiving *outside* the secondary system here, with General Educational Developments Tests (GEDs) and adult education and community colleges and other ways you can recycle yourself after you go through the relatively monolithic secondary system that we have.

Michael Cohen:

Eva, talk a bit more about multiple awarding bodies. How would you envision them getting created here? Because it seems like an interesting idea. God knows, we haven't set out to create an awarding body, but …

Eva Baker:

Well, I have sort of two sets of deeper understandings: One is Australia, one is the U.K. There are probably those in the room who know more about both of those than I. But because I was the chair of the group to review the A-level exams, which were multiyear and multipolitical, as you can imagine, I had to learn a lot about how those tests got made and who set cut scores and whatever.

"Multiple awarding bodies" is another term for examination boards in England. While there used to be many, they have consolidated to three or four, depending upon who you think is a player. Right now, for instance, the Educational Testing Service (ETS) has won the award to do some of the national assessment in England (not A levels), a contract that was held before by Pearson, because they bought an examination board that was already existing. These boards are independent companies. They agree to explicate and work together

on standards that are put in place by various groups, sometimes governmental, sometimes not. But there's an agreement on the standards and promulgation of those standards. They then create what they think is the right curriculum to meet those standards, and professional development, and then people opt for the package, in general, as I understand it. There's some notion that, "Oh, well, if you take math with this board, it's a better shot than this board." But the universities also know, so if you expect to go to a top-tier university, you probably ought to be working with a certain examination board.

It's a little different in Australia, where they have a "Board of Studies" managing the syllabi but tend to contract this examination function out to one or two major providers. What's interesting in different Australian states is that they weight the choices kids make, so that if you choose "easy" social science, English literature — all my favorite things — instead of very high-end math and science, there is no way that you're getting into a med school, no way you're getting into certain kinds of programs. They are weighted and put on some sort of conflated scale, which also worries me a little.

But the way these things are handled in England (and this is what Checker certainly won't like) is that there is a group called Qualifications and Curriculum Authority, now led by an Australian, Ken Boston, and they serve as the regulators They convene people who are the people who write the exams, who review what's being done, who meet with the awarding bodies or publishers.They have found an uneasy kind of agreement. It's now a tradition there. So that, I think, in some ways fits better with our view that there shouldn't be a monolithic "it" that everybody has to pass managed by one test publisher, who may or may not be as transparent as we would like. I'm not advocating this approach fully, but just setting a range of options.

The question is: If you don't trust the federal government to develop the common test, who do you trust? Wouldn't you want to leave it up to private industry?

A Single Test, or Competition to Create Multiple Tests?

Lauress Wise:

I wanted to push a little further on this notion of commonness, and one of the things that Checker said that concerned me a little bit, common tests. The question is: If you don't trust the federal government to develop the common test, who do you trust? I think that we could make huge gains in relieving the capacity problem by having a set of specs for one fifth grade math test instead 50. But wouldn't you want to leave it up to private industry to compete to be innovative and creative as to different ways of developing a test to those content specifications, with maybe some accreditation on the back end to certify that they had or had not developed an adequate test? Otherwise, who would develop the test and who would decide that?

Chester E. Finn, Jr.:

You're saying develop uniform common standards, but allow for multiple tests built off those standards.

Lauress Wise:

Competition for the best possible tests. Let states, or whomever, pick among different certified tests of these common standards. Especially as you push it down into the lower grades, there are a lot of tests to be developed in every grade and subject, and it's one thing to do an Algebra II test, but multiply that by huge amounts of scales. Would you want a single entity doing all of those tests?

Chester E. Finn, Jr.:

Two concerns: One is where does certification come from? Second, if each state, in effect, continues to hire its own test contractor, which has typically been the practice up until now, then you don't get comparability across the states, and you get this finagling-with-cut-score problem as states decide what is politically acceptable for the proficiency level rather than what's desirable for the proficiency level. I worry about those problems accompanying the model you were sketching.

Lauress Wise:

It's just I think there's a huge difference between asking companies to develop "the" fifth grade math test versus the Delaware fifth grade math test, or the Idaho fifth grade math test, and so on. So that I think you could manage the situation much more effectively with regard to common adequate tests if you had the common standards that the tests were being built to.

Michael Cohen:

The question is: Whose common standards? Who certifies the tests? I think there are real advantages to not having a single test that one publisher develops and we're all betting that it's a good test. Competition, I think, among the test developers, among private companies, can hopefully lead to good quality, particularly if it's a more focused competition, rather than 50 tests in every subject, every grade level. But we, at present, don't have a mechanism to organize that kind of competition.

John Bishop:

There's a tendency for Americans to say competition is the solution for all problems. But many of our current educational problems are caused by the competition to become the dominant college entrance exam. The basis of that competition is you persuade colleges, or state governments, state university systems, to say, "This is the test that we pay attention to." So that's one constituency, and the second constituency is students. We can't have them in the room for more than three hours.

If you select the college entrance exam based on how long the test is, the price of the test to the student, and the convenience of colleges, you have a lousy system for selecting

people for college that does not take into consideration the interest of the educational system to set high standards and to have broad standards. We end up with, until very recently, no essays as part of the examination system.

In all the other countries we've been looking at in Europe, these systems of these tests are to the specification of the Ministry of Education in the nation, and it's taking into account: What do we want people to know, and how do we think we can assess that? They believe they need essays to assess a lot of the things they want to assess and, therefore, that means the cost of doing the test, and the time that kids are engaged in taking the test, is an order of magnitude above the American pattern. You take a test for a whole day for one subject, and then you might take eight subjects. So you would be taking tests for, at the end of high school, two weeks, and in the whole school system the seniors are all doing that.

That's the way it used to run in New York. At the end of twelfth grade, those exams were three hours, subject by subject, and there were at least 10 or 12 different subjects that were being offered.

Gordon Ambach:

Twenty-six.

College entrance standards and tests are the default national standards at this point.

John Bishop:

Twenty-six. You cannot have a high-standards system if kids don't spend a lot of time taking the test. And competition has to be on the right basis.

College Admissions Tests as De Facto Standards

Michael Cohen:

You've got several things going on here. One is that the college admissions tests are developed and selected on an entirely different mechanism than the K-12 system.

John Bishop:

In the U.S., not in Europe.

Lynn Olson:

Actually, Wayne, this is a good time to call on you, because of two issues, both of which just got fleshed out here. One is we have one system of standards and tests K-12. We have sort of a separate system of college entrance standards and tests, which are the default national standards at this point. That's one issue that's been raised. So there are other systems besides Achieve out there, actually, that have some weight. Go ahead and weigh in on this.

Wayne Camara:


Wayne Camara

Two points: The higher ed issue is very complicated, and efforts to bring the higher ed community together to get consensus — I wouldn't want to put any money on that. We've got all kinds of schools. We've got schools that say, "A kid with a 700 math score fails in our engineering program. We want 750 on the SAT IIs." There are lots of very good schools that have different cut scores. Even when they don't, just look at the AP, which is pretty much a national standard in some ways: There are colleges out there that will accept a 3, there are colleges that will accept a 4, there are colleges that want a 5, there are colleges that will accept a 3, 4, or 5 — but only for two courses. So I think that part of it is real difficult.

I think, honestly, the SAT and the ACT are examples of where things do work quite well. These are two fairly different tests in some ways, but colleges can use them interchangeably in ways that I know the federal government has been unable to look at multiple tests and make comparisons. I think the SAT and the ACT are very high quality, comparable, and they allow choices.

As long as there's no commonality among standards, a core, you can't have that common test. If you have a core, you can add to it with modules and enhancements.

The way out of this really, unfortunately, is that you do need common standards. I don't think you have to throw state standards away. But I don't think that you can say, "Believe the Arkansas standards, or the Massachusetts standards." You do need states that are part of this to say, "Here are our common standards. Now in Massachusetts, we tacked these on. But we began with a minimum set, and we accept these." Until you have agreement about common standards, I think these efforts will be very difficult. If you could get ADP or some other group to say, "Here are 20 math standards for eleventh grade." Now, if every state wants to add to them, that's okay, but only at that point can you develop a common test with common items. As long as there's no commonality among standards, a core, you can't have that common test. If you have a core, you can add to it with modules and enhancements.

I know this is where the policy and the technical side come up to confront each other. The only other point I wanted to make is that the problem in a lot of these efforts is that we are about trying to reach consensus and bringing as many players as possible to the table, 30 states, 50 states. When we do that, we have to come up with something very vague, very amorphous, that we can agree to — not specific. Would it be better to begin with very specific requirements? For example, what's the purpose of the test or the standards? If it's college readiness and work readiness, perhaps it's accountability, and maybe a secondary purpose is for placement, for admission for college. And that's it.

So if a state comes in and says, "Well, we want to use it for graduation purposes,"we're comfortable in saying, "Absolutely not. It's totally inappropriate for that purpose." Because when we open the door and we allow every state to come in with their own purposes, we simply overburden that system to a point where the technical requirements, the course requirements, the policy requirements will just simply make sure that that proposal falls like a

house of cards. I would be more comfortable in moving this forward and saying, first of all, "This is the explicit purpose of the test and the standards, and that's it, guys. If you want more, do more on your own. But this is the purpose. It is for accountability, college readiness, and work readiness. It is totally inappropriate for high school graduation purposes, and if you want to use this test for that, you've got the wrong consortium." I'd like to look what would happen if we limited the focus.

Michael Cohen:

When we worked with states and put the Algebra II test together, they did agree on the purpose as a college-ready indicator to provide some consistency in what's taught in Algebra II within and among the states, and the ability to compare performance across the states. Those are the purposes. Now, what we did not do is say, "You may not use the test for any other purpose." And I honestly don't know, as I think about where we are in the process, whether we could enforce that. I mean, I suppose we could throw states out of the consortium. Honestly, it hadn't occurred to us at the time, and there's only one state that, again, is thinking about doing something other than the common purposes that we all agreed to, and I don't know that they're ever going to get there.

The question of what model you choose that has enough transparency to get the public's buy-in is really important.

Bethany Little:

A couple of thoughts on the NCLB piece, these little pieces that are coming out here and talking about who does what. I think on some level, that is what we have now. There are certain people who have tests for outcomes at the end that the marketplace decided on — the SAT or ACT or what have you. I think part of the challenge for the education community is the transparency of that process, in the sense that nobody has particularly bought into that. That isn't a set of outcomes that they understand where it came from or who decided that. So I do think that the question of what model you choose that has enough transparency to get the public's buy-in is probably really important.

Michael Cohen:

You mentioned that transparency is important and that neither the SAT or the ACT have it, and so there's not a lot of public buy-in to it. I think it's true that those tests are not as transparent as some state tests, or at least state standards-setting processes, are. But arguably the SAT and the ACT have much more public support than state standards do, because they actually count for something in a way that people care about.

But because they come out of a different perspective, they don't attempt to align them with the standards that states are setting. That's where the rub comes in. Cindy will probably have a word or two to say about that. There's a gap between what the test measure and what the state standards are that makes it harder for the K-12 education system to buy into those tests, though it's quite easy for the higher ed systems to value those tests.

Chester E. Finn, Jr.:

Well, I'm pointing out the obvious, but the SAT and ACT don't have a lot of traction in the primary and middle grades. They really are creatures of the end of high school. They don't create a testing system that makes sense in an NCLB environment, where you've got a lot of push on grades 3 through 8 right now.

Data-Driven vs. the Product of Consensus

Lorraine McDonnell:

I'm struck that, in a session that's titled "Models for Setting Standards," we got so quickly got to testing. I think there's an issue there. But I want to come back to the models for setting standards and ask Mike a question: He said that these would be data-driven and not a product of consensus. That's a very different model than the states themselves have used where it is a political process, and underneath is the essence of the word, that the community has to come together and say these are what's important to us, and it's not just the technocrats, or the business, or the cultural conservatives, or whoever. So I wonder how you can set common standards without having it be, at least partly, the product of consensus among these different societal groups.

From left: Allison Armour-Garb, David Abrams, Lorraine McDonnell

There clearly is still a substantial element of consensus-building there. But what people are agreeing to is what their best interpretation of the evidence is.

Michael Cohen:

Good question. First of all, there's clearly a small "p" political part of this, and there's a fair amount of consensus building that goes into this. But fundamental to the question that people in each state are asking, those who are working with us, is, "What's the evidence that you've got about what young people need to know in order to succeed after high school?"

So when you're looking at a college-ready measure, for example, part of this research that we did in the first round of the Diploma Project, in which we worked with faculty who teach first year credit-bearing courses in largely broad-access institutions in five different states through a variety of means, was to try to get them to be as explicit as possible about what work young people actually do in their classrooms, so we can identify the prerequisite skills. What does the research tell us about what data can we find about the relationship, for example, between high school course-taking and test score performance on the one hand, some metric about what students learn, and postsecondary success? We looked heavily at the work that ACT had done in their own studies to identify what young people need to know in order to be college ready on the ACT measure. So we tried as much as possible to make this data-driven, as opposed to what largely went on in the 1990s, which was pulling experts together in each discipline and asking, in effect, "What can you people agree would be desirable for young people to learn?"

Now still, when you go through the kind of process that we've gone through, there clearly is still a substantial element of consensus-building there. But what people are agreeing to is what their best interpretation of the evidence is. I'll tell you, the dynamics in those discussions is quite different from what we saw in the 1990s. There was very little debate in the states among these higher ed, business, and K-12 folks who were involved in defining standards — very little of the culture wars that went on in the 1990s. The debate that goes on is, "What's the evidence that kids really need to know all that level of advanced mathematics in order to succeed after high school?" It is very much about the evidence. It's very much not about the politics of it. I think that's a big difference.

Implementing Standards-Based Curricula

Lisa Hansel:

Some districts can take a set of standards and write a curriculum, and some can't. And that has a big impact then on what kids actually learn and how that plays out in the tests.

I have a similar comment. I'm a little uncomfortable with going directly from standards to tests, so I'd like to hear both of you just say a little bit more about all the intermediate. Who writes the curriculum? What about the instructional guides? What about professional development? What are we going to do about the textbook industry? Because I think that one thing that has been demonstrated over the past two decades is that some districts can take a set of standards and write a curriculum, and some can't. And that has a big impact then on what kids actually learn and how that plays out in the tests.

Lisa Hansel

So I'm wondering: Would it be appropriate to say that states should take on more of this burden in terms of instructional guidance? Because right now, there are too many districts where it just falls on the teachers. Whether you're a brand-new teacher or a veteran teacher, you've got to figure out what to do with your kids, and you've got to go home and write your lesson plan every night. And that seems to me to not leave enough time to focus on individual student needs and how to teach the material, as opposed to writing the material first. Even if we had much better standards, there's still a problem there. The most specific set of standards still leaves an enormous amount of flexibility in terms of how that actually gets taught.

Michael Cohen:

I can tell you how we've dealt with or not dealt with that in our work. We have not tried to develop a common curriculum across the states. What we've done is start with the standards for end of high school. We have articulated them into a series of course descriptions, so at least we can figure out this is what kids need to know at the end, what should they learn in Algebra I, what should they learn in Algebra II. But we've not gone beyond that. Frankly, we don't have the capacity to do that. Maybe we ought to be pulling states together and agreeing on curricu-

lum and lesson plans and the like for courses, but we've not. It would be hard for us to do this, I think, until states were themselves willing to step up to the plate to do that kind of work and create those kinds of tools. All too often, after the standards and the tests, they're local control states. It gets left to someone else to do. And I think that's a gap we should be filling.

Lisa Hansel:

So you do think it's a role that the states should take on, even if they don't want to? Is that it? Or am I pigeonholing you too much?

Michael Cohen:

I think states need to do more in this space than they currently do. I don't think that means they need to create a curriculum that everyone then must implement. You could create a model curriculum. You could pull a consortium of all the districts together to create one on work. There are a variety of things that could be done. Some states are fairly aggressive in that, and lots of states are standing back. But it's pretty clear that if we leave that kind of task — curriculum, formulating assessments, professional development — to 16,000 school districts, lots of them won't create anything close to what's needed.

The most specific set of standards still leaves an enormous amount of flexibility in terms of how that actually gets taught.

Chester E. Finn, Jr.:

Or it gets left up to three million teachers. We could easily divert ourselves into a discussion of the teaching workforce here and what it needs to do its job. In a lot of other countries when I look at their standards for a course or a subject I am struck by how skinny those documents are. A fair number of countries seem to assume that if the standards are laid out in broad outline form, the teachers will be up to finding ways of teaching them — either because they know their subject so well, or because they have access to other sources of curriculum and instructional ideas, materials, lesson plans, and things like that.

I think the states' foremost role here is to get the standards right and get the assessments right. Then the fill-in — everything else in between: teacher training, curriculum, textbook, professional development — is going to play out differently in different places. Some schools will do it, some districts will do it, some states will do it statewide. In some places, it won't be done well at all.

Federal vs. State-Led Efforts

Lynn Olson:

So I hear a little schizophrenia here, because we're saying, on the one hand, we want high common rigorous standards. But the more we get to implementation, it quickly de-

volves back into 50 states and maybe 16,000 districts all doing their own thing, but somehow it's going to add up to commonality. And I'm not seeing it yet, Checker.

Chester E. Finn, Jr.:

It's commonality of expectation. It is commonality of what you mean by proficient in sixth grade English.

Lynn Olson:

So let's push a little bit again on who is choosing that commonality. It's not the federal government, because you decided you don't want the federal government near this. Is it a certifying agency? Do we create something new? Is it a group of 12 states that come together and hold hands and decide they're going to go down one path?

Chester E. Finn, Jr.:

It's not as though we started out saying, "Let's have common standards for all states." We've been responding to demands from states who themselves want a greater level of commonality.

Three small states in New England have already decided they are going to do this in common and have done this in common. So there is a tiny little prototype already visible in the United States that is being done beyond the state. Once again, I think this could be done by Achieve. This could be done by the Chiefs. We could go down the list of existing bodies or entities. We could also create something that we don't have today. This is, I hope, why we're here today, and other conversations like it in the months ahead. Because this is the problem we need to solve. We need to do all the imagining we can about mechanisms for filling in this pretty substantial gap in the middle.

Michael Cohen:

We're backing our way into these issues. I'm speaking of Achieve. It's not as though we started out saying, "Let's see if we can have common standards for all states and a common test." We've been responding to the demands and pressures from states who themselves want a greater level of commonality than we have now. We haven't set out to design a mechanism for solving all these problems. But the reason we have the problems now is because the states have pushed us to take some steps. As we do that, all these other issues come up right away. There aren't immediately apparent solutions to them.

Scott Marion:

Lynn, I'm glad that you brought out this schizophrenia because that is something that I was concerned about. I'm with the Center for Assessment, and we are very involved in the New England Common Assessment Program (NECAP) — it's not a great acronym. If you go to Google Earth and you look at these three states, they all fit inside many of the western states.

I can tell you that this wouldn't be happening if it wasn't for both desperation and the force of the personalities of the three assessment directors who are keeping this going. It's not because of any grand scheme in service of some model for how to do this larger. They've actually learned to work together to get along. So there's this piece about how this works. If we switched out some of the personalities (like a lot of what happens in education), I'm not convinced that this would work. Now it's generating some institutional history, and that's hopeful.

I can tell you that we had hoped that Charlie DePascale (Senior Associate, Center for Assessment) in our shop — who has been intimately involved, has been with ADP — was going to work just one, two days a week at the most with these states. And he is essentially 60 percent time at NECAP, and half of it is as a family therapist. And if you know Charlie, that's not a great role for him.

So that is one piece. But this schizophrenia really bothers me. Checker, you talk about this cut score finagling across states, and I agree that's a problem. You haven't convinced me yet why federal standards in certain areas are a bad thing.

When Secretary Spellings is replaced by Secretary Weaver in the next administration, I don't think the standards should be done over again. I don't think that when George Miller takes over from John Boehner the standards should change.

And there's a couple of pieces of that. How much of the school year should be taken up? Lorraine was talking about some common physical education standards. We don't need to be thinking about that. We don't necessarily want the entire curriculum.

I agree with Lisa that we need more specificity. But this is the conundrum. In other countries that do this well, they have very skinny standards. But they also have very skinny textbooks, because they do have a common set of standards. We know that a lot of curriculum is text-driven. In this country all publishers are trying to be all things to all states, or at least a few states. My 10th grade daughter has back problems now from carrying her backpack around, these things are so heavy.

So we point to these international examples, but then we back away. We say, "Yeah, it works. We like this part of it." We will use them as an example when it serves the purpose. A lot of them do have national standards that are government sanctioned and government run. They get some efficiencies (Laurie Wise and I were talking about this at breakfast) with textbook publishing and professional development and teacher training. There are certain efficiencies we get now.

Of course, there is always the question of who gets to decide what the standards are. In this current administration, I might not be in favor of who they might be bringing together. I am happy with Achieve running some of the show right now. So, I will just leave it at that. What's so wrong with having a single agreed-upon set of standards? I agree with John and Eva, let's not make it all the way through high school — but through 9th, 10th grade, and then have some choice? What's so wrong with that?

Chester E. Finn, Jr.:

Nothing's wrong with it. It's just that I don't think the federal government should do it. I don't trust anybody there to do a good job of it. I don't think it should change when Secre-

tary Spellings is replaced by Secretary Weaver in the next administration. I don't think the standards should be completely done over again. I don't think that when George Miller takes over from John Boehner the standards should change.

Today, we don't have a federal mechanism. We can talk about the National Assessment Governing Board if you like; that's the closest facsimile of one today that is connected to the federal government. But this is fraught with extraordinary political difficulty, lack of credibility, lack of consensus, and it isn't going to work very well.

As for NAEP, to repeat a point I alluded to earlier, I think it should stay outside and be the universal auditor, not the basis for high-stakes decisions about schools or districts or states, for that matter.

Lynn Olson:

I would be curious about what some of these mechanisms are in other countries that let them get beyond their current political abuse of power.

There are two studies recently that both raised the question of the relationship between performance standards and what states are actually accomplishing in terms of student performance. And both of them say to me very loudly that there is no clear correlation.

Chester E. Finn, Jr.:

Yeah, that's a really good question.

Theory of Action

Mitchell Chester:

I just want to push on the presumption that going after the standards and the tests is sufficient. I suspect no one here would claim that. But I want to kind of push this a little bit because I think we are operating from what I would consider an inadequate theory of action, i.e., that the main game needs to be getting the standards right and getting more consistency on those. We are talking, in part, content standards, but largely performance standards on these tests.

From left: David Abrams, Mitchell Chester

There are two studies recently. Checker, your organization sponsored one, and then there was the Institute of Education Science (IES) study this past summer, that both really raised the question of the relationship between those performance standards on the tests and what states are actually accomplishing in terms of student performance. And both of them say to me very loudly that there is no clear correlation. Now, that probably doesn't mean that we shouldn't worry about the standards. But the problem, it seems to me, that we are trying to solve, kind of that means/ends issue on the table here, is how do we get more students through the K-12 system with the kind of knowledge, understanding, and skills that are going to serve them well beyond high school?

Neither of those studies gives me real confidence that knowing where a state sets its performance standards, in and of itself, gives me good insight into whether the degree to which they are accomplishing that versus other states. The IES study was very direct about that and talked about the fact that you couldn't predict statewide performance based on where the cut scores were set on state tests. While that wasn't the main conclusion that the Fordham study drew, I don't think you'd see much of a relationship if you compare performance to what you inferred about the variation of state standards.

So I just wanted to push that we're operating with a very inadequate theory of action, only focusing on the standards. That, in and of itself, is not going to get us there. It is probably necessary, but insufficient.

Lynn Olson:

Well, Mitch, it sort of comes back to what Lisa's been talking a little bit about the notion of curriculum-based tests that have curriculum and implementation attached.

Michael Cohen:

It also goes to that point Susan raised early on: We've been at this standards and testing business for 30 years without necessarily seeing the results. I think that's partly because we had often operated as though standards and tests are both necessary and sufficient when they're clearly not.

Unless there is at least significantly more attention right now paid to professional development, teacher preparation, curriculum, instruction, and all those things — if that doesn't happen, then, although it's important to get the standards right (because you don't achieve what you don't expect), that's not enough. Common, state-specific — it doesn't matter. If the other pieces don't follow, neither will the achievement.

Common, state-specific — it doesn't matter. If the other pieces don't follow, neither will the achievement.

Lynn Olson:

So we don't want the federal government involved, but we need curriculum, professional development, and common standards and assessments. Where is the money going to come from? Who do the partners need to be?

K-8 Standards

Phoebe Winter:

That's a little bit about what I wanted to talk about. We've talked a lot about high school, which, although it's very, very complex, is much more contained — in terms of what the courses are and the issues are and things like that.

When we get to details, we've never said anything about K-8 — or not much about details. Scott brought up: Who is going to do it? Who's going to set the standards? That's important. I can almost buy common grade-level standards at K-8. Almost. I'm more of a grade span kind of person myself. But common assessments at that age, I just think we'd really have to think through. New England Compact did it. It hasn't been easy. It's mostly been because of money, not because of some common philosophy (from what I understand) of education, or anything. But for 50 states, or even 40 states, when you get to those ages, states have very different philosophies about where they want to spend their money, whether they want to do multiple choice testing or open-ended testing. Well, look at how Nebraska was. (It's moving in the other direction — the collapse of Nebraska.) Especially at these ages. Sometimes it might make more sense to be doing assessment along the way. I think Wyoming has a system that assesses at different times. So the philosophies of education, particularly at the lower grades, are a lot different.

A number of states have expressed interest in getting help creating vertical alignment of their standards. We're curious to see how much appetite there is for common grade-by-grade or grade-span by grade-span expectations.

The other thing, of course, is stakes. You can't talk about assessment without talking about stakes. What we might do in a high-stakes environment would be very different from what you do in a lower-stakes environment. Again, especially at K-8. In high school, you can think more about student responsibility and all that. But most of us get softer hearts when we get down to the little ones, saying it's probably the schools and the teachers that are failing, not the students.

Lynn Olson:

So, Phoebe, are you arguing for a set of common national standards, but then states having different approached to the assessment?

Phoebe Winter:

I don't know what I'm arguing for. I'd like to hear some discussion around what people think about what you do in K-8. I'm not including just 3-8, because K-2 is becoming more and more standardized, or standard.

Lynn Olson:

Mike, you're moving back down the line, so why don't you talk a little about what the K-8 program is?

Michael Cohen:

We have taken the end of high school standards in math and English and back-mapped them. In math it's grade by grade down to kindergarten. In English language arts it's actually grade spans down to 4th grade, I think. There are a number of states that have basically expressed interest in getting some help from us and creating some kind of vertical alignment of their standards. We're about to begin some work to help states do that. In fact we're

curious to see how much appetite there is among the states for having common grade-by-grade or grade-span by grade-span expectations. We'll pull a group of states together to let them work on their own standards in the company of others, so that they can have those cross-state conversations and see where they end up.

For us it's really important that the state focus on what it thinks its standards ought to be and that the commonality is something of a byproduct of that. Because if it's not, if you start out saying let's all come together and make sure you have the same standards at the end of the day, unless it's three tiny New England states who are trying to save money on testing, the odds of getting there are pretty slim.

So we don't know what the appetite will be. If there is an appetite for common standards then we can move from there. The common tests will require agreement on the purposes of the tests, the formats of the tests that is all things that you mentioned and then you don't know how many states you'll wind up with at that point.

Cynthia Schmeiser:

Cynthia Schmeiser

Just a quick note, again, on some of the points that I was thinking about relating to the standards and assessment issue. There are a number of states that have also adopted other approaches. For instance, there are 14 states that clearly have adopted college readiness standards that are related to a longitudinal system that we offer. But some of those same states are also ADP states.

It's going to take all of us. If you line up the college- and work-readiness standards that states are adopting today, whether it's ACT or College Board or ADP, there is a whole lot of commonality. But I do think we've got to look at a more national collaborative.

I guess my point is, if we're really going to get this done and that is if we're going to adopt a policy position of college- and work-readiness for high school graduates, it's going to take all of us. It's going to take all of us, in multiple approaches, to get to college and work readiness. And I'm not so sure they're so diverse. If you line up the college- and work-readiness standards that states are adopting today, whether it's ACT or College Board or ADP, there is a whole lot of commonality. We're not talking apples and oranges here. But I do think we've got to look at a more national collaborative, and start talking together and moving forward together, in order to get the uniformity that we want.

I would also say that we need to take a step back and start thinking about what is the objective evidence of our success. Because if the standards are going to be judgmental to a large part, then where's the proof in the pudding that, in fact, if we educate students to those standards that they will be college- and work-ready? We have to bring data as criteria of our success into this conversation now, lest we find ourselves focusing on standards, and we prepare kids to those standards, and guess what? We're where we are today. They're still not ready for college and work. So I think we have to bring up data-based criteria that we're going to use to judge our success.

The last thing I want to say — and I know I'm an assessment-trained person — but near as I can tell in the few years that I have been in this business, nary a test has ever really changed a huge policy. So we've got to think about alignment through the system and really focus on the hard question: How are we going to increase the number of kids in this country who graduate high school ready for college and work? That is the real issue. The hard questions around this are curriculum, instruction, teacher support, and so forth. The test itself, yeah, it'll take our temperature and tell us where we are, but how are we going to get at the real hard work?

The NAEP Model

Thomas Toch:

As someone who has endured England's A-levels, I can attest to John's description of them. They are grueling, day-long exams in each subject based on a year or two years' curriculum. The standards implicit in them — actually, they're quite explicit — are just so much higher than anything we aspire to here. It's striking. They are much more difficult than the AP exams.

With that said, I do think that standards are very important, and I think it's appropriate that we do discuss them. Think of NCLB. The standards there are designed to create a floor. There's lots of rhetoric in the law about lifting the ceiling, but in fact the law is merely designed to raise the floor for some percentage of our lower-achieving students. So if we want to make good on the rhetoric of the standards movement, then we really do need to set the bar higher and find ways to allow students across the achievement range to get there, to move up. So while I think that backward mapping from high school is potentially a very good idea, in the short term, we've got an accountability system that deals with 98 percent of the kids. I'm not sure what the enrollments in the Algebra II are, but they're a little smaller, so far.

We really do need to focus on the tests that deal with the largest number of kids and those are the reading and math tests that are given across the states. One way to think about that is, in fact, if we improve the quality of the test, that would send a signal that would have a beneficial effect I think on the standards, on the expectations, and therefore on what goes on in the classrooms. I don't think we can ignore that.

I think NAEP is a good example. NAEP has very good multiple-choice questions, it has short answer, and it has a fair number of long-form questions. I'm working my way to a point here, and a question for Checker. I do think that NAEP is a model for national standards and the National Assessment Governing Board (NAGB) is a model for a standard-setting mechanism. We have three different levels of standards reflected in that process that is a democratic process, to go to a point made earlier, that involves practitioners at all levels.

Chester E. Finn, Jr.:

But modify it.

Thomas Toch:

It really seems to be a model, so my question is, why not?

Lynn Olson:

Why not NAEP?

Chester E. Finn, Jr.:

Well, there are a couple of answers. One of the reasons that NAEP is good at what it does is because it is only a tool of information. It is not a tool of accountability. There are no stakes attached to it, and it can afford to be pure and relatively demanding in its expectations because nobody's fate is actually attached to it. That is one issue.

The second issue is that when President Clinton tried to attach voluntary national tests to NAGB, things began to crash and burn. Mike can recount that saga in far more grueling, gory detail than I can.

The other thing is that if NAEP were a national test, either it's got to be compulsory for 50 states, or the states that don't take part in it get left out of the entire measurement system. My view is that if there's a national test and, let's say, 40 states end up in it and 10 don't over the next decade, the 40 and the 10 should still get compared on NAEP. That's a very healthy thing. It might turn out that some of the 10 that don't come into the national test are doing a better job and having higher achievement. We wouldn't know that if there weren't a common metric external to them all, which I think is NAEP'S role. I really don't see any Congress elected in the 21st century doing mandatory national testing for 50 states whether they want it or not. I just can't imagine how that would happen.

Thomas Toch:

Well, you could make it voluntary and you could give states a strong incentive to participate.

Chester E. Finn, Jr.:

And then what happens to the ones that don't? Are they just noncomparable because we've used NAEP for the 37 that do?

Thomas Toch:

Well, you could set up a parallel system. I guess when I referred to NAEP, I was referring to the feasibility of a government-supported independent body that has pretty wide credibility, and therefore a mechanism that could indeed create very strong incentives for states to participate. It would be strong, both sticks and carrots.

Michael Cohen:

It turned out in 1997, 1998, whenever it was, when the prospect of NAGB overseeing the development of tests (just for individuals — whether they counted or not was another matter), the credibility of NAGB, as it was constituted and appointed then, began to go down. Part of the political debate — if you remember, Checker — was, "Who's on NAGB?" And once we got that solved, "Who gets to appoint them?" I mean, the fights over who's in charge inevitably — because it's the same fight over NESIC. I can't see a mechanism that has the federal government overseeing or creating any of this without having truly mind-numbing fights over who's in charge that are utterly divorced from what's actually going to happen. It just seems built into the DNA of at least this part of the federal government.

I can't see a mechanism that has the federal government overseeing or creating any of this without having truly mind-numbing fights over who's in charge.

Chester E. Finn, Jr.:

And the very next event is then they'll have to have Senate confirmation for sure. Right now they don't. Just picture that process. Part of what we're dealing with here, unfortunately, is the larger breakdown in American governance. It goes way beyond education and leads to paralysis on so many fronts right now. I'm with Mike, I just don't think it can work without some chaos and catastrophe following, and I'm a great fan of NAGB; I helped invent it.

Gordon Ambach:

I wasn't going to say anything about NAGB, but because of this exchange, I think it's important to note that both Mike and Checker are correct in saying that as soon as the stakes are raised for NAEP (for example by using NAEP for individual testing instead of sampling), there will be all kinds of political pressures on NAEP that could destroy the importance of NAEP trend lines and its integrity.

If any persons had predicted in 1990 or 1985 what happened with No Child Left Behind, they would have been considered nuts. The extent to which there has been "government creep," if you will, a national government assertion over vital issues of elementary and secondary education that has occurred in No Child Left Behind, just could not have been imagined. And yet, they did happen. (In the interest of full disclosure, I supported NCLB. The American Federation of Teachers and the Council of Chief State School Offi-

cers were the two education organizations that supported the accountability provisions of the legislation, because we had very good reasons to support the whole package.)

I don't discount what might happen in the 21st century. Now this is the cautionary note: somebody creates the tools for comparing standards or performance for one purpose. Those tools can be adapted or adopted by a federally organized agency for different uses and purposes. It's like having a toolbox. If the box has a tool that is respected as being something that could be used to get comparisons across the states on the issues that we've talked about, the temptation is for someone to say, "Oh, that's been successful, why don't we have the federal government pick that up?" That might happen, and the fact that there is a history of federal legislation picking up various tools and making them requirements should put up a caution. I think this is a very difficult prospect for us to consider throughout the day in the context of what structure of federalism is desirable.

21st Century Tools for Making Standards Work in the Classroom

Eva Baker:

I have so many good lines I will now not give you because we don't have time. But my real concern, I think, is that we are talking about procedures — that is standards, assessments, curriculum, professional development — at various levels of granularity. We understand that consensus gets built at the highest level of generality, typically.

There are tools out there that would allow consensus among states.

So, to link a little bit to what Gordon said, I think there are sets of tools that are available, but not in use, that would change the way we think about things, so that we are acting like we're in the 21st century and we're not repeating Ralph Tyler's work in the 1940s. Those involve not just standards, but deeper concerns about learning and how and what it takes for students at various levels of expertise, or lack thereof, to get to the next place. I don't care whether the teachers are going to help them do it. You can have technology help through the ether. I don't care. But I do believe that there are tools out there that can help us explicate the domains of interest at a very granular level that would allow consensus among states — not in terms of how many got to a certain point on the same scale, but how many have different features of these knowledge pieces or skill-based whatever you want to call them (content ontologies are what they are called), but how much a New York kid on their test has these kinds of things represented.

Because the current approach to mapping alignment between tests and standards is woefully global. That's because the policy makers, test publishers, and teachers who have to use them don't have the tolerance for greater levels of detail. To make standards work in an instructional or learning environment, one really does need to get down to the actual level where learning takes place, the students and what they see and do. The discussion and action cannot hover at the levels of school organization and governmental organization.

Lynn Olson:

We may come back to that this afternoon when we talk about tests more.

Opportunity-to-Learn Standards

Gordon Ambach:

I wanted to come back to your point, Lynn, about curriculum and the point made by others here about resources. Just to provide a little historic reminder, the development of standards, particularly at the national level, came about after 1989 and the determination in the national goals that the U.S. should be number one in the world in everything in education by 2000.

Chester E. Finn, Jr.:

Math and science. We were going to be second in art.

The extent to which there has been "government creep," if you will, a national government assertion over vital issues of elementary and secondary education that has occurred in No Child Left Behind, just could not have been imagined.

Gordon Ambach:

How many of you remember the national celebration in 2000? There wasn't a single thing done in the U.S.Congress. There wasn't a single thing that the president or anybody else did. It was gone. But, the important result from that 1989 summit on goals was the boost it gave to the concept of standards. At that time, for the most part states didn't use that word; they used "curriculum guides" or other terms.

Now, in the days of NESIC and the National Council on Education Standards and Testing, the concept of standards had two parts. One was student achievement and the other was the good old "opportunity-to-learn standards" — the naughty term that got pushed under the rug as fast as some could get it there. In fact, the main debate at that point was, "Should you couple these together?" The resolution was, with South Carolina Governor and Co-Chair Carroll Campbell pushing on it, "The academic standards for students should be developed at the national level, whether voluntary or compulsory, and federal action here should be voluntary. But opportunity to learn is the business of the states and the localities."

So this created a bifurcation in the concept of what we should really be getting at with standards, namely what performance is desired and how much will it cost to get there. You bundle the whole business of having standards to spur improvement in order to raise student achievement. The debate after the bifurcation narrowed to how much mileage the education system could gain just by ratcheting up the academic standards and the tests." What has been lost is estimates of the reality of, "How much does it take in increased opportunity to learn to get desired increments in the performance on achievement?"

My question here is to both Mike and Checker: Where do opportunity-to-learn standards come in now with respect to what you're doing in Achieve, or in any of your designs,

Checker, by way of how to make sure that there is a specific burden on the school, the district, the state, and the national government in addition to the students?

One last point on the findings of state comparisons. We knew, when attention to comparisons was started, that Mississippi and D.C. were at the bottom and that students in states close to Canada were at the top. One improvement strategy was to provide incentive for students to move close to the Canadian border. We're refining comparisons, but it isn't changing the dynamic of the scores.

Lynn Olson:

We've managed to actually talk very little about No Child Left Behind this morning, which has been a great relief. I do want to spend a few minutes toward the end talking about, given these structures we're struggling toward, how do we need to change that context to make any of this more or less possible?

Michael Cohen:

Just briefly, in our conversations in Achieve we just haven't heard the word opportunity-to-learn standards for a long time. We've forgotten about them — not the issues there, but the idea of standards. We did talk about what it will take to create the tools and supports that are necessary to help kids get over the higher bar. But to be perfectly honest, an organization like Achieve doesn't have the capacity to take that one on. So, go back to the earlier conversation, it's the state's responsibility, and it's fulfilled to varying levels across the states. We don't have a mechanism for getting the federal government to play a bigger role in that.

Michael Cohen

As long as there are states that can look at you with a straight face and say 87 percent of the kids in the state are proficient today, there's mighty little pressure on anybody to deliver a better education.

Chester E. Finn, Jr.:

I was pretty much limiting myself to what I thought was today's topic, which is standards and accountability systems — not the delivery system by which we teach kids so that the results on those standards and accountability systems improve. I think that's an all-week topic, not a six-hour, one-day topic. I think you'll have some radically different ideas in the room as to what are the best strategies for altering the delivery system to get there. I just think that, as long as there are states that can look at you with a straight face and say 87 percent of the kids in the state are proficient today, there's mighty little pressure on anybody to deliver a better education.

A Competition Underwritten by Foundations

Susan Traiman:

Just a few quick things and then a question. Despite Checker's example, we still have a substantial percentage of students who can't attain today's dumbed-down standards, and we have a substantial number of educators who say it's impossible to get them to the dumbed-down standard by any certain date. So we have to think about that as we develop the new system.

Second, I think we need some clarity about what we're aspiring into. I was really struck by an opportunity to meet the Minister of Education from Singapore. He said that Singapore's great accomplishment is a high level of average performance.The U.S. has peaks and valleys. So what Singapore's trying to do in education reform is to maintain a high average performance, but get peaks without getting valleys. We are trying to get a higher level of average performance while I assume still maintaining peaks. But there is still this dishonesty as we talk about the standards because it's in the context about the percentages we get at each level. So in any new models we develop, which was our topic today, how do we think about what percentage are we trying to get where?

Private funding for three, four, or five development exercises, each with a deadline, by which time standards and aligned tests would be developed, and then see who wants to use which ones — I think that'd be swell.

And then, from the *Fiddler on the Roof* line, "If I Were a Rich Man." If I could wave a magic wand and not do this in a federal way, I think I'd use the model that the federal government does use when it's going to build a new bomb or a new nuclear submarine, where very substantial resources are given to three or four competitors. Once they get pretty far down the line then they come and only one goes forward. If we could get some of our major foundations to say, "We're going to underwrite four groups to go forward, and then it's up to the marketplace to determine which one advances," because right now the marketplace doesn't work. ACT can't go forward without knowing somebody's going to purchase it. The same for the College Board or any of the textbook publishers. But their incentives — not College Board and ACT, but the textbook publishers — are to achieve consensus and become more mediocre, in essence.

So I guess my question to you is, Mike, you've said five or six times, "Achieve doesn't have the capacity." If Achieve and a number of other groups got more capacity, could this whole process be accelerated outside the federal government?

Michael Cohen:

Probably. When we designed the Algebra II test, we basically put out a request for proposal (RFP). Here are the specifications, we're looking for someone or someones to bid on this, then we'll pick a winner. If we, number one, did the same thing for a variety of instructional tools and did it in a way that you suggested that we may pick more than one winner, we could figure out a procurement mechanism for that. That would probably speed things up. It is frankly not a role that those who founded or govern Achieve thought we were well

suited for, and I'm not sure we're the right organization to do that. But if that were done I think we'd be in a different spot five years from now.

Chester E. Finn, Jr.:

As with the multiple exam boards in England, this is a perfectly legitimate approach. Private funding for three, four, or five development exercises, each with a deadline five years from now, by which time standards, and aligned tests, grade by grade, in at least core subjects, would be developed, and then see who wants to use which ones — I think that'd be swell.

Congressional Politics and NCLB

Bethany Little:

I think what Congress will do here is a really interesting question because they're of two minds on this. First of all, Congress loves cheap national policy, and you don't get a whole lot of it. You can't do cheap national farm policy or cheap national transportation policy, but you can do cheap national education policy if you got the standards right — in their minds, at least. So they see ways to play with the standards and get massive national changes without having to invest anything significant. And that's scary because most of us would agree that's not exactly where we're trying to take the system. But there is a fascination there.

Congress loves cheap national policy and you can do cheap national education policy if you got the standards right — in their minds, at least. And that's scary.

I think if you talk to most members of Congress who are looking at this issue — and they're all looking at this issue on some level — most of them privately agree that there's a need, and there are some who are out there even more publicly. It shouldn't go unremarked that two presidential candidates have national standards proposals out there, in different ways. So this is definitely a conversation they're having. Publicly, they're terrified. It's one of those "everyone's racing to be second" situations. Like Congressman Michael Castle saying in a hearing, "I'm not going to use the phrase national standards, but somebody has to start saying it."

And then there's the issue of international competitiveness, which is really weighing very strongly on Congress right now, with its efforts to pass a competitiveness bill, etc.

So then they come to NCLB, and I think the big question to them is: "Is this an opportunity, or is this a reason to dump the national standards issue over the side of the boat before the boat sinks?" There are significant players at the table who think that the national standards issue could sink the NCLB boat and would rather see NCLB move forward than move national standards. So they're putting a lot of eggs in that basket. "Leave national standards aside so we can get NCLB done."

Some interesting leadership from the Education Trust put another option on the table as well, which is around: "Can you use the desire states have for flexibility in the system, com-

bined with the NCLB restrictions, to create a moment of opportunity here?" They suggested that states that adopt a higher college-readiness-aligned set of standards be given an extension on their timeline to get all kids there. The logic is clear. If you have higher standards, you need more time to get kids there. That has really ratcheted up a conversation that's happening mostly internally. There's not as much public debate going on about this, but there is a significant conversation in Congress about, if they take that direction, where should it go? I do think that this whole question of what the model is, this is the right time to have that conversation, because Congress is potentially ready to act or potentially ready to turn tail and run. They need some vision on how to get to national standards if they do decide to go there. You two, I think, are leading a lot of that debate nationally. I'm interested to see, where do you think they'll go, and where do you think they should go?

Michael Cohen:

A couple things. One is, you're right that there's strong sentiment in Congress that getting NCLB reauthorized is hard enough; don't add national standards to the debate. I think they're right, although it's pretty clear that keeping national standards off of the bill hasn't actually accelerated its progress. So it might be time to reassess that, although I just want to underscore things that I think both Checker and I said before: The prospect of the federal government getting involved frontally in this scares the daylights out of me. The odds that they will write the Ed Trust provision in a way that helps states move forward rather than gets in the way itself is in doubt. This is just not something that they are really very good at doing.

This is the right time to have that conversation, because Congress is potentially ready to act or potentially ready to turn tail and run. They need some vision on how to get to national standards if they do decide to go there.

Chester E. Finn, Jr.:

As for the political prognostication, I have no idea what they're going to do. I know what they ought to do. That's clear. The biggest error in NCLB, in retrospect, is clear. In modern management terms, it's that they ended up being tight with respect to means and loose with respect to ends. That ought to be reversed, and NCLB ought to give states enormous flexibility with respect to how they deliver education and how they intervene where it's not being successfully delivered. But all ought to be measured against a common metric of standards by which everybody can be compared.

Lynn Olson:

Except we don't want the feds to set the standards.

Chester E. Finn, Jr.:

Correct. Now you've got it perfectly summed up as to what needs to happen. You've got the Rockefeller Institute's agenda for the next three years exactly summarized.

Where Do We Get the Money?

Richard Nathan:

You said a minute ago, Lynn — and I'm thinking of institutional invention and maybe getting a little beyond what this panel should do — where do we get the money? So with the little bit of time that remains, maybe Checker and Mike would say where do we get the money? There's a lot to do here.

Chester Finn:

Well, the development work ought to be privately financed by a consortium of private foundations putting together a hundred-million-dollar kitty or whatever is the requisite sum for getting this done in more than one place over the next five years. I don't have any idea what the relevant sum is, but it's not billions. It's millions. And therefore, it is fundable by the private sector at the development stage. Once developed, then we see what the appropriate role of federal funding may be in its implementation.

The biggest error in NCLB, in retrospect, is clear. In modern management terms, it's that they ended up being tight with respect to means and loose with respect to ends.

Susan Traiman:

And the specs could be written in different ways. Some of the specs could be written to develop a system that could be administered via the web. There are a whole variety of ways that potentially, then, could keep implementation costs down, but raise a whole lot of other equity issues.

Michael Cohen:

You know, the implementation costs, when you step back, are really not that high. People complain that testing is expensive. It is a tiny fraction of a percent of per-pupil expenditures for a tool that…

John Merrow:

That is because they're cheap, lousy tests.

Michael Cohen:

But, you know what? If you built better tests, they would still be cheap compared to per pupil expenditures.

Chester Finn:

I think the tests are about $85 a head right now per subject.

John Merrow:

Hartz spends ten times more testing flea powder and kitty litter than we spend testing kids with NCLB.

Michael Cohen:

And if you spent ten times more, you'd finally be getting to a real amount of money. If you doubled or tripled what we're spending on testing, as a fraction of per pupil expenditures, it's still pretty tiny. And while that's a hard argument to make to state legislatures, for a tool that we've been relying on for 30 years as a tool for improving instruction, you'd think we could at least make sure we have a high-quality tool. And the cost of doing that, again, compared to the cost of educating, is tiny.

From left: Mitchell Chester, David Abrams, John Merrow, Robert Linn

If you doubled or tripled what we're spending on testing, as a fraction of per pupil expenditures, it's still tiny.

Federal Grant Programs; National Professional Organizations

Bryon Gordon:

I almost hesitate to bring this up. I think there is a federal example that relates to what we're talking about here today, and that's the Reading First program. And I'm hesitating more because of how it played out in implementation. But when you look at how the federal government actually got involved in the content area of the subject, it was years of building consensus in the research community, and eventually the national reading panel.

I guess my question would be, is there a way to approach that in the area of standards and assessment where — obviously, I think there's pretty much consensus here that you don't want the federal government to be writing those — you can abstract from those to some good practices that the federal government can adapt? Where the federal government is effective is either coming up with the money or attaching conditions to the money that is doled out that can make sure that those practices are followed and that states have the incentive to follow those practices?

Lynn Olson:

So is Reading First a model?

Chester E. Finn, Jr.:

Keep in mind that NCLB is an optional grants program that no state or district has to subscribe to. All that happens if you don't subscribe to it is you don't get the money. If you don't play by the rules, you don't get money. But nothing bad happens to you, and you're not held up in ignominy for comparison purposes and things like that.

Also, I think it's important to note, as you just did, that there was 40 years of pretty good reading research underlying this. I don't think the same can be said about standards and assessment today.

Michael Cohen:

There is an important federal role in just funding the research that would lead others to better, clearer, and more rigorous, more useful standards. That would be an important thing for the federal government to do.

From left: Paul Goren, Margaret Goertz

Margaret Goertz:

One of the things that struck me, I think both from reading the paper and listening to the conversation, is nobody's talked about the national professional organizations.

Lynn Olson:

You're talking about the voluntary national standards?

Margaret Goertz:

Right. One of the things that did come out of the 1990s, with federal funding, was the development of some of the voluntary standards and the National Council of Teachers of Mathematics (NCTM), which most states basically picked up. So if you look at the math standards, even though there's variation, you see the categories that come out of NCTM. Less so, obviously, in language arts because of the reading wars, and a little bit in science. So it struck me when we were talking about the multiple awarding bodies, or whatever: Do you see a role for these organizations? Actually, Mike, I'm curious, in the work that Achieve has done as you go from Algebra II and map backward, has there been interface or interaction with NCTM, either at the state level or the national level?

Michael Cohen:

Yeah, there has been, at both the state and the national level with NCTM. And there, we've taken slightly different approaches. Our math experts work on their project, we get their advice on ours. So there's a level of consistency there that I think is important.

One thing we learned in our work — we didn't set out to do this, but if you compare what we've done with states with what happened in the 1990s with the standards that the national organizations developed (you're right, most states have picked up on the national models and did something with it) — what we wound up doing through our alignment work with the ADP benchmarks was much more close work with states. We do very detailed comparisons of their current standards with our benchmarks, and as they work on revisions, similar comparisons. I think what that's led to is a way to using national benchmarks that is much more impactful on what state standards ultimately look like. I don't think any of us had that idea in mind in the 1990s, but it turned out to matter a lot in terms of getting a level of consistency across the states.

Lauress Wise:

It seems like in the session this morning, the common content standards and some commonality — maybe evidence-based performance standards, perhaps multiple ways you could be proficient, etc. — have to come first, or it doesn't make sense to talk about improving the testing, opportunity to learn, opportunity to learn what, etc.

Why wouldn't you want to fund state collaboratives, with federal grants or with foundation money, working in consort with the national professional organizations or Achieve or other groups? Because the buy-in of the states is obviously very critical.

So, just to throw out a model, why wouldn't you want to fund state collaboratives, with federal grants or with foundation money, working in consort with the national professional organizations or Achieve or other groups? Because the buy-in of the states is obviously very critical for this. The thing that the NAGB standards don't have is any real opportunity for states to have a voice in this. So why wouldn't that be the mechanism that you would prefer?

Chester E. Finn, Jr.:

It could be one.

Michael Cohen:

Yeah. It's a competitive grants program. I think that would be a standard form of federal involvement.

Lynn Olson:

So something we might agree on.

Summing Up

Lynn Olson:

I just want to say it sounds like we've had a couple interesting models on the table this morning: multiple examination boards, voluntary state collaboratives, NAEP, and NAGB,

which I heard a little less enthusiasm for, at least up here on the podium. So it will be interesting on pick up again on what some of those models are as we move into the afternoon.

So this has been great, obviously. People still have their hands waving, there's lot of questions going on. I think our time is up, though, but we can continue some of this in the afternoon. Thank you.

Richard Nathan:

Richard P. Nathan

Thank you very much to the panelists and to Lynn. I want to just make a comment that strikes me as I'm listening. The first book I ever wrote at Brookings was on revenue-sharing, anybody remember it? We went out to states and we looked at what was going on.

I got a letter from the president of the League of Women Voters of Sarasota County, Florida, and she said, "I've written a paper about your study of revenue-sharing. Could you send me some information?"

I was very young then, much younger. So feeling flush and having Ford Foundation money, I sent her our Brookings book. She wrote back a letter. She said, "Thank you very much for your book. I'm still confused, but on a higher plane."

I wrote to myself in my notes — and I'm listening and learning, and a lot of these acronyms are still not my acronyms — "Dick, be careful. Don't draw conclusions." This has been very rich.

Second Panel: Intergovernmental Approaches to Testing Oversight

John Merrow:

I am not going to bother introducing these two household words to you. I thought that it was remarkable pushing forward in this morning's session and it was great pleasure to listen to Checker, Mike, Lynn, and some of you, but they also raised some questions that I think have come to the point for all of us to grapple with. Let me ask you two guys, what have you learned here this morning? Dick said it well with that little anecdote, that he's confused at a higher level. That's indeed where I am. How about for you guys?

100 Percent Proficiency by 2014

Robert Linn:

It's part of what is causing this pressure to have dumbed-down standards. If we got rid of this 100 percent proficiency idea, I think that we could have differentiated proficiency standards.

Well, I guess my big takeaway — the one thing I didn't hear that I wanted to hear — was that part of our problem is with the claim that 100 percent of the kids are going to be proficient in 2014.

John Merrow:

You think we'll get there earlier?

Robert Linn:

I think we'll never get to the high individual proficiency standard that the people in this room were talking about. If you compare Singapore's results on international assessments to the NAEP standard, about a quarter of the students are below the proficient level in math. So if that's what's happening in Singapore, which by many other accounts is doing a fabulous job with education, I don't think we're going to get there. And so, it's part of what is causing this pressure to have dumbed-down standards.

John Merrow:

Is this because some politician is setting proficiency standards?

Robert Linn:

Right. Right. So, but then what I did hear, I guess, things that I definitely agree with is having something that is nonfederal, whether it's in Achieve or the chiefs or other mechanisms that were mentioned, as something that could create standards. If we got rid of this

100 percent proficiency idea, I think that we could have differentiated proficiency standards. John (Bishop) mentioned that, in many cases, you have to have five or six gradations. So if you are going to have a test like the AP —

John Merrow:

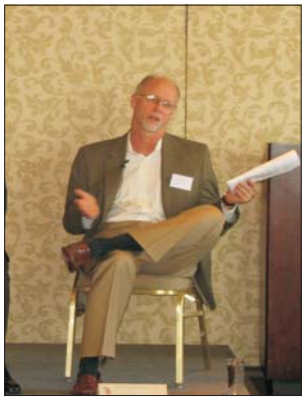
You're talking about the performance standards? Same content standards? Just different performance standards?

Robert Linn:

Well, I think you probably need both differentiated at the high school level. At the elementary, 3-8 standards, I don't think you need that.

Thomas Toch:

I was struck by the consensus in the room — a room full of experts but people with different perspectives, different protocol, information, and the like — on the need to create more common standards. If we juxtapose the starting point of this morning's conversation to where we were five or six years ago, as contemplated when we began to build NCLB, we've come a very long way. And I think it's in part because the No Child Left Behind Act has demonstrated to us the challenges faced by a decentralized system in a political and even economic environment where we are trying to raise standards. The conversation is very different today than it was five or six years ago. The broad acceptance of the notion that we need more centralized standards is striking.



Thomas Toch

What happens is the standards are hard for teachers to make sense of. Teachers look at a test item and say, "Oh, you want the kids to be able to do that? That's what that standard means."

Tests as De Facto Standards

John Merrow:

You know, it's interesting that you say that, but it strikes me that the conversation we are *not* having is that what I see here is a contradiction — Tom, your reports from Education Sector's talk about the spending on No Child Left Behind tests. Fifteen cents out of every $100 goes for these tests. And the tests, don't they in some way become standard-setting?

Robert Linn:

I think they do. In many cases, they have a more prominent role than the standards *per se*, because that's what really counts. That's where the rubber meets the road on how kids

do on the test. So you may have excellent standards and you may go through a process that shows alignment, according to some criteria, of the tests to the standards. But our ways of judging alignment are less than terrific, in my view, and so you may end up with a test that is much less ambitious than the standards are.

John Merrow:

Isn't this a huge contradiction? I mean, is all this talk for naught if these tests are shaping what's actually happening in real schools?

Scott Marion:

What happens is — I think what Lisa was getting at before — the standards are hard for teachers to make sense of. And what they see in the test is a way of instantiating standards, in a way. Teachers look at a test item and say, "Oh, you want the kids to be able to do that? That's what that standard means." In the best set of circumstances, tests could be used to help clarify the standards and translate them into what teachers might want to be doing. That's in the best cases. In other cases, they could result in narrowing and a race to the bottom.

We have placed tremendous importance on tests as a lever of reform. Yet, we've invested very little in the infrastructure that we need to ensure that the tests reflect the high standards that we aspire to.

Obstacles to Higher-Quality Tests

Thomas Toch:

It's true that we have placed tremendous importance on tests as a lever of reform. And yet, as John and Mike suggested this morning, we've invested very little in the infrastructure that we need to ensure that the tests reflect the high standards that we aspire to. What we get as a result is an ironic situation. We see tests focused on recall and restatement of facts, rather than analysis and synthesis of information, and so they calibrate their instruction accordingly, and we get a level of instruction in the classrooms being reduced rather than focused on the high standards that is the aspiration of NCLB.

Susan Traiman:

One of the differences between the U.S. and the other countries we've been talking about is we have a much more litigious society. Do you think some of the reason we have the tests we do is because there are lawsuits and people are going to bring cases against tests that might be graded in a more subjective way?

John Bishop:

The Regents has had essay parts of the exam, and these are published. The tests are now made available to everybody, so you have to recreate the Regents exam three times a year,

so that increases the cost of devising the test. But the real constraint is grading the test, because a quality test cannot be graded just by computers — or at least we don't have the algorithms to grade essays and short-answer questions.

John Merrow:

So it's a money issue.

John Bishop:

It's money and commitment. Grading the Regents exam was one of the responsibilities of the teacher of the course and they would go off in a group of people and grade the exam over a weekend. It was done within a week, and you had the results back in time to hand out the grade at the end of the course. So they had a one-week or two-week window to get the grade from the exam. The teachers did the grading in teams.

John Merrow:

One of the differences between the U.S. and other countries is we have a much more litigious society. Do you think some of the reason we have the tests we do is because there are lawsuits?

So you don't buy the argument that it's because we have a litigious society?

John Bishop:

I think that's a contributor, but I think basically people want the answer quickly. If it's to be done by a large testing organization, that's, I think, the wrong way. I would prefer the teachers do the grading. Here's why. One, teachers are very involved in the grading in other countries. In Canada, teachers do all the grading of their end-of-high school exams. And they are graded over a weekend or just in a short period of time.

Susan Traiman:

John, for their own students?

John Bishop:

No, they get together. Generally, everybody comes to the capital city and all the tests are brought in there and then this group of people does this question and there is more than one person grading the exams.

Chester E. Finn, Jr.:

Like AP scoring.

John Bishop:

Right. Like AP scoring but with the AP, they take forever. There's no reason why you couldn't structure it so the grade could be done more quickly, because Canada does it, and many other countries do it as well.

Thomas Toch:

But NCLB has made that very difficult to do because, as you know, there's a provision in the law that requires the results to be distributed to the school but also to the parents before the start of the next school year, so that parents and the family know whether their kids are eligible for Supplemental Educational Services tutoring and/or placement in another nonfailing school. So, as a result, you get pressure on the front end from state legislators.

Ohio is a good example of this, where the state legislators a couple of years agreed that the test had to be administered in May rather than March to give the educators in the state more time to prepare kids for the test. The legislature also demanded that the results be produced two weeks more quickly than in the past. So you get a compression of the timelines, which creates a very powerful incentive to use test questions that can be scored by machines quickly.

The short answer is no. You don't have enough psychometricians. A more centralized system would allow the existing pool of expertise to focus on building fewer, higher-quality tests instead of being diffused across 50 states.

John Bishop:

Yeah, if you're not willing to spend the money to hire the teachers to spend the time grading them in teams. And you've got to have schools out during this period for at least the teachers of the courses that are being tested.

John Merrow:

So, if the goal is common standards and common tests, the question is what are the obstacles? You're saying one is money for good tests. Is that fair? What about the test makers? Because one of the points Allison made in the paper is that — well she didn't say it, but they're all working for Hartz testing kitty litter because it pays more. In other words, do we have enough psychometricians to create tests?

Robert Linn:

The psychometricians don't do the hard work of grading them. But it's the content experts, really, who do the writing of the items, the hard part.

Thomas Toch:

The short answer is no. You don't have enough. The logic of that problem that takes us to a more centralized system, where you allow the existing pool of expertise to focus on

building fewer, higher-quality tests instead of being diffused across the system where you've got to create 50 different exam systems. You would be much better off, and the logic of this problem is that we centralize.

From left: John Merrow, Robert Linn, Thomas Toch

John Merrow:

We started out with standards, and here we are with tests, because the tests are setting the standards. How do we get back to this question of agreeing on what the standards are? What would a good accountability system look like?

Phoebe Winter:

I think all these things enter in together and interact with each other, and it's a very complex issue. You've got the issue of objectivity: Nobody thinks essays are objective. You've got the issue of money: It costs more to score these things.

We have high policy stakes attached to these assessments, which leads to this idea that we don't want to change anything. States are afraid to change what they are doing, which is a lot of multiple-choice testing.

On the other hand, all the states jumped into doing writing assessment back in the day when there weren't all those stakes attached. That cost money, and that had people talking about objectivity then. I don't know what the answer is, but one of the issues, I do believe, is the stakes that are attached to things. We have these high policy stakes attached to these assessments, which leads us into this idea that we don't want to change anything. States are afraid to change what they are doing, which is a lot of multiple-choice testing. There's outside pressure to keep it to multiple choice or change from relatively good testing programs to ones that are all multiple choice, from the objectivity thing. But if we could rephrase the idea where the stakes are educational stakes as opposed to policy stakes (I'm being Pollyanna, I know), then that kind of leads to the idea that what you have to do is have better assessments and be assessing the constructs in the ways you want our kids to learn them. We have the technology now. I mean, if all our talk here is for change a policy here, change a policy there, and we don't get to the point where the assessments are actually assessing the constructs we think the standards are requiring, then it's just cosmetic.

Maybe there is no way to change the assessment world. And, no, we don't have enough psychometricians. Yes, we do have a lot of content experts out there. But being a reformed psychometrician, I can say that you need to have educated psychometricians who understand the importance of policy and the importance of content.

Thomas Toch:

Underlying your comments is a questioning of the legitimacy of accountability, of holding educators responsible for the performance of the kids. I think that, if there is anything that is widely agreed upon in the current debate around standards and testing, it is that

the notion of local standards is, to a greater or lesser extent, an oxymoron across the country. We need someone to hold the state people responsible for their results.

Phoebe Winter:

I'm not saying that you can't have good assessment and accountability in the same place. Accountability is not going away, and I'm not sure that it should. I think perhaps the way it's structured could go away. But even within the same structure we have now, which is probably going to stay for a while, with these externally imposed tests or externally imposed assessment requirements, we could still do a better job of actually assessing what it is we want kids to know — because, as you said, the tests are the standards.

I live in Virginia, where we're a multiple-choice city, and it has affected my children's education. I've seen it — firsthand, of course, and I know my observation is only anecdotal. But if we don't change the assessments, then all of this is just kind of really good, well intentioned, but won't really get us anywhere.

John Merrow:

We all agree that the test is the signal of "what my kids need to learn," since stakes are attached to it.

So you have to change both. I think this morning they said common standards, common tests, and a common cut score. Just not necessarily a common curriculum. Is that a general agreement?

Phoebe Winter:

No.

Margaret Goertz:

I was going to say, this is really about test use because I'm thinking if we go back to after the Improving America's Schools Act of 1994 (IASA) where you were holding the schools accountable. States like Kentucky or Vermont were doing much more innovative things, and several of you in the room also evaluated those assessment systems. There were some validity issues. I think what pushed everybody into multiple-choice testing was testing every kid grades 3-8 and having to have those tests scored by X date.

So I think if you back up: What would a good accountability system look like? Do you have to test every kid, every year, every grade? And then what assessment can actually give you that information? I would lay the blame on the accountability system at this point, not the assessments.

John Bishop:

But New York State has retained tests with longer answers and problems in the math portions that require multiple steps and you have to write out the answer, despite having to

test every kid, every year. So, it's a matter of commitment to your model of what the instruction is supposed to be about. We all agree that the test is the signal of "what my kids need to learn," since stakes are attached to it, from the school's perspective. If it's answering multiple choice questions only, it leads teachers down a path we don't want them to be going down, and I think you send a very bad message. You trade off the greater effort the teachers will put in because of accountability for a narrowing of what they emphasize, and it's probably cut back on having kids do writing because we are not assessing that.

Federal Role

Thomas Toch:

This year, roughly half the students in the country being assessed under NCLB will never see anything but multiple-choice questions. As someone suggested, it doesn't mean that has to be the way things are. Massachusetts has a fairly sophisticated system with high standards and a range of questions, but the key there is that it costs Massachusetts $30 million a year, and $8 million of that they get from the feds. So the state has made a commitment that many other state lawmakers in other states don't make to standards and the infrastructure you need to be true to those standards.

You trade off the greater effort the teachers will put in because of accountability for a narrowing of what they emphasize.

John Bishop:

So if you really believe in standards and the content of the test is important, the feds should essentially say, "Okay, we'll give you $8 million if you have a multiple-choice test, but if the test meets these specifications, we'll give you more money."

John Merrow:

So there's the federal role: to provide more resources for better, more complicated tests?

John Bishop:

That's one way. It's also incentivizing states to decide to improve the quality of their tests, but leaving it completely up to them. It's a very straightforward thing in terms of the rules that you might establish.

Michael Cohen:

I just want to pick up on a couple of things there. John made most of the points I wanted to make. The discussion about the requirement for a rapid turnaround of tests and tests given at the end of the year: The New York Regents example suggests that turnaround time and giving the test at the end of the year is not necessarily a driver to low-level multiple choice. It has to do with the commitment you make to scoring it.

Look, you took the New York State Regents exam in finals week, with essays and real problems to solve, and you got the results on your report card. Right? You didn't get it the following November. But they did it because, as John pointed out, they had the scoring done locally by teachers, with a verification system, and they produced results.

Thomas Toch:

That proved it can be done.

Michael Cohen:

That's right. My point is it can be done. So the question that I think would be worth focusing on is: If it can be done, why isn't it done more often?

John Merrow:

Which may have to do with trusting teachers.

John Bishop:

No, it has to do with scaling over time.

David Abrams:

A couple of quick points that I've heard that I did want to respond to. One of the things certainly that has me concerned is the discussion of tests and testing families as though they're all the same. I want to separate out your K-8 system and your 9-12 system, particularly from my state, which has a K-8 system and stakes are not student-centric but system-centric, and then your 9-12 system, which in many states obviously you have a student-centric stakes component.

We got to testing very quickly this morning because it's that visual manifestation or representation of these abstract concepts. Our discussion this morning about standards was never really about policy, which standards are, but rather curriculum and performance indicators, which we sort of went to very quickly in terms of the discussion of: What should an Algebra II class look like? What should an English language arts class look like?

Tests can obviously have different types of psychometric architecture and that is one of the reasons why you see differences in turnaround time. Turnaround time in the Regents testing program has nothing to do with some of the issues you're all talking about. It has to do with the various distinct decisions to use a certain type of psychometric architecture that allows the test to be scored, e.g., pre- vs. post-equating.. But there are tradeoffs in any type of decision for testing.

John Merrow:

Do you lose something by insisting on that quick turnaround?

David Abrams:

You know, one of the things about turnaround (and this is something I have to live with on a daily basis) is that everyone wants to test quickly and I constantly say to the board and, of course, the superintendents, "You want your tests accurate. Speed is not the issue in these large scale systems." So we did a set of validity studies to make sure that the decisions that we made in terms of how we cut the Regents exams for a pre-equated model were stable, so your classification accuracy was solid — because after all, these are exams that are tied to diploma conferral.

From left: Robert Brennan, Cynthia Schmeiser, David Abrams

I want to raise one last point and then hand it back to my colleagues here. We keep talking about the tests *per se* as if they stand alone and that every test a student takes in the context of these accountability systems has nothing else associated with it for student decisionmaking. Yet these are all conjunctive systems. Look at any of the states that are sitting in the room. Your tests are conjoined with requirements for seat time, course time. We talked a little bit before about teacher grading, and the issue, I think, that Mike also has alluded with the Algebra II exam: Would it be part of the final exam? How much of that final exam would it part of? How does that contest with your internal grading policies? And all of us know it would be very hard to have a state grading policy. Even when you're in high school, you see grading policies differ from teacher to teacher who teach the same class. I just want to put back out on the table the conjunctive nature of the systems.

I use this example now, and it's accurate, and people chuckle at it. But I'm a former high school English teacher and I taught summer school every year that I taught. I can take everyone in this room this summer around the state of New York and show you kids who passed all their Regents exams — good, bad or indifferent — got through the whole system, but couldn't graduate because they skipped gym, didn't pass their phys ed requirement. So the issue should be: Should we get rid of gym because it's preventing kids from graduating, and it would give us more time for testing? Oh, I'm sorry, strike that from the transcript — my sense of humor. Or do we respect the fact that these are large, complicated systems with various sets of requirements for kids? The testing, in some respects, at least, gives us the opportunity to give the credential meaning at scale and to look at the issues of comparability — which then ties back to external exams like, say, how NAEP, SATs, and ACTs can be used to calibrate.

Thomas Toch:

I guess I would argue in response that NCLB has given us a national case of conjunctivitis.

David Abrams:

There's an antibiotic for that.

Thomas Toch:

In the sense that the signals it sends, that it's causing schools to get and teachers to receive and act on in their classrooms are not the signals that we would hope that they would get. It's just not incentivizing the kind of instruction that we all here agree we should have. To answer Mike's question, the reason why the states aren't doing as John [Bishop] suggests, the reason they're not taking more time and committing more resources is because they don't have incentives to.

John Merrow:

You're suggesting that No Child Left Behind is an obstacle to a high, demanding national or common standard?

Thomas Toch:

I think I am making that argument.

The reason why the states aren't taking more time and committing more resources is because they don't have incentives to.

John Merrow:

Mike, do you think so too? What's your reaction?

Mike Cohen:

Yes. Well, NCLB is an obstacle to common high standards. You could have common low standards, but it wouldn't be a preferred move on my part.

John Merrow:

But No Child Left Behind is an obstacle to high, demanding standards?

John Bishop:

It takes only a small tweak to change that.

Predictive Validity Studies

John Merrow:

I want to go back the question I think that Cindy asked this morning: Does the American Diploma Project have any data that indicates that these standards bear any relationship to kids' readiness to do college work or go out to live in the workforce?

I mean, my litmus test for this is, it seems to me most parents have a common sense question. "What's my kid going to be like when he's 24?" That's really what they want to know. And is there anything that you're doing in the schools that bears any relationship to making him or her capable of earning a living, being a good parent, etc.? Do we have any data that the ADP stuff has any connection at all to what's important?

Michael Cohen:

Sure. First of all, what we don't have yet are predictive validity studies because it takes standards and measures, and then you have to follow kids. We will have that with the Algebra II test over time.

But what we know, at least, is that the standards that have been developed reflect the prerequisite skills for doing college-level work. That's pretty clear. It's clear from the analysis we did in the original states. It's also clear when 15 to 20 states have gone through their own conversations, with their own data and their own higher-ed systems, looking at the skills and the ADP benchmarks and what they know about what it takes to do college-level work. There's some independent confirmation of that.

Is it as tight as it could be? No. But it's tighter than we have for any other standards that states have been setting right now.

John Merrow:

Are you comfortable with that answer? Is enough research being done?

Cynthia Schmeiser:

I think the research that Mike referred to being envisioned in the future, where you're actually looking at performance of kids who've taken the Algebra II and how did they end up doing when they got to college algebra, is important research, you bet.

The *Test Standards* and Test Use

Wayne Camara:

I wanted to actually talk a little bit about some of the concepts that Allison mentioned in the paper. I know we're really up here in the stratosphere with some policy issues. But one of the issues that Allison raised in the paper was really about the *Standards for Educational and Psychological Testing* (the *Test Standards*) that are jointly developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (NCME). We're talking primarily about accountability. I would note that in educational instruction, I don't know if I could improve the policy that much. But I do think that, in terms of some of the technical problems of tests

and test use, there's a lot of room for improvement. As professionals, we may not have any influence over NCLB, but I think we can probably, as a group, improve the testing, and indirectly improve policy.

There's one example I wanted to give you. (I do agree with what the paper said.) I went to a state where they use the SAT for statewide accountability, and I was never asked so many questions about predictive validity studies in my life! That was great, and I just said, "Well, you know, a lot of the stuff you want, we've done, and a lot of it we don't have. Can you show me any studies you've done on the tests you've been using for the past decade?" And the answer was no, but they wanted it from us.

In terms of accepted standards, however, I do want to say that I think, over time, what we have to do as a profession is we have to take back our tests — how the tests are used. I would not have said this 10 years ago. But I work in a testing organization, and I think the policy in NCLB has forced us into lowering the quality of how tests are used.

The proposition I put out there is: With the next revision of the *Standards*, I very much would like to see us come up with five, six, or seven standards that are mandatory. They're mandatory for all types of tests — whether they're certification, employment, clinical, or educational — and they work for both computers and paper.

An example of one can simply be that if you're going to develop a test, that you have to be willing to explicitly state the appropriate purpose of the test, and upon request you have to provide some evidence supporting the validity. Whether we agree that there's validity evidence there is another issue.

There are tests out there right now that will not tell you why they're out there. You cannot get the publisher or the state to define the appropriate use, nor to define the inappropriate uses. When the Buros Institute goes to them and asks for validity studies, their response is that they're proprietary and they're not going to share them. So at a minimum, I think, a step that could affect accountability is to have three to four to five standards. I use the 95 percent rule: 95 percent of psychometricians and professionals will agree that they apply and those would be mandatory. You don't have to meet them, but if you don't meet them, you cannot claim to be meeting the standard for educational and psychological testing.

Lynn Olson:

What's the enforcement mechanism?

Wayne Camara:

Well, one enforcement mechanism is every RFP that I've seen says, "The publisher must meet these standards." Now states are saying we must meet these standards. If we come up with these standards, the Buros Intitute is very willing in reviewing the SAT for example: "Wayne, you don't have any statement on your Web site that tells me the purpose that the SAT's supposed to be used for. You have not met the minimum requirements."

John Merrow:

So is this a call for an organizational action?

Wayne Camara:

Well, I think it's just a call for us as professionals to look at ways that we can go back and use the *Standards* not in an enforcement manner/mechanism, but so they prevent everybody out there from saying, "We meet the *Standards*" and it's a Rorschach test.

John Merrow:

The tests are being misused, and who's complicit here — the state education departments and the test makers?

Wayne Camara:

I'll take some share as the developer of tests that we haven't done enough. But I think we've allowed the test users to take our products without enough control. We need to exert, as a profession, a little bit more control on this.

I think we've allowed the test users to take our products without enough control. We need to exert, as a profession, a little bit more control on this.

John Merrow:

How would the Association of Test Publishers react to this?

William G. Harris:

Obviously I agree with Wayne that there certainly is a need for some type of self-regulation, from the standpoint of being able to say that test publishers and test users have communicated well enough to know that the tests are going to be used for the purpose for which it's intended. That's what they say. Obviously, we can't regulate it from the standpoint of once they violate it, being able to provide for some type of penalty.

John Merrow:

But do you agree the tests are in effect being misused?

William G. Harris:

In some cases, yes.

John Bishop:

But the misuse is not under the control of the people who develop the test necessarily.

Lynn Olson:

But they don't back out of the contract.

John Bishop:

That's right. Remember, tests have multiple uses. In the context of accountability testing, what you want is a test that measures what you want kids to learn, regardless of whether it validly predicts. Achieve went through an extra hoop. But if the state decides people should read *On Liberty* and be able to talk about it, then you have a test where you ask some questions about that. That's a valid test. It measures what the state said that they wanted kids to learn.

NCLB Peer Review

Phoebe Winter:

Phoebe Winter

To some extent, for accountability systems, the peer review standards that are used to guide peer reviews of the states' assessment and accountability systems have already done that. Now, the degree to which they are enforced, and how they're enforced, and whether those are the right ones to be assessed, is a whole different set of questions. But, for example, it says the state must have a purpose for assessment and specify what they are. They say that you have to have validity evidence, consequential validity evidence, all the good reliability, comparability, equating.

Scott Marion:

Susan, you know that that hasn't been attended to.

Phoebe Winter:

It's been attended to. Whether it's been attended to well is another matter. As much as I am not a fan of the testing system required by NCLB, one of the things it has done has made states pay attention to this. State agencies that would never before have thought of this — that have just gone and bought the Stanford Achievement Test or whatever Pearson happened to bring in and said was aligned — are at least thinking about technical quality — and validity, even more importantly. I am not saying it's good, but it's better than it used to be.

Wayne Camara:

I don't think state tests would pass muster on independent audit. Almost every state test is giving subtests on factors like geometry and algebra on uni-dimensional tests, when those two factors are correlated at .85.

Phoebe Winter:

Oh sure, no, I'm not saying they pass. I'm saying that the feds have outlined what they're saying, to some extent from the *Standards*, what the important standards are. And they do require standard errors around the subscores — theoretically, not that any state does it.

Robert Linn:

The peer review process has caused states to do a whole lot of work that they weren't doing before. The documentation the typical state submits is voluminous. Most are smart enough to put it on a CD, and then I can get to it a little bit faster, I can even search it.

Part of the problem I have with the peer review process is that the measurement is concerned with process as opposed to the end quality. That's particularly true with the standards. If you have the right groups of people there developing the content standards, that passes muster. So it doesn't really look at whether or not those are good content standards — not that we'd all agree on whether they're good — but there's no quality on that side of things.

Part of the problem I have with the peer review process is that the measurement is concerned with process as opposed to the end quality.

On the tests themselves, they do require a bunch of evidence on various aspects of validity. In response to those requirements, states are doing things that they weren't doing before. They've independently contracted for alignment studies and the like. As I kind of implied earlier, while I have some difficulty with the trustworthiness or the depth of the alignment judgments, it's nevertheless of better quality.

Thomas Toch:

The peer review process is process-oriented. It doesn't really measure how good the test questions are that come out of the process. There is no auditing function.

John Merrow:

What do they do in peer review if they don't audit?

Thomas Toch:

There are others who know more about this than I. They make sure that there is an appropriate mechanism in place to ensure that standards are aligned to state tests.

Phoebe Winter:

For example, if you have a multiple choice test and the reliability coefficient's 0.75 — the response to the state on my teams (I don't know about anybody else's) will be, "This reliability is unacceptable," or, "Why is it so low and what are you going to do about it?" Or something like that. So it's not just process that they measured the reliability, it's that things seem reasonable. But I just wanted to make sure people understand that it's not just process.

Thomas Toch:

Who enforces the response to your putting those problems on the table?

Phoebe Winter:

I have no idea what happens after we do that.

Thomas Toch:

The U.S. Department of Education (not to be too critical) believes philosophically that this is very much a state process, and it has declined, at many junctures, to engage in this conversation around money or around standards-setting.

Phoebe Winter:

No. I was on a team that was sent to a state when they needed to look to see if they could get computerized adaptive testing to pass muster. So they will send people out. That was only one person, one team. I don't know what they do in other states. I know there's been a response to at least one state.

David Abrams:

I want to say, peer review was very intensive and very invasive in many ways. I can speak to New York State and say that it resulted first in one testing family being entirely rebuilt on the fly, and then two substantial policy changes. So the notion that the technical rigor of the review that peer reviewers went through not having subsequent impact with the states isn't true. In many cases, the assistant secretaries were discussing this with us and the commissioners. You can go state to state and see the types of changes.

Margaret Goertz:

I want to go back to a point that Paul [Goren] and I were having a little side conversation on. This gets us back to NCLB and the reason why these reviews are very process oriented. Remember at NCLB, the feds come in and they cannot approve the content of the standards.

It's very clear they can only approve the process. It's the same thing with the assessments. So, in a way, it's not the fault of the process as much as the context.

So I think if you were to do the separate thing, it would be sort of above and beyond, outside NCLB. Again, you get that tension between not wanting the feds (I'm looking back at Checker) to be making these calls, and what we really would want them to do.

Federal Oversight vs. a *Consumer Reports* Approach

John Merrow:

Tom, is this a place for that Consumer Product Safety Commission-like organization that you talked about in one of your papers?

Thomas Toch:

Yeah. I proposed that we perhaps consider a "National Testing Quality Commission" along the lines of the Consumer Product Safety Commission to perform the auditing function that Wayne suggested we need.

I like the idea of a national organization as long as I'm on the board.

John Merrow:

What is your reaction to Tom's idea of a national organization?

Phoebe Winter:

I like the idea of a national organization as long as I'm on the board.

Chester E. Finn, Jr.:

If there were a few spare psychometricians around, you might contemplate a *Consumer Reports* approach to this for testing and test use that would give greater transparency to some of the tests than is available today. You might get some useful information out there about the quality and legitimate uses of individual tests.

John Merrow:

So is your notion that this consumer agency would actually test the tests like *Consumer Reports*?

Chester E. Finn, Jr.:

The *Consumer Reports* model is different, in my view, from the Food and Drug Administration (FDA) model. The FDA officially says you can use the drug for a specific disease

but not for this other disease, and it's regulated that way. It drives the pharmaceutical industry crazy, of course, but it is a government regulatory function. I was talking about much more of a private sector regulatory function.

John Merrow:

Consumer Reports issues the report that says, you know, "This car does work. This popcorn popper doesn't work."

Robert Linn:

There have been suggestions along those lines, and even some efforts. One of the issues that's faced is: How do you get the data to really do the audit? There's not a clear mechanism for doing that. So that's a problem that needs to be solved for something to work like that.

Lee Cronbach suggested a long time ago that there ought to be a different kind of psychometrician's function, which was an evaluator of the tests and the test uses. But the idea has never gotten beyond that stage, that I know of.

Technical Advisory Committees

John Merrow:

Are there more people who want to react to this?

David Abrams:

On Tom's idea about the oversight board: I remember this discussion last year when Tom's paper first came out and this was widely debated at NCME. I always thought to myself, "Everyone is missing one fundamental aspect of the discussion: states' utilization of technical advisory committees (TACs).

Now, when you work with measurement experts like I do as a measurement administrator, I find that when I meet with my TAC, many times, they'll approach a problem, and there may be a number of differing processes, approaches, or steps that they'll recommend that I can take that (I want to go back to Wayne's point) adhere to and are aligned to the *Test Standards*. As a measurement administrator, which is what I do for a living, I go nowhere without the *Test Standards*. We write nothing without the *Test Standards*. At NCME I asked them, "Yes, please do another version of the *Standards*. Now you've got obviously computer testing, comparability studies, and test use." So you won't get any guff from me on that, just the opposite. We rely on the measurement professionals and researchers to help us make those good decisions.

John Merrow:

So you're saying you're not guilty of what he was talking about earlier? The misuse of tests?

David Abrams:

I have some disagreements on where you get use and misuse, but it goes back to the sweeping generalization — sometimes in the paper and sometimes today — that either all the tests are bad or all of them are good. I mean, the tests vary so much from state to state. I'm still very nervous about, say, a set national test or national set of standards because it's so hard to get that agreement. Every time we start talking about performance, we're not really talking about standards anymore; we are talking about rigor of what we want students to demonstrate for what they know and can do.

I would like for us, as we engage for the rest the day, to think about the roles of the state TACs: those independent technical advisory committees that speak to state measurement administrators, to then go back and advise and counsel their commissioners, secretaries, or state superintendents on how to make those right decisions — because I think, in some respects, Tom, we have infrastructure there already.

I would like for us to think about the state TACs. In some respects, we have infrastructure there already.

John Merrow:

At the state level, you're saying.

David Abrams:

Yes, at the state level. And they're each used differently. I'd ask Bob [Linn] to speak to it. I am so indebted to how smart he is. He's a member of the New York State TAC. So I just wanted to come clean when I ask him to talk about the role of TACs. I am looking around the room and I see some of my psychometric colleagues who I know sit on multiple TACs; maybe some of them can talk to what it is like as they intersect from state to state.

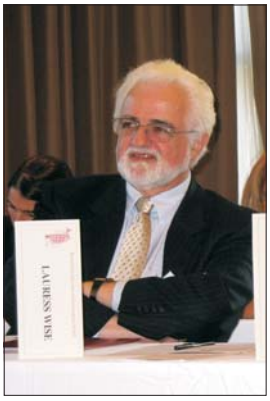
Chester E. Finn, Jr.:

Sitting on TACs is painful.

Margaret Goertz:

I only sit on one, but I want to second what David said. In the state where I happen to sit on one, I'm there as the accountability person, not as the assessment person. All the other people are psychometricians. The state that we advise wants everything in writing from the TAC, so that they can put it in their portfolio to defend what they've done. They will not do something unless they get the opinion of the TAC.

Lauress Wise:



Lauress Wise

I sit on three TACs. We've heard three different models for how you might enforce the *Test Standards* and make sure that the tests and their use are valid and appropriate. One is a federal model, whether it's peer review or the Consumer Product Safety Commission. One is the *Consumer Reports* model; I think Wayne mentioned the Buros Institute is a group that does test reviews already. The third is the TAC model.

If we wanted to pursue the TAC model, I think that there might be opportunities for professional development for TAC members and for some dialogues about the appropriate roles that both the states and the TAC members would engage in — because the TAC is only effective if the state people are willing to ask the right questions and listen to the answers. That might be a model that we could work on, if there were funding.

I just point out that the Center for Assessment, when it first started, had this concept of a "Super TAC," which was sort of this interlocking directorate.

John Merrow:

Explain TACs to me. How does the TAC work in New York or Connecticut?

Scott Marion:

It works differently in different states. The Center for Assessment sits on 15 or so TACs. We coordinate them for about a half a dozen states and then just serve as members on other states' TACs. They operate very differently, as Bob Linn knows, because Bob probably sits on 15 by himself.

We see that, in small states with not great capacity, it's often turned over to the test contractor, and they have no interest whatsoever. I was at a meeting one time and the contractors — the program manager and the psychometrician — were sitting there the whole TAC meeting and didn't take out a pad of paper or a computer. I thought, they're just like "All right, it's under our contract. We've got to bring these guys in and just let them say what they want to say. If the state asks us to do anything, we're going to charge them a couple of hundred thousand dollars."

John Merrow:

This is the fox in the hen house?

Thomas Toch:

Exactly, yeah.

Scott Marion:

When I was in Wyoming, I ran my own TAC, and we actually relied on the TAC to work as this independent agency. One of the things the Center does is, we actually serve as that liaison between the state and the contractor and the TAC, and so we like to think we get better advice that way.

The other thing — Laurie's right — we sort of did these sociograms of TACs. There's the Andy Porter (dean, Graduate School of Education, University of Pennsylvania) TAC, that sometimes includes Roger Trent (former director, Division of Assessment and Evaluation, Ohio Department of Education), and Bob Linn's on a bunch of those, and the Center, and the WestEd ones. So there does need to be some coordination.

John Merrow:

That would argue for something. Maybe it's what Tom proposed, but something which has some larger function.

Scott Marion:

About the Consumer's Union model: You can go buy a Mazda Miata and a Nissan and go out and drive them and do all the things. But when you try to do this in a testing system, you're not going to bring in 10,000 kids and administer the test to them as this independent agency. So you are very reliant on the existing data and the existing use.

If the prime problem is really the state testing systems, then the focus should be, as much as possible, on how do you provide maximum help to improve those systems?

John Merrow:

But somebody needs to ask those questions about the validity of the test.

Design Principles

Gordon Ambach:

From left: John Bishop, Lisa Hansel, Gordon Ambach

I have two observations about principles that might be included as part of a potential future agenda. This conversation swings back and forth between rather general principles about what should be done and then some very specific recommendations about some of the operational aspects of it. I want to focus on general principles.

One principle is that if the prime problem is really the state testing systems, then the focus should be, as much as possible, on how do you provide maximum help to improve those systems?

Tom, with all respect to your suggestion about some kind of consumer board, which in effect is a control board, I think we have to very rigorously deal with the question as to whether we're going to get more progress here out of interventions that will create support and assistance, or those that are designed to basically put on controls.

John Merrow:

So you're saying you don't want common standards? You'd rather improve every state? I'm not sure what you're saying.

Gordon Ambach:

It is not an issue of whether I want common standards. Even if there are common standards, there is a whole array of issues about test use, which is what Peg was getting at a little bit ago. You can have common standards, and maybe even common tests, but then you face a whole bunch of issues within each of the states about how they use them; what's their mix of reliance on teacher evaluation, district, and school evaluation; and the state's evaluation.

We have to deal with the question as to whether we're going to get more progress out of interventions that will create support and assistance, or those that are designed to basically put on controls.

I'm not addressing the issue of common tests. I don't particularly support the idea of trying to develop a whole lot of common tests unless you've got real committed voluntary partners who wish to pool their resources do it.

In 1980, Bill Honig, who was superintendent (of public instruction) in California, and Ralph Turlington, who was the secretary of education in Florida, and I sat together to figure out whether we could make the New York Regents examination system useful to those two states. They were both interested in doing that. It didn't take very long in those conversations before it completely fell apart. It fell apart because of a lot of operational problems, not the least of which is potential student theft of your tests. My worst days as commissioner were in those Regents weeks. There were 26 tests out there, all in locked boxes in every single high school in the whole state, 1,500 locations. Maybe it's more nowadays. And the fear was a breach of security.

David Abrams:

It hasn't changed, commissioner, and you're correct.

Gordon Ambach:

They were in private schools as well as public schools. What you did was you prayed every night that you weren't going to get those calls at four o'clock in the morning from the press that some school had found out that they lost the key or the locked box had been pilfered. Anyway, I don't want to go on about this.

John Merrow:

If the idea is to understand obstacles to the common and high standards and common tests, security is one.

Gordon Ambach:

It is one, especially if you want to try to run a system that is based on course examinations. It's big enough when you're doing 1.5 million tests, which is what we give in any one week. But if you want to spread that to two or three more states, then you can just imagine — different time zones.

If the states do want to go to commonality, to go back to your point, John, I don't have a problem with it. But I would not try to devise strategies where your objective is to get them to commonality. Okay? I don't know why you're puzzled; that's pretty clear. May I make my second point?

John Merrow:

No, stick with your first point. You say you don't think people shouldn't be pushing states toward commonality. I thought it was pretty clear that the system itself is pushing toward commonalities right now. They just happen to be way down here. So the alternative is not commonalities or no commonalities. The reality is, you're moving toward this race to the bottom or race to the lower middle.

If the states do want to go to commonality, I don't have a problem with it. But I would not try to devise strategies where your objective is to get them to commonality.

Gordon Ambach:

What the system is pushing for is solutions to common problems. That's not pushing toward a common testing system.

John Merrow:

But a common low denominator is where you would seem to be headed.

John Bishop:

But it's 50 different tests.

Gordon Ambach:

Look, you cannot get caught up in this notion that if you go national with a test, it's necessarily going to be better and higher in quality than what's out there for most of the states. You cannot. You have to go, I think, on the basis that the standards will move toward reduction toward the middle.

Now, the second point does have to do with part of the context for this discussion. Allison's paper very, very carefully indicates that it's not for or against national standards. It's not a paper for or against No Child Left Behind provisions. It's an attempt to try to outline what our problems have been and how you might devise ways to deal with them. I would observe from the conversation the following: It will not be possible to strengthen the accountability system in this country without significant revision in No Child Left Behind's accountability design.

John Merrow:

In what way?

Gordon Ambach:

Well, I could give you several different ways that have already been mentioned. I could summarize them, but I won't.

All I'm trying to do is to suggest a principle: You cannot deal with a whole series of instrumental fixes on accountability unless you deal with what it is, explicitly, that is being driven now across the country by the formula and the design of No Child Left Behind. Unless there is change on, for example, that the whole target is proficiency, or change on the whole target is just reading and math (that's how schools are now measured in this country — on Adequate Yearly Progress (AYP) and its scale-up to 2014), it will not matter what we do with all of these instrumental designs. You're still going to be left with what's driving the whole system in this country and will through the next reauthorization if we stick where it is. There will be a whole series of things that will, in fact, impede any kind of accomplishments which might otherwise be gained. I suggest it as a principle.

What the system is pushing for is solutions to common problems.

John Merrow:

Okay. So if we're identifying obstacles, that's a very clear one.

Gordon Ambach:

They're principles as to how you try to design an agenda from here.

Have Instruction and Standards Gotten Worse Under NCLB?

Susan Traiman:

I think it's really important that we not overgeneralize and that we are clear where we have information and where we don't. Tom, there isn't one shred of evidence that instruction is worse since No Child Left Behind. There's this myth somehow spread frequently by


Susan Traiman

teacher's organizations that creativity was flourishing in our classrooms before No Child Left Behind when teachers were free to do their own thing. There's no evidence, no research that there's any diminishment in instruction.

Also, in terms of everything going to a low common denominator, Checker's study (*The Proficiency Illusion*) showed enormous variation in test scores between states. So we can't overgeneralize that they've all gone down.

The other thing is: The District of Columbia looked around and said, "Massachusetts is considered the state with the best standards and assessment. We're going to adopt their standard. We're going to adopt the Massachusetts Comprehensive Assessment System." Much higher. Very good. Hasn't made a shred of difference in the D.C. system. Hopefully it will under Michelle Rhee (chancellor of the Washington, D.C., public school system appointed in 2007).

But one of the things we have to think about in creating this new system is: What would be different? You know, when it's ratcheted up and when it's more common, what would be different that would get us to a different outcome? I know that's not what this meeting is about, but it's got to be in the back of our minds, or else why do it?

Thomas Toch:

I'll react. The logic is pretty overwhelming, whether there are studies that get into classrooms and show differences in behavior today versus six months ago. That research is very difficult to do, and there are so many other factors that can influence what's going on in schools. It would be very hard to suggest that, in fact, nothing has changed. But the logic to me is pretty compelling, Susan, that if you focus on low-level skills, teachers are going to do that stuff.

This is only an anecdote. I was at a school the other day where I was given a memo in which the principal said, "We need to focus on these eight kids, because if we get these eight kids up over the hurdle, then we are going to achieve Adequate Yearly Progress." So that kind of rationalization goes on.

Susan Traiman:

But let's just take that school. Before NCLB, they probably never noticed those eight kids — or half the other kids. It's just the implication that it's gotten worse, and there's no evidence of that.

Thomas Toch:

I'm not saying it's gotten worse.

Susan Traiman:

You did say that.

Robert Linn:

There is evidence that there's been more focus on reading and math at the expense of other subjects. And there is evidence that the so-called bubble kids are getting more attention than the high-end kids or the kids that, in some sense, Title I is about in the first place — that is, the ones that are really low — especially in a state that happens to not have succumbed to lowering their standards.

John Merrow:

Gifted programs are being cut.

Robert Linn:

If you look in South Carolina, for example, where they have kept ambitious standards, you've got a lot of kids who are so far below that the teachers know there's no chance of this kid is ever getting over the bar, and so they don't get as much attention.

John Merrow:

No offense intended, but I find that this argument flawed to say, "You can't prove that No Child Left Behind has made things worse." The operative question ought to be: "Is this what we want? Is the way you want schools to run?"

Susan Traiman:

I'm just objecting to Tom's overgeneralizations.

John Merrow:

Let me just finish. Talk with Bill (William L.) Sanders (senior research fellow, University of North Carolina) about the data, about the impact of instructional practices on highly capable, gifted but poor kids, usually minority kids. Essentially, he drills down into this stuff indicating that, pre-No Child Left Behind, quite often the gifted programs were the stimulation these poor kids got. They didn't have parents who took them to Europe or got them pianos. They got all the stuff at school. They're not getting it now. His data say those kids are regressing towards the mean. That's just one piece of evidence.

We surveyed 500 Teachers of the Year, the last ten years' worth of Teachers of the Year. It's not one of your scientific studies, because they either responded or did not. But 170 responded, and a significant number of them talked about, essentially, they would not

go into teaching today, given the conditions they'd be forced to work under. At the Teacher of the Year convention, they spoke up at the press conference about the impact.

I spend a lot of time in schools. You see this drilling, drilling, drilling, and it's particularly with poor kids. It's not a question of pre- or post-NCLB. It's a question of is this the thing you want? Is this getting in the way of creating the kind of school system you want for your own children?

Chester E. Finn, Jr.:

Well, it's fun to sort of demonize NCLB as the source of this. The goal of standards-based reform, which antedates NCLB in many states, was, of course, to do similar things. The question is whether the tradeoff is working. The tradeoff is to have standards, to know whether kids are learning things, and to have an accountability system that rewards success or intervenes in cases of failure. That's the basic theory of standards-based reform.

Back to Front: Theresa Siskind, Paul Goren, Chester Finn, Richard Nathan

Most states embarked upon this before NCLB came along, but did so somewhat differently from each other: different kinds of standards, tests, interventions, and rewards. But my own instinct is that this variety is a net plus.

They're doing what they can with the limited resources they have. They're doing drill and kill because they do not have the resources — either the numbers of people or the knowledge within those people — to teach better and differently.

John Merrow:

If the goal here is to identify obstacles to higher standards and figure out if there are entities that can move that forward?

Chester E. Finn, Jr.:

Then you notice, for example, that a single (proficiency) standard is a problem, and if you're redoing it, you can install three standards in your system.

John Merrow:

What about if, as Gordon says, security is a problem, cheap tests are a problem?

Chester E. Finn, Jr.:

Okay. So then you go list the ten things that need perfecting, that need reworking.

Bethany Little:

I want to just continue to build on what's driving what here. I think the NCLB question and what it's driving is really critical because it does give us some points of commonality across the states. On some level, it should be having similar results across the country — but it's not. That's in part because we do have a federal system, where the state and localities are driving what they do with that law. So it's one thing to have the set ofrequirements in the law. They don't all have to have only cheap tests (see Massachusetts). They don't all have to do drill and kill. They're doing what they can with the limited resources they have. They're doing drill and kill because they do not have the resources — either the numbers of people or the knowledge within those people — to teach better and differently.

So the question is: Is it really that NCLB is driving these things worse, or is it that we're not putting in the other supports in the system to make them better? I think it's that second part, the opportunity to make them better, that NCLB, and hopefully a new generation of it, will drive.

This goes back to the question: Is it driving towards common? I think it is driving towards common, and I think one of the reasons is because of issues of economies of scale. If you don't have to do it 50 times over, you can do it better.

The people writing the No Child Left Behind Act have no idea what the TAC systems do and what the state systems look like.

I think there are also issues of transparency. Right now there was a big discussion here about what the TACs do and how the state systems look. I can tell you, the people writing the No Child Left Behind Act have no idea what the TAC systems do and what the state systems look like, and their actual tests. But there's a curiosity — "Are they good enough? Are the standards high enough? Are the tests good enough?" — being driven by this sudden move towards commonality that I think will actually begin to produce a desire to see this happen.

Going back to the fundamental question here, I think the national test issue is very much on the table, and I'm not sure why we would recreate it 50 times over a little better. I think we're on this path, and I'm not sure why we wouldn't want to be on this path.

John Merrow:

I think a lot of people are saying we should be on that path.

Validating Tests vs. Validating Accountability Systems

Mitchell Chester:

John, they encouraged me to stand, to wave my arms. I actually know the answer to whether NCLB has helped or not, but I'm not going to give it.

Instead, I'm actually going to pick up on one of the last things Susan said and tie it back to what was, at one point, the topic of discussion here: how to help states develop better tests.

Susan's comment was, I thought, pretty provocative about importing a test that is getting very high marks, the Massachusetts test, to another locale. What should we expect the impact of that to be?

So, one of my concerns with the discussion that we started from, do we need some kind of consumer council or outside validation of tests, is that I think there's a difference between validating tests and validating the systems in which they reside. I think it would be really easy to put too much energy and effort into validating tests that might miss the forest for the trees.

Because, in fact, if you take the same test and put it in two different states, two different policy situations, you may get hugely different results. Same exact test. You know, we'll get better data as time goes on from Washington on the extent to which just simply bringing in the test makes a difference.

But I want to illustrate that in another way. My understanding of the data over time from state NAEP suggests that some states have made pretty good progress over time, and some have not necessarily. Two of the states that typically get cited as states that have made pretty good progress, in general as well as in terms of bringing the low students up, are Texas and North Carolina.

There's a difference between validating tests and validating the systems in which they reside.

Susan Traiman:

North Carolina, no more.

Mitchell Chester:

Well, what's interesting is if you work backwards from those studies, if you took the impact of the state system as your judge and if, in fact, you agree that those are two states that made a lot of progress (I don't know if folks are going to agree to that or not), then you might say, "Well what kind of tests did they have? You know, how did those tests play in?"

I don't think there would be too many folks who have watched tests (I'm kind of going back to the earlier generation of tests in those two states) who would argue that those were tests to emulate. They had no open-ended items. They were all multiple choice. Most folks would say they aimed pretty low. If you only looked at the tests, you might assume that those are two states that have it wrong. Yet there seems to be a lot of evidence that, at least for a while, those two states were doing something right.

Lynn Olson:

I want to go back to the topic of the afternoon, intergovernmental approaches to oversight of testing, and to pick up on some things that people said. A lot of the discussion about TACs and about the *Standards* have to do with technical quality assurances within the testing community itself about whether the testing community thinks its tests are technically good enough.

That's a very different level of discussion about testing oversight than what the public needs to know, and might want to know, about the tests that are being used, and doesn't get at issues about should it be a essay test instead of a multiple choice test. It's asking "Is the technical quality of the multiple choice test, okay?" So I think we have to think about that when we talk about what kind of oversight do we want, and of whom.

Then the second question, again, picking up on Gordon's: I think we need a little more discussion about what's the difference between a consumer report card, essentially an accountability mechanism, and a consumer products commission or a National Institute of Standards and Technology (NIST), which also has some capacity-building component. Are we going to get further down the line by chastising states where testing companies that are doing the wrong thing, or by putting some effort into capacity building to do a different or better thing?

My argument is that where the commonality should take place is at agreement on how tests get made. That is, not at a detailed level, but the goals and procedures that are followed to make sure the tests meet the goals that one has for purposes, use, and validity,

John Merrow:

That's a rhetorical question.

Lynn Olson:

Well, no, I think it worth asking. Which are we most in need of now? Because there are two different proposals on the table.

Thomas Toch:

I'd say we need both at the same time.

Lynn Olson:

But should you have one organization that does both?

Eva Baker:

I agree with the goal of commonality at the most general level, but I think that maybe we're focusing on the wrong detail, that is, moving the test around.

If you go back to the *Test Standards* which many of us know quite well, and some of us will know better soon, there is a chapter — Chapter 3 — which is about how to develop tests. That chapter, I think, could be updated, because there is some new learning that follows out of the ways computer systems get developed with reusable components.

So my argument is that where the commonality should take place is at agreement on how tests get made. That is, not at a detailed level, but the goals and procedures that are followed to make sure the tests meet, ultimately, the goals that one has for purposes, use, and validity, rather than just the idea of shipping around tests that are good ones and saying,

"Oh yeah, that works and that doesn't work." That seems to me to be at the wrong level of detail for us.

Secondly, I would think that we all would want to be pushing on the test publishers or others to be more inventive about how they are going to create assessments that will fit 10 years out. Because right now, we're still, I think, way behind the curve compared to the way other big organizations, or small manufacturing organizations, work in anticipating changes. The test publishers, in general (not any in this room, of course), have really repeated in slight refinement what they've been doing for years and years and years. This is about the quality issue of the assessments.

The last point I want to make is that I'm delighted that we're talking about assessments instead of Adequate Yearly Progress. However, I take Bob's comment very seriously that, as long as the Adequate Yearly Progress thing is there, it's going to limit the creativity and exploration that's available for tests and assessment development.

Releasing Test Items, Test Prep, and Teaching to the Test

John Merrow:

Right now, we're still way behind the curve compared to the way other big organizations work in anticipating changes.

You all know so much more about this than I do. Let me ask you this question. I was talking to some people in physics about this idea of testing. Is there something wrong with this idea that you develop, let's say, 2,000 items that test your knowledge of the various parts of physics and you release that test? You say to all the kids and teachers in the country, say, "The physics exam is going to be 17 or 100 or some number of these questions." They're all good. I mean, they change a few numbers. You can't memorize them all. Is there something flawed in that notion? I mean, that's the way the Department of Motor Vehicles (DMV) works. You know, I never thought I'd hold up the DMV as a model.

John Bishop:

Yeah, but it's the same 20 questions always.

John Merrow:

The DMV said, "Here are the things you need to know. We're going to ask you 15 of these 35 questions." Is there some flaw in that?

Phoebe Winter:

Yeah, but that's what standards are supposed to be, though. They're supposed to be those 2,000 questions. You could do that. You could make a fortune.

John Merrow:

Is there something wrong with that, besides putting test publishers out of business?

Phoebe Winter:

To the extent that they were good questions, meaningful and actually had something to do with what you wanted kids to know and had a bunch of different formats and all that stuff, and had rubrics that showed how the responses should be scored, and examples. I think it's a wonderful idea.

John Merrow:

So then there's a role for the Gates Foundation or the Spencer Foundation to take one piece of the curriculum — maybe it's Algebra II, maybe it's fifth grade math — and say, "Go ahead and teach to the test."

Phoebe Winter:

But is that the most efficient way to get where you want to go?

John Merrow:

I don't know.

Phoebe Winter:

I don't think so.

David Abrams:

Don't do that.

Lauress Wise:

This only works, of course, if there's a common set of specs that we're trying to build the test around. If there are 50 different states each with their own frameworks, you're not going to find a common set of items.

John Merrow:

But Algebra II is pretty much Algebra II. That we generally agreed on.

John Bishop:

If each question is of differential difficulty, you need to pretest all those items on a population that is representative. The people who are taking the test would have to be incentivized to try, which is a problem in secondary education. Then, using IRT, you could construct the scale.

John Merrow:

What's IRT?

Chester E. Finn, Jr.:

The subway system in New York!

John Bishop:

Item Response Theory. Basically, you have to find a way of taking a question that is more difficult and using it to create a common scale, so that the test is comparable from year to year.

John Merrow:

Some questions are more difficult. That's always true.

Phoebe Winter:

This is sounding strongly like Northwest Evaluation Association, so I'd prefer not to go down this road.

Scott Marion:

John, I have memories of sitting on my parents' patio in Long Island, New York, studying for the Regents exams and going through the Barron's Test Prep books. I did very well on those exams. They figured out, if you went back far enough, you would hit questions on your exam that were close enough. So the question is: Is that real learning, or is that just test prep? So it's 2,000 items, maybe 10,000 items, and I just have this fear that some of these schools, they're going to start on question one on day one of the year.

John Merrow:

You have to have leadership.

Robert Linn:

How many copies of this SAT are released now, Wayne?

Wayne Camara:

We have seven administrations a year. We release three of them.

Robert Linn:

Right. But the collection of books is probably up to 50 or 100 SATs.

Wayne Camara:

From left: Richard Nathan, Judith Koenig, Wayne Camara, David Shaffer

Yeah, with the new test it's not as many, but there are easily a couple of hundred disclosed real SATs on them. When we got into writing, equating with constructed response is very difficult. So we have a squad of people that meet and ETS said the best way to equate that essay is to develop 200 essays and we release them all on the Web so that everybody knows what they are. They will be pre-equated. And who is going to memorize 200 essays?

Well, guess what? We did a quick survey and we discovered that if we release 200 essays, there are not just a small number of schools that would be memorizing, not the verbatim answers, but rubrics to try to answer 200. We thought, you know, it won't work. To spend that kind of instructional time bringing kids through tenth grade English classes looking at a different essay on the SAT every day would be a horrible disturbance. So I think, unfortunately, if you put 200 or 2,000 items out there, there will be educators and parents who will insist that their kids memorize these.

Scott Marion:

If they did one a week throughout their high school career, they'd get through about 100 essays.

Eva Baker:

If they did, that would be great.

David Abrams:

I don't think it would be a very good idea. I think Tom's anecdote a moment ago that raised the point about leadership was a classic example. We've all seen this — that is, you're a building principal and there are 500 kids in the building and all 500 are unreported, versus you're a building principal and you see eight you have to get over the bar to meet AYP. So you won't get the leadership at scale. You'll get misuse.

What's more important (and I say this as a former teacher and a measurement administrator) is the release of the samples or of exams so people see the tasks, the cognition that you're measuring, the primary content that goes into those cognitive processes. Once again, the sweeping generalization (and I appreciate Susan bringing it up) that all the tests were simplistic and factual recall is not true when you look across all states. No student — undergraduate, K-12, or graduate — should ever walk into an assessment not knowing the structure and the expectations and/or the time signatures. That's part of good instruction. It's not necessarily just drilling and killing the kids. So I agree, you want to put a lot of material out there.

Scott and I had very similar experiences as kids, because I grew up in New York too. I didn't sit in Regents classes in high school just day in and day out doing Regents tests. I did the core. We read, we wrote, we thought, we debated, we argued — which is what we want in our schools for our kids, a good comprehensive education.

I'm not going to put out 2,000 physics items (or any number of items in any discipline) and say, "Here you go, here's where the test is going to come from." What you put out are your test specs. You put out your problem types. You put out your core curriculum.

John Bishop:

And examples of past items. But you are not going to replicate that one over again.

David Abrams:

Yeah, because first of all, consider the confounding issue of a kid who had access and kids who don't. So you already have confounded the results if you release your test questions before you give them and then give them in a stakes-generated environment. So when we go back to litigation, I could ask my TAC how it feels. I could ask Bob and other members. I know how now I feel about this counsel, which is that I wouldn't recommend that approach.

Thomas Toch:

Just for the record, I did not say that all states have multiple-choice only tests. I gave a couple of examples. Massachusetts is a good example. I did suggest that we are trending in that direction, and that's problematic, just for the record. I would ask you, Susan, if you

would rather have your kids in a system of the sort that we've been contemplating here today or in the current system.

Susan Traiman:

There was an interesting piece in yesterday's *The New York Times Sunday Magazine* and it was about preschool. Basically, it said that most people, like the people in this room, want for preschool for their children: the whole child, the play, and being very nurturing. Poor parents, less well-educated, tend to want that they're going to learn their ABCs, etc. Of course, there is a balance between the two, but the more affluent kids tend to be exposed to those verbal skills, the ABCs, without having to go to preschool and so it actually may be a benefit to the poor kids in terms of closing the gap. Researchers will argue back and forth.

The answer to what you are saying is, I really believe that one of the main missions of our schools is to make sure that all kids get a foundation. They need to be at a much higher level than it is now. So I can't say whether I am comfortable in the current system or in the new system. It has to do with whether the educators in that system are smart — in other words, that they understand if you teach a rich curriculum, you don't have to worry about the tests.

John Merrow:

That's a two-prong issue too. If you get a preschooler, then the parents want their kids to learn the ABCs, if the educators know the way to do that with a rich curriculum, as opposed to, "Okay, what is this, everybody, what is this? Quite often, that's a capacity issue.

Susan Traiman:

Right. And that's what I am saying. It's not one or the other. It's educators who understand and principals who let their teachers do that. They don't have such a lack of confidence in their teachers.

John Merrow:

It goes back to what Lynn said also about providing the resources.

Thomas Toch:

That contradicts the assumptions of NCLB. NCLB is predicated on the notion of incentivizing educators to do the right thing.

John Merrow:

And punishing them if they don't.

Thomas Toch:

Exactly. And currently we are incentivizing them to teach to a lower level of expectations. We are setting a floor under which we argue as a nation that we do not want these students to fall. And that's a very different model than when we say we want all students to achieve a high standard.

Susan Traiman:

Or a true growth model like what Bill Sanders talks about, across the whole spectrum — from the bottom moving up, the middle moving up, the top moving up.

John Merrow:

My goal is to try to list as many obstacles to moving in the direction of common standards as possible. So far I've heard security, cheap tests, misuse of tests, an NCLB mentality, a kind of "gotcha" mentality. I think this goal of higher standards for all by 2014, in some ways, gets in the way, at least from what I was hearing this morning. And someone else had said higher education was in fact an obstacle to figuring out how we get to some common standard.

Standards-based reform is really the driving force. NCLB is just an effort to extend or impose that on the states that weren't already doing it.

Who Benefits From the Current System?

John Merrow:

I am wondering who benefits from the current system. The current system seems flawed in so many ways. Generally, in journalism, you follow the money. Who benefits from what we do now? Can anybody take a quick crack at that?

Margaret Goertz:

I'm going to show my age because I've been going out into schools since the late1970s when we had the basic skills movement. I'll take my own home state of New Jersey, which was an early basic skills test state. I bet if I looked at the basic skills test from 1978 and compared it with our current state test, there would be a big difference, and it would be in a positive direction.

So I agree that we have a lot of problems with No Child Left Behind — maybe the ones that Gordon pointed out — and we have to ratchet up. But I think we have been on an upward trajectory over the last 30 years. I think if we looked at the NAEP scores — we all focus on the percent proficient — but if we look at the percent below basic, that has really dropped. So I think, basically, we've been moving up the basic part and now we want to say, let's move up the proficient part. So I just want to make that one point, that we've been sort of two steps forward and one step back.

John Bishop:

Standards-based reform is really the driving force. NCLB is just an effort to extend or impose that on the states that weren't already doing it. When you compare the states that had good standards-based reform systems and had exit exams, those kids learn more.

We complain when there is only a two-point gain on the NAEP; that's two-tenths of a grade level equivalent on the NAEP. If you can get, in just three years or two years, a gain of two-tenths of a grade level equivalent, that's a big accomplishment. The gain in North Carolina and New York is like one and a half to two grade level equivalents, controlling for other things, in math in eighth grade since 1992 to 2003. So there have been huge improvements in math performance, and it tends to be in the states that have the strong accountability systems. So it's working, but not to our satisfaction.

John Merrow:

NAEP scores went up more in the five years before No Child Left Behind than they've gone up since.

There have been huge improvements in math performance, and it tends to be in the states that have the strong accountability systems.

John Bishop:

No, I don't think No Child Left Behind is the thing. It's standards-based reform and exit exams that are doing these things.

Phoebe Winter:

Who really does benefit that I think that we have some information on (I don't know if we have numbers) are students with disabilities and English language learners. It has nothing to do with testing or anything like that. Well, it does to the extent that they're included in the assessment systems, which means states are paying attention to educating them — especially for kids with severe cognitive disabilities.

John Merrow:

So there are clear benefits to students with disabilities.

Phoebe Winter:

Students with disabilities and English language learners, which is a kind of obvious thing. I don't agree necessarily with all the policies, but it has brought these kids in, and it's made educators and testing people think about what it means.

I think psychometricians have benefited, because we have to start thinking about what it means to devise tests that appropriately assess kids coming in with different strengths and weaknesses. I think that's spread across the whole field.

We have to think about teaching English language learners, which makes us think about teaching kids with reading disabilities that makes us think, "How do I get the kid with Ateention Deficit Disorder?" I'm not saying it's perfect or anything. But these kids, I think, have benefited.

Chester E. Finn, Jr.:

Well, the testing industry clearly benefits from having multiple tests, rather than just one.

John Merrow:

You probably already know this: The chief source of the revenue for the Washington Post Company (which also runs *Newsweek*) is Kaplan.

Chester E. Finn, Jr.:

And state bureaucracies benefit from having multiple tests. Every state bureaucracy, department of education, has a testing division.

Scott Marion:

I don't know if Mitch (Chester) and Teri (Siskind) are really benefiting by these additional tests.

Theresa Siskind

Chester E. Finn, Jr.:

No, by having their own, as opposed to being part of a multistate, interstate or international tests.

Lauress Wise:

Only if being bigger and more bureaucratic is better.

Chester E. Finn, Jr.:

Well, a lot of the bureaucrats think it is, present company excluded, of course.

Scott Marion:

But the amount of people that they are adding is far fewer than the proportion of tests they are adding. The workload in those units — this is why we have the quality problems.

John Merrow:

And so, I'm just asking you, is it an economic benefit to testing companies?

Chester E. Finn, Jr.:

I think there's an economic benefit. And publishers that produce the textbooks that accompany the tests and the curriculum in the different places are using different tests.

John Merrow:

Are these oxen that would be gored?

I think the governance entities benefit — not so much from having their own tests but because they benefit from being independent governance agencies.

Chester E. Finn, Jr.:

I think so.

John Merrow:

Should we move in a direction of a common system?

Wayne Camara:

I think we need to look at NCLB though. It was not just publishers who lobbied for this law.

David Abrams:

Actually it cut their profit margins.

John Merrow:

But that isn't the question. The question is who's benefiting from the current system.

Wayne Camara:

Policy makers. As a superintendent or as a governor, I'd much rather have my own tests subject to standards than have them done nationally, because then I can show each subject that my kids are proficient in. If I use the SATs or the ACT, I don't look that different.

Chester E. Finn, Jr.:

There is some political benefit to having your own. You can fiddle with them. You can claim accomplishment on your own terms. There are various things you can do to make it come out the way you want it to when you've got your own. I also think that multiple tests benefit folks who don't necessarily want full transparency about performance by American kids and schools. Because it's so bloody hard to compare and publicize results in anything resembling a uniform metric. So if you don't really want to change your ways, you're better off with a lot of confusing signals in the system.

John Merrow:

Why, you cynic! I never expected that of you, Checker. In all the years that I've known you. Mike, who benefits?

Michael Cohen:

Oh, I agree with what's been said. I think the governance entities benefit — not so much because they necessarily benefit from having their own tests but because they benefit from being independent governance agencies, whether it's 50 states or it's 16,000 local school districts. If you think about the political fights in the 1990s to try to create some kind of national standard and/or test, much of the pushback came from advocates for state or local control. Not just of tests, but of education.

From Left: Mitchell Chester, Michael Cohen

If you think about the political fights in the 1990s to try to create some kind of national standard and/or test, much of the pushback came from advocates for state or local control of education.

Mitch Chester:

NCLB's disaggregation requirements have changed what gets discussed, what gets highlighted in Ohio in very profound ways — around racial/ethnic lines, around poverty status. So I think that's not been trivial. In regard to the kind of common tests and common standards, you know, this may be the catalyst to get us there. Without requiring all states to build their own, we may never have gotten the impetus to even have the conversation.

John Merrow:

So economics could be a part of it too. Final thoughts? I am not sure we've done what we were supposed to do in terms of figuring out the structures that are needed, so you have 30 seconds.

Thomas Toch:

I'd say Mitch's point is a good one because, in effect, he's suggesting rightly that government reacts to problems rather than functioning proactively. NCLB's problems, coupled with our shared commitment to higher standards for all kids, may in fact drive us. That's what's pulling us together here today to find ways to truly leverage improvement in the system through the accountability and assessment systems, and that's a good thing.

Robert Linn:

I would agree with Mitch's point that, in fact, NCLB has had some very positive effects. I know I'm known as a big critic of NCLB, so I should say that, and it's largely around disaggregation.

I also think Peg is right that the tests that we had in the 1970s were less ambitious than the tests we have now. There was a period in between there when they were even more ambitious than they are now, and so it's kind of a curve.

I think maybe part of what this session might be about is it could trigger that curve to go back up on the trajectory that it was on earlier. Part of the reason that it went up was people were talking about more ambitious standards and wanting to have tests that did that. A lot of states introduced open-ended responses, and then No Child Left Behind has pulled back on that — not in all states, but in quite a few states, so that has damaged the trend line. Having cooperative arrangements such as Achieve or the Chiefs that would help push it in the other direction, even if all the states or most of the states don't buy into that, at least we have a model there that could be beneficial.

"If we don't hang together, we'll hang separately."

John Merrow:

I am curious. How many of you think that five years from now we will have moved significantly forward in the direction of higher common standards and perhaps common tests? Could I just see hands if you think we're moving in the right direction?

John Bishop:

But it isn't NCLB-type things. It's the expansion of AP and the International Baccalaureate (IB). It's a whole variety of things going on.

John Merrow:

In 1754, the Colonists were fighting among themselves over all sorts of things, boundaries and so on and so forth. Ben Franklin tried to unify them with something called the Albany conference. He called them together and they refused to unite. Franklin said, famously, "If we don't hang together, we'll hang separately." What actually unified the col-

onists was the Stamp Act and the Tea Act. It may well be that NCLB may be the equivalent, and this is the shot heard around the world in education.

In any case, I appreciate the opportunity to spend time with all you folks.

Richard Nathan:

John, thank you. You mentioned the Albany Plan of Union, I want to say two things about it. One, I wasn't there. Second, right across from our building we look out at the Albany Plan of Union Street, and it was because of the French and Indian Wars that Franklin brought them there more than anything else. But anyway, thank you much.

Many have mentioned Allison's paper. What should we do to it? Where might we go with it? I want to tie together what Checker and Mike Cohen talked about in the morning session with the kinds of things that were introduced here and focus just a little bit on what we can take from this. Not tell you everything I've learned today — that's more than I can do. But what we might take from this and where we might go with our particular interests in institutional machinery, institutional change, federalism, and what the states are doing, with the Diploma Project and Achieve out front, and other things that could be part of it or go along with it, as Tom suggests, in the way of institutional considerations and possible options.

We've got maybe 18 months. We are not going to decide this, as Lynn reminds me, this year and right away. So with your help and involvement, I'll sort of probe to get some input from you on how we can, in our work, draw on this very good discussion.

Implications of Today's Discussion

Richard Nathan:

Let me gather us together. A man Checker Finn and I once worked for said, "Bring us together." That's not exactly the phrase, was it Checker? Checker and I go back a long ways. Lots of wonderful, warm memories. We knew more people than I ever thought I would know who went to prison. I always say, if you want my story, I worked in the first term of the Nixon administration.

Chester E. Finn, Jr.:

The first two years is the best way to say that.

Richard Nathan:

Under a man I admire and learned a great deal from, George Schultz, in the Budget. And we worked closely with Checker when he worked with Pat Moynihan.

First and foremost, thank you for coming. This was a really good day and a wonderful group of people who know this subject and obviously care about it deeply and contributed to what we sought to do. I feel it worked very well to do it here today.

So let me remind you, why did we invite you to come here? We invited you to come here because we bring particularly an interest in federalism, states, and institutional structures to many areas of domestic policy — particularly in times like these when the national government is so preoccupied. So I want to try to link some of the things that people talked about and call on a couple people, particularly with this institutional change, institutional invention, states and federalism orientation that led us to ask Spencer and the Joyce Foundation to support this work that we are doing.

Let me start by saying that I'm proud of Allison and proud of the fact that a lot of people reacted to — and I never know who to believe, but I think I looked in their eyes and I thought a lot of you meant it — that her paper did help us and did bring something new to this national policy making NCLB debate and subject matter. I want to remind you what the paper does, and then I want to talk about what we're going to do next — which is not a lot, but some things that we do feel we can do and we will do. I hope you'll help us and participate with us. Then I want to call on a few people, particularly on institutional and federalism dimensions of our good discussion. Then I'm going to ask Allison to comment a little bit on what she thinks about how this day has gone.

Allison's paper says we should consider a third way: "Reliance on one or more intergovernmental entities involving states and educational experts in a collaborative process to identify…" and it goes on to speak about standards and testing and all the things that we talked about. Then in the paper, on page one, she talks about the role of collaborative entities. The biggest and most important one that we've talked about a lot today, Achieve and

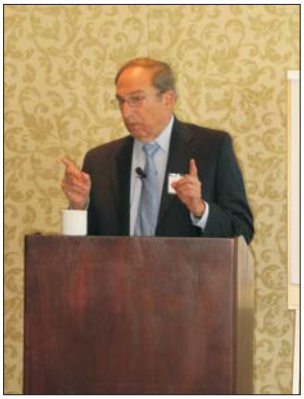
the American Diploma Project is, of course, the work that Michael Cohen has done. He's the first person that I'm going to call on to say what he thinks about the implications of our discussion for the state role in educational reform and accountability going forward. I also want to call on Checker because he had comments about the way in which we could, or should, think about reaching out for material resources — public, private, state, federal, blends, not blends, foundation only — to support things that we are considering.

Then I'm going to call on Tom, because he mentioned something that the feds could do — a role that may or may not be compatible when you think about institutions and institutional change and institutional invention. He mentioned auditing and supporting research. Is there or is there not a way in which NIST — a huge organization which is where thousands of people work collaboratively with industry to come to agreement on standards — could be a model? It's been out there a long time, and nobody ever heard of it and it doesn't rile up all the politicians. So maybe there is federal role, and I'll ask Tom to speak about that.

Then I'm going to call on Eva because she gets the highest mark for having mentioned part three of your paper, which I wanted to particularly remind you is right here, "possible functions" of the interstate and other entities that you know about, and maybe new and maybe federal entities. We're the institution people.

I'm going to call on those four people and then Allison, and then open it up for you to tell me what we federalism wonks should do in the coming months about these issues that you have done so well today to educate me about. I had a lot to learn. I came here with more to gain than anybody else in this room, I think.

Following Up After the Symposium



Richard Nathan

Now, what are we going to do with all of this? We have a transcript, and one thing we're going to do is publish and disseminate it along with Allison's paper. We're very proud of what she did to write it and to draw on you and to pull people together for this symposium.

Secondly, Allison and I will write a new paper drawing on today's rich conversation. I was for years chairman of the board of the Manpower Demonstration Research Corporation (MDRC). The board chair before me was Eli Ginsberg. At the end of the meeting, he would say, "This is what we did," and go, "one, two, three, four, five, six." And I thought, "Eli is so smart. Why am I not as smart as Eli? Why couldn't I do that?" And I can't. The reason is, I think, Eli didn't really do it. I think when I listened harder, it just sounded good. I don't think he know everything that everybody discussed in a day where we covered so many things as we did today. So, as I said, I'm not going to try to tell you everything I've learned today; instead, Allison and I will devote some time to writing a new paper.

The third thing, so that we take advantage of today's conversation, is we're going to have Lynn Olson write about what she thought, as someone who follows the political and policy process.

A Bigger Puzzle

Richard Nathan:

I'm first going to call on Michael Cohen. Michael, we talked a lot about what you're doing and what you want to do and what the role of a collaborative of the states, the Diploma Project and Achieve, and maybe other groups is. Thinking about this conversation, what might you want to say to us that would be very valuable input to us on how you take all this in.

Michael Cohen:

We are being pushed by the states to do this. We did not go to them and say, "Don't you think it would be great to have common standards and tests?"

Well, this is all very simple actually. I think, like you were at the end of the first session, I feel confused at a much higher plane than I was when I came here. There are some things that I think are pretty clear to me, but my big takeaway from this conversation is all the other things that are more complicated now than they were before. Let me illustrate.

First, the things that are clear: The work that Achieve has been doing with a critical mass of states, I think, will enable us to create an existence proof that it's possible for states to have common standards and for some large number of them to have a common assessment. One of the things that makes that possible, by the way, is that we are quite deliberately being pushed by the states to do this. We did not go to them and say, "Don't you think it would be great to have common standards and tests?" They've been pushing us along, and that dynamic, I think, is very important for the success we've had so far. Hopefully, also the success that we'll have in the future. So that's the easy part, around the standards and some critical mass of states on one or more tests.

What I've been struck by in this conversation is as this is moving quite rapidly, there are a whole set of other issues that there is nobody really positioned to attend to. So I was taken, for example, by the idea that came up several times that there really ought to be a number of different collaborative efforts underway. Even if we got a critical mass of states to common standards, there would be a number of efforts to create different tests aligned with and reflective of those standards. But there's not an obvious set of candidate organizations to do that and no one in charge of organizing it.

There was lots of discussion about what else is needed to bring about results, in addition to standards, assessments, and accountability: curriculum, instructional tools, formative assessments, a whole variety of things like that. The private sector will create some versions of those, but there's no multistate effort to do that, and again, no obvious candidate to do that.

On the issue of the federal role, I think, as Checker does, that our ability to make progress on this agenda is enhanced without the federal government being heavily involved in it. The less involvement, the better. Nonetheless, current NCLB requirements provide some of the environment in which this work occurs. Some argued that getting all kids proficient by 2014 is, at some point pretty soon, going to put a real roadblock in front of any effort to create higher standards. So, number one, what can be done to create some more room for states? Number two, there's a big difference between a critical mass of states and a 50-state effort, or even a 45-state effort. What kind of incentives are going to be necessary to provide more states to move in this direction, if you think it's the right direction in which to move. And at what point, if any, will the federal government be necessary to create those incentives?

That last issue is a relatively active conversation between Achieve, the Alliance for Excellent Education, the Hunt Institute, the Chief State School Officers and the National Governors' Association, though I wouldn't tell you that we've got it all figured out yet by any stretch of the imagination. To the extent that it's focused on the federal government, Congress may do us a favor and give us time to figure it out. But these are fairly complicated issues. I think it was Checker who said at the end of our session, we're in agreement that this kind of effort is best accomplished with little or no federal role. But there are a lot of things that need to be done that one would naturally otherwise turn to the federal government to address, and there's not a really good substitute for that.

What kind of incentives are going to be necessary to provide more states to move in this direction, and at what point, if any, will the federal government be necessary to create those incentives?

So I think what I'm taking from this is, as we continue to move ahead with the states with which we're working and others that may join with us, that there's a bigger puzzle that's going to require more attention over time than we've given it so far. And I think it's going to take a lot of us working on that to figure it out.

Richard Nathan:

That's a very good list. Thank you very much. This symposium was timely, and we've got time. They're not going to reauthorize NCLB this year. After a presidential election, it takes a while before you get people lined up and and working on issues that don't involve the war in Iraq (editorial comment). So there's maybe 18 months for people to take a hard look at other arrangements, other designs, other federalism and institutional entities, collaboratives, and collaborations.

A Design Competition

Richard Nathan:

Checker, you commented in the conversation with Mike Cohen this morning about where the money might come from: Foundations? Other sources? In my business (and I've been in and around this business a long time), you have to think about who would support,

or how we would support, activities such as those that Michael Cohen just listed or others that you might want to mention.

Chester E. Finn, Jr.:

Generally, the day has dwelled on three megaproblems. One is the problem of variable and discrepant standards from place to place. The second is the problem of weak and unsatisfactory tests from place to place. The third is the problem of missing structures for solving the first two problems.

There is some time to think about these things. There's an enormous need here. I think we need some creativity and some imagination. We could just use ADP efforts, but it's already clear from the conversation today that there are other organizations and entities, even in the room, that are dealing with pieces of these problems in different places.

I think there's a need for some kind of design competition to come up with multiple strategies, solutions, and structures. Allison lays out four categories of possible structures at the end of her nice paper and elaborates a bit on three of them, and more work needs to be done in that area.

I think we need some creativity and some imagination. There's a need for some kind of design competition to come up with multiple strategies, solutions, and structures.

I really do believe that there's foundation money to be had for what, in my mind's eye, is maybe a two-step design competition. The first step is really cheap. It might be nothing more than paying a lot of people $1,000 each to write brainstorming papers about an outline of a solution or a structure. You get 20 papers in for $20,000 and you evaluate them. And if at least four of those are any good, you might have the basis of a persuasive argument to make to the Gates Foundation, the Spencer Foundation, the Joyce Foundation, or the Broad Foundation — you can go on with this list — to actually invest in the much more ambitious design competition that it would take to flesh out some of these ideas into something that might actually be operationalized.

I think it was Susan (Traiman) who referred to the substantial payments that go to aerospace and defense industry contractors just to do the design for a new helicopter or something. We may have to get to that point, where we are able to rustle up the X million dollars (but again, as I said, it's not billions) from the private sector to underwrite a design competition for some structural arrangements that might help us solve the first two problems having to do with standards and tests.

The results from the first round, the twenty papers for $1,000 each, might bomb, so to speak. You might discover there's nothing there worth pursuing. But if there are half a dozen good ideas, then you can go back to the private funders and say, "Hey, we have here the makings of a design competition worth having. It would be good for the country over this two-year period to actually roll that out in an imaginative way."

There are lots of partial answers floating around this room and around the country, but nothing's been thoroughly developed, had its pluses and minuses weighed, and been batted around by people who want to argue for and against different options here.

From Left: William Harris, Thomas Toch

If the Broad and Gates Foundations can devote $60 million to the "Ed in '08" strong American schools campaign that they're currently embarked upon, which, as far as I can tell, is little more than an elaborate public relations effort to inject education into a presidential election that doesn't really want to deal with it, then there's a lot of private sector money on the education issue right now. I don't think that's the biggest barrier for design.

Implementation's another whole story. That requires political buy-in and things like that, and it does involve NCLB. But I don't think there are major obstacles to design, other than imagination and very modest amounts of investment.

New Standards Project

There are lots of partial answers floating around this room and around the country, but nothing's been thoroughly developed, had its pluses and minuses weighed, and been batted around by people who want to argue for and against different options here.

Scott Marion:

I was thinking about this design competition. Would people get deducted points if they plagiarized the New Standards project? There's a pretty good history of that. It worked for a while and looked like it was going to take off, and then it didn't. There's got to be something we could learn from that effort.

It involved good standards, technically strong assessments, great design teams, good content experts, an articulated system. So why aren't we talking about New Standards? I know there're lots of things wrong with it. I'm just listening to what you're suggesting and thinking, is there something we can learn from that failure?

Chester E. Finn, Jr.:

It might have been ahead of its time. It might not have had the right kind of political sensitivity. It might have just had "educator sensitivity" to what people thought they'd like to see kids learn, rather than the creation of structures by which states and federal governments make decisions about accountability in a standards-based reform era. I don't know.

Yes, an interesting paper could be written on what's to be learned from previous efforts. Parts of Allison's paper, of course, involve previous efforts at federal involvement with standards setting. Somebody also mentioned earlier today that the New American Schools initiative produced an effort to reinvent the school that didn't lead to much. It, too, might usefully be autopsied.

Scott Marion:

Mike, did you think about New Standards in ADP?

Michael Cohen:

Yes. In fact, we've got a significant part of the New Standards team that work for Achieve.

Scott Marion:

That's what I thought. We're going to get them through this time!

Gordon Ambach:

What are they telling you, Mike, about what not to do?

Michael Cohen:

Well, the folks that we have did the technical work around assessment development. The issue now gets joined with the growing interest in performance assessments, for which New Standards might well have been ahead of its time. In the conversations we're having with them, essentially creating content area-specific design specifications and rubrics, in order to be able to have performance tests for which you can get a consistency and comparability across settings, turns out to be particularly important.

New Standards was too much for the state to actually buy into, because it meant pushing aside too much of what they had.

I think we learned a fair amount (although I don't know that I know how to articulate it off the top of my head) about how to create a common effort by states. In fact, I said before that we are being pushed by the states, rather than having gone to the states and saying, "We have an idea for you." That's a big difference. I remember well the difference between going from state to state selling them on an idea, versus having chiefs call me to find out if I'm really as serious about this as they are. When I get those calls, I know that the dynamic is right and we'll be able to keep moving. That's a big change from before.

Gordon Ambach:

Just watch out how big a package you create. That's the key. New Standards was an all or nothing deal.

Richard Nathan:

Politics, a little politics in this corner.

Gordon Ambach:

New Standards was too much for the state to actually buy into, because it meant pushing aside too much of what they had. The critical issue there is: How much of a package can you get, and what increments can you support?

Information for Consumers

Richard Nathan:

I was going to call on you, Tom. You mentioned an audit function, and other people mentioned other possible NIST-like (or anyway under-the-radar-like federal functions) that would fit well with, and perhaps augment, things that are in this mix that Michael Cohen suggests is already out there and evolved.

Thomas Toch:

Backing up a little bit to the first principles of our conversation today, if we had a different group of 40 or 45 folks in this room, we would probably have a different set of assumptions that we're working off of. It's important to note, therefore, that this group generally agrees that 50 disparate sets of standards and tests — very disparate — is probably not ideal in terms of driving the system where we want to go.

There's not wide agreement over that. There are some who take a very strong federalist stance who reject that just in principle, the notion that sort of started our conversation today. But we do seem to agree that centralization of some sort, whether top-down, if that's the right phrase — that's loaded so we probably shouldn't use it — but having the federal government somehow be involved in that process, or a bottom-up movement of the sort that Mike is leading with Achieve, will take us to where we want to go to create that as a more centralized, more coherent, and more efficient system.

I think that, at bottom, there is the notion that standards have to come first. We tended today to talk a little bit about tests and testing, but we need to keep in mind that if you don't have strong standards, rigorous standards, then the tests, by definition, are not going to be as good as we hope them to be.

We talked about two different things today and it's important to keep that in mind. One is about what's the most efficient way to create a new, more centralized system. The other conversation was around how do we ensure the quality of tests, the sort of audit function. Those conversations overlap a lot, but they're not exactly on the same subject, and we tended to conflate them a little bit today. We all were doing a little bit, and it led to a little bit of confusion.

So to me (and this was the focus of the paper that I wrote last year), the question is: Assuming that we have good standards, assuming that we have testing systems, then how do we ensure on an ongoing basis that they are valid and reliable; that they test what we want them to test; that they separate the distinguished, high-achieving students from low-achieving students, or smart kids from less-smart kids, and the like?

There are different ways to go there. I propose this notion of the Consumer Product Safety Commission-like entity, and I use that model in part because of the emphasis on the consumer nature of this. There's very little information out there for parents, and the busi-

ness community, and the general public at large — the sort of primary consumers of this information — that they can understand in a readily digestible way.

One thing about NAEP that is very important we might consider in this model are the content maps that they have, which equate individual score numbers, like 320, with a particular skill, an ability to negotiate New York City with a map, that represents the skill level there. That kind of consumerist information (I think Checker was implying that when he talked about a *Consumer Reports* model) is very valuable, in addition to ensuring that we have the technical strength within each test that we need.

So you can create a new entity. You could build, as Allison suggests, on existing entities, and Dick suggested the same thing, perhaps through the National Research Council Board on Testing and Assessment. Or you could go Checker's route, which would be to have an outside independent entity that is truly consumerist in nature.

But in every instance, my thought is that we produce information that makes us confident as consumers, and as policymakers and taxpayers, in the rigor, and the validity, and the meaningfulness of these standards and assessments as a way of focusing attention on the higher level standards that we all aspire to.

There's very little information out there for parents, and the business community, and the general public at large — the primary consumers of this information — that they can understand in a readily digestible way.

Eva Baker:

About the validity issue, or an external group to do test evaluation or consumer reporting, I have my own experience at Center for Research on Evaluation, Standards & Student Testing (CRESST). We did, for six years, have a test evaluation project where we collected measures and did what we could. We had terrible times, as Bob knows, trying to get additional data. We would use public data, and then we would collect more from the schools, and now that's almost impossible. We gave it up because of lawsuits from test publishers who didn't like judgments that their test only got three dots instead of five dots, and wanted to contest that. So I think that the way all of that plays out at the detail level would make a difference in terms of how useful it would be.

Diversifying Qualifications for High School Students

Richard Nathan:

Eva, you mentioned section three of Allison's paper, about functions that should be part of how we think further about intergovernmental approaches to education accountability reform.

Eva Baker:

Well, the conversation I think you're referring to takes the strategy that one doesn't create something new in the face of all this other stuff that's in place, but picks and chooses one's spots. My suggestion was to focus on diversifying the qualifications and certifica-

tions for high school students, in terms of either test performance or certifications that could be awarded by business groups, government groups, or community groups. There would need to be an overarching manager that would have to certify those groups. This is very much in the spirit of the merit badge discussion that we had. It's a way of broadening the choice for kids, making high school more instrumental for them, recognizing their adultness, and giving them something usable that may actually have real stakes for them in the future.

We need to step outside of the educator perspective, because I think that's far too narrow. If we brought in business people, military people, folks who have a different societal purchase, then we might be able to create some opportunities for secondary school students who may or may not decide to go on to postsecondary opportunities.

So I had been thinking that what we wanted to do was to forge a coalition. I assumed it would not happen through the government, although the government could fund it through developmental mechanisms. But it could be a consortium of foundations, of groups sort of like ours, of groups like those represented by the people in this room — however it made sense. It would include the guilds or the professionals that would have to be there to certify and to get the buy-in from their people.

The other point I wanted to make was about the awarding bodies and the notion of multiple players at the state or national level. I think that there's a way of doing that, a way of handling those competitions and a way of providing incentives for getting the very best options, but allowing those options to be a little different so that people can adapt them to their own needs.

The reason I thought this was a reasonable place to start was I didn't want to deal with my own reservations about what's going on in K-8 as far as testing, assessment — the fact that some of that is hardened beyond cracking, I think, at this point. So this was partly a strategy to get there.

Richard Nathan:

Thank you, Eva. John Bishop talked about that, too. But I said I would ask Allison now to make some comments about the meeting and what she's thinking about as she listens to and talks to you about the paper.

Allison Armour-Garb:

Thank you. Well, I've been trying very hard to listen rather than talk, and I've been comfortable in that role. But I'm interested in hearing what the states feel that they want and need, in terms of help for strengthening their accountability systems. I know that one thing that Mike feels has been important in Achieve's success is that it's driven by what the states want.

Allison Armour-Garb

Of course, No Child Left Behind puts certain incentives on states in terms of how they're going to structure their accountability systems. Nevertheless, Achieve's been able to get states interested in setting high standards for end of high school, even if the incentives in No Child Left Behind to some degree work against those.

So maybe there are things that states' chief school officers and accountability people would like to try to put their heads together and work on, even if it's not necessarily completely compatible with the incentives under No Child Left Behind. What direction do they want to work in? How do they want to try to drive the agenda for strengthening accountability systems, moving forward? Are they interested in getting help implementing the kinds of standards that Eva and Bob Linn came up with, standards for accountability systems, more broadly than just the *Testing Standards*? Is there something else that they want to work on?

More research is needed on everything from the impact of policies to technical details. We need a structure for setting the agenda and prioritizing the research to be done, and then reviewing and synthesizing the research at the other end, so that we actually advance knowledge and practice.

I don't know that states are going to ask to be audited. But maybe there are things that some of the state representatives here might want to talk about in terms of capacity building, or ways of leveraging resources, the expertise of the psychometricians and other accountability experts that they can get input from — on a group basis, not just with their individual technical advisory committees, but working together.

A District Opt-In Model

Susan Traiman:

This is sort of half-baked in my mind, but if you look at what's happening among home schoolers, many of them, because of the access to the Internet, are using common curricula and common tests. Could there be different models where, for example, districts could opt in? A district could say, "We want to hold our schools to the common national standard, as opposed to the Maryland standard." Or it could be like an IB diploma for an individual that is recognized across state borders. Maybe there could be networks of schools that want to be part of the network of schools that give common national test, or something like that. There are all sorts of problems with state responsibility, but I don't know if that's something we could explore further.

An Appeal for More Research

Lauress Wise:

I would just like to add an appeal for more research. More research is needed on everything from the impact of policies, to technical details like how content standards could be

better aligned with detailed theories of cognitive development, and so on. In particular, we need to think about a structure for setting the agenda and prioritizing the research to be done in the first place, and then reviewing and synthesizing the research at the other end, so it's not just a hundred unrelated research studies that reinvent each other, but that we actually advance knowledge and practice with cumulative research results.

Richard Nathan:

The Department of Education sponsors a lot of research, in my opinion (editorial comment follows), much too heavily influenced by very narrow classical experimentation, underlined. What I think of when you spoke about a research agenda is a research agenda that takes cognizance of what states would like to work on, and together might work on, in a way such that federal support would be appropriate and useful.

Council of Chief State School Officers

Phoebe Winter:

There is a model that is a start for state collaboratives, and it's State Collaboratives on Assessment and Standards out of the Council of Chief State School Officers (CCSSO). They have states come together voluntarily — and actually pay to be there — to discuss and do research on issues related to assessment and accountability. There's actually one called "Accountability Systems and Reporting." It's a good model.

States are willing and want to get together to work on these issues. It's just having a vehicle to do it.

The thing that they don't have, I think, is the money and the support to do things like Laurie was talking about: pulling together the research, putting it out in a way that will help all different states. But it's a model to start with and build on, maybe do something different with.

I think the states are willing. I don't want to talk for all of you, but I will. States are willing and want to get together to work on these issues. It's just having a vehicle to do it. There's got to be an incentive that doesn't then take them away from doing something else.

Gordon Ambach:

The Council of Chief State School Officers has a number of different consortia, which were much more extensively used in the 1980s, and the 1990s in particular, and then the function dropped off. But you really want to use CCSSO as a key source. You really need to go and talk with Gene Wilhoit, who is the current executive director. Colleagues here are working in various networks that are provided by CCSSO. There's machinery there, and a lot of experience about how you fathom what's desired and then how you go about working on it and get solutions which match the interests that the states have.

Test Validity Depends on Use; Problems Under NCLB

John Bishop:

Under the *Test Standards*, the design of a test depends upon the uses to which it is to be put and the decisions that it is intended to inform. We haven't discussed those decisions that it's intended to inform carefully in this. We focused on the quality of the test, regardless of the particular uses we intend for it.

Actually, my view is if we're going to have school accountability, it should be based on value-added measurement, not the kind of levels measurement and looking-over-time AYP approach that NCLB implemented. Increasingly, states are building the capacity to be able to switch to value-added. I think whether the tests that you're going to have are to be used in a value-added framework or not will really influence the nature of the tests and the way you want to run the system.

Another very critical feature is this accountability system with multiple hurdles, i.e., all groups within the school need to get above a certain standard and then the school is okay, but if any one of the groups fails the standard, the school goes into this category of being a "school in need of improvement." I think the *Test Standards* say that conjunctive approach is wrong and it should be compensatory, and you should be using a more reliable measure of the performance of the school for making these kinds of judgments. We didn't discuss that too much.

So I just wanted to bring up two issues that I think are very important: (1) The tests are not an independent thing, and (2) *Test Standards* regarding testing incorporate the issue of whether your intended use is a value-added use or a level approach versus a multiple hurdle approach versus a compensatory approach. Also, they always recommend multiple measures, i.e., some additional measure besides the particular test that you've designed, such as teacher grades or some other indicator.

So there are a lot of issues that we did not discuss here that I think are very important and would influence what should be done. I just wanted to put on your radar screen that the world is even more complex than we described it earlier.

Richard Nathan:

Allison's nodding. It's on her radar screen.

Technology Push

Eva Baker:

A lot of the talk today was about consulting people, see what their needs were. The states, represented with the work in Achieve and the Chiefs and other places, may provide

us with an appropriate requirements-driven agenda. That's what they want now, what their pressures are now.

In contrast to that "requirements-pull," what I hope we might do is either help them anticipate what they might want later. As our good friend, Tom Glennan, first said, have something called "technology push," where you'd say, "Here are some new ideas. Is this something you're interested in? How would that work in your environment?"

Richard Nathan:

Lynn began with that in the first session on five years from now. I think that's a good reminder.

I want to say one more thing. I'm going to hang around a little while, Allison too, if there are things you want to whisper in our ear, or send to us, or influence us about. We are very pleased that you came, and we thank you very much.

"Here are some new ideas. Is this something you're interested in? How would that work in your environment?"


STAIRS

Intergovernmental Approaches for Strengthening K-12 Accountability Systems

A Framework Paper circulated in preparation for a Symposium convened by
The Nelson A. Rockefeller Institute of Government
with the support of the Spencer Foundation and the Joyce Foundation

October 15, 2007

by
Allison Armour-Garb
Director, Education Studies

Rockefeller Institute of Government
411 State Street
Albany, NY 12203

Table of Contents

1 Introduction

Five years into the implementation of No Child Left Behind (NCLB), there is much uncertainty about the impact of the federal law and next steps the nation should take to advance educational performance and accountability.

Defenders of the law say it has spurred better performance and focused attention on the educational needs of underperforming groups of students. Critics say its emphasis on testing has distorted the learning process and overburdened state education departments. Some argue for greater federal control of standards and assessment. Others say states should be given greater flexibility so that they may address problems unique to their citizens and pursue innovative approaches.

This paper will consider a third way — the development of one or more intergovernmental entities involving states and educational experts in a collaborative process to identify national standards and oversee testing, without direct control from Washington. The paper has been prepared for a Symposium to be held October 29, 2007, entitled "Intergovernmental Approaches for Strengthening K-12 Accountability Systems," convening leading experts to discuss the need and prospects for this possible alternative.

Some such entities may include intergovernmental collaboratives that already exist — for example, the American Diploma Project Network and the State Collaborative on Assessment and Student Standards. Others might include ideas such as Thomas Toch's proposal for a national testing oversight agency modeled on the Consumer Product Safety Commission, Daniel Koretz's suggestion that consortia be established to conduct research-based evaluations of testing programs, and variations on the idea of an agency to audit testing.

The paper raises questions to focus our discussion at the Symposium:

- Is there a role for a collaborative federal-state institution to set national standards?[1]
- Is there a need for an existing or newly created intergovernmental entity to oversee states' educational accountability systems — testing in particular? What sort of structure might work? What functions could it perform — e.g., technical assistance, review, accreditation, or research? What enforcement role might it have, and how would the right degree of political accountability be achieved?
- What features would such collaborative governance structures need to be successful? How could they balance federal, state, and local interests — and what role would the private sector play (i.e., employers, testing organizations, foundations)?

To help inform our consideration of these questions, the paper looks at some of the structural problems that weaken educational accountability systems. It then covers some of the functions a collaborative entity could perform to address those problems. Finally, it considers existing models and some of the alternatives that have been proposed.

---

1 This paper does not analyze the pros or cons of national standards; its focus is on finding better ways to implement policy, not on arguing for one set of policies or another.

2 Structural Problems in Educational Accountability

Stakeholders are grappling with many structural problems[2] in educational accountability — problems that might be addressed by intergovernmental mechanisms:

- Sound accountability systems are difficult to design and operate, and many state agencies lack adequate access to, or budgets to pay for, the expertise they need to implement such systems.
- Policymakers, educators, and testing companies face incentives to lower standards, cut corners, and "game the system."
- The public lacks a clear idea of the performance and effectiveness of the various components of the education system, because standards and measures of performance vary widely from state to state and between states and the National Assessment of Educational Progress (NAEP).
- The variation in curriculum standards across states poses difficulties for teachers, students, and school leaders who relocate across state borders; and poses challenges for test developers, schools of education, and textbook companies when they decide what material to cover.
- There is not enough good research on large-scale assessment, its effects on student learning, and its other consequences.

These are not problems that can be worked out by tweaking isolated policies, such as when a standard or cut score is set too high or too low. Rather, these are structural features of the educational accountability sector that probably require changes in institutions and incentives. This section describes these structural problems and explains their relevance to the need for and design of intergovernmental arrangements for standards-setting and oversight of accountability systems.

2.1 Shortage of Expertise

Educational testing is a highly technical field. Its practices are governed by the *Standards for Educational and Psychological Testing* (often simply called the *Test Standards*),[3] which are jointly developed and periodically revised by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education[4] (See the text box on page 115, "What Do the *Test Standards* Cover?") Creating and implementing a high-quality testing system that conforms to the *Test Standards* is difficult, labor-intensive, and expensive. This is particularly the case because the highly

---

2 Thomas Toch, *Margins of Error: The Education Testing Industry in the No Child Left Behind Era* (Washington, D.C.: Education Sector, 2006), 6.

3 American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing* (Washington, D.C.: Authors, 1999).

4 In 2002, in response to NCLB's emphasis on accountability, the National Center for Research on Evaluation, Standards, and Student Testing and the Consortium for Policy Research in Education published their *Standards for Educational Accountability Systems*, but these standards have not yet achieved the canonical status of the *Test Standards*. Additional sets of standards, checklists, and other guidelines aimed at test developers and test users have proliferated in the era of standards-based reform — some of them emanating from the federal government in a recent push to help states implement NCLB's testing provisions.

What Do the *Test Standards* Cover?

The *Standards for Educational and Psychological Testing* are jointly developed and periodically revised by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. They address the following professional and technical issues:

Part I: Test Construction, Evaluation, and Documentation

1. Validity
2. Reliability and Errors of Measurement
3. Test Development and Revision
4. Scales, Norms, and Score Comparability
5. Test Administration, Scoring, and Reporting
6. Supporting Documentation for Tests

Part II: Fairness in Testing

7. Fairness in Testing and Test Use
8. The Rights and Responsibilities of Test Takers
9. Testing Individuals of Diverse Linguistic Backgrounds
10. Testing Individuals with Disabilities

Part III: Testing Applications

11. The Responsibilities of Test Users
12. Psychological Testing and Assessment
13. Educational Testing and Assessment
14. Testing in Employment and Credentialing
15. Testing in Program Evaluation and Public Policy

Source: American Psychological Association, "The Standards for Educational and Psychological Testing," www.apa.org/science/standards.html (accessed October 8, 2007).

technical *Test Standards* can be properly followed only by experts trained in measurement theory and statistics,[5] known as psychometricians, only a few of whom enter the field each year.[6]

The surge in demand for testing under NCLB has led to a critical shortage of testing experts. Because states and districts across the country are all mandated to comply with the law's testing mandates, the need for testing experts is widespread and decentralized. Unfortunately, federal policymakers imposed these new testing requirements without a full appreciation of the technical challenges they would pose, both for the

5 Robert L. Linn, "Following the *Standards*: Is it Time for Another Revision?" *Educational Measurement: Issues and Practice* 25, no. 3 (Fall 2006), 54-55; Koretz, telephone conversation with author, June 20, 2007.

6 Toch, *Margins of Error*, 9.

professionals who would write the tests and the laypeople charged with implementing them.[7] Of course, many of the government officials who write and implement education laws are education experts, broadly speaking, but relatively few trained psychometricians work in the education bureaucracy at any level. Not only does this help to explain the weaknesses in NCLB's testing provisions, it is a problem that should be overcome before enactment of any future government policies concerning testing.

The shortage of expertise has troubling consequences. Testing experts who work for private companies typically receive much higher salaries than those who work in the public sector. Those who work as consultants or in academia may have the freedom to choose their own projects, collaborators, and work schedules.[8] Education agencies — which typically offer neither competitive salaries nor flexible working conditions — have therefore been hit hardest by the shortage and have experienced high turnover in these positions.

As a result of this difference in hiring power, it is understandable that many state education departments lack staff able to design a technically sound testing program. Most states purchase their tests from commercial test publishers, and many states rely on outside consultants for advice. Yet some states lack sufficient technical know-how to supervise outside testing contractors effectively.[9] When states work with independent testing advisors for just a few days per year, officials may not know the right questions to ask, or — not grasping the significance of consultants' recommendations — may fail to implement them.[10]

This results in what economists call "asymmetric information." The officials who write education laws and buy tests, due to a shortage of experts on their payrolls, may not be aware that the tests and testing services supplied by contractors are running afoul of the *Test Standards*. Even if the testing companies are producing technically defensible tests, the government's uses of the test results or the conditions of administration may violate the *Test Standards*. The affected students and the public are even less likely to know that assessments may be violating the *Test Standards.*

2.2 Perverse Incentives

Before the implementation of NCLB, politicians, education officials, teachers, and test publishers stood to gain in varying degrees from increases in student test scores, whether or not those scores represented real gains for children. But NCLB has increased the stakes — it added both carrots and sticks — and thereby increased the need for effective oversight of educational accountability systems. On the incentive side, the law ties states' Title I funds to compliance and has made billions of dollars available to test publishers. It also in-

7 William G. Harris, "The Challenges of Meeting the *Standards*: A Perspective from the Test Publishing Community," *Educational Measurement: Issues and Practice* 25, no. 3 (Fall 2006): 43.

8 Rather than work full-time for a single testing company or government agency, many testing experts opt to offer their services as consultants to multiple agencies. For example, they serve on states' "technical advisory groups" or "technical advisory committees"; work with state collaboratives such as the American Diploma Project Network and the New England Common Assessment Program; or provide technical assistance and review through the U.S. Department of Education's NCLB peer review process, its Assessment and Accountability Comprehensive Center, and its LEP (Limited English Proficiency) Partnership.

The multistate perspective that such arrangements give many testing experts may be seen as a fortuitous consequence of the national shortage of psychometricians (see footnote). If every state could afford to hire and retain its own staff of experts, those experts would likely develop a more bureaucratic and parochial mindset.

9 Toch, *Margins of Error*, 9, 20.

10 Lauress Wise, interview with author, July 20, 2007.

creases the visibility of high-scoring districts and those that post proficiency gains for all students and disadvantaged subpopulations. The sticks are the negative consequences (programmatic and political) that flow from failing to show "adequate yearly progress" in the percentage of children who score proficient on the required tests each year.

NCLB has posed new financial, practical, and technical challenges for both testing companies and states, and with those challenges come strong incentives. The testing market is dominated by highly competitive for-profit corporations whose primary motive may be to increase profits by satisfying client demands. NCLB has greatly increased the amount of criterion-referenced testing that states must conduct, and it is difficult and expensive for test publishers to create and score so many new tests within NCLB's tight timelines. Moreover, each state sets its own curriculum standards, and its tests are supposed to be tailored to those specific curriculum standards. The fact that testing companies must custom-design aligned tests for each state has multiplied the challenges that states and testing companies face in implementing NCLB.[11] The norm-referenced tests that were more common in the past did not require customizing or annual "refreshing" and were therefore much cheaper to produce.

These new demands are "squeezing testing company profit margins"[12] and have created pressure to cut corners — to design instruments that test lower-order thinking and are generic (i.e., weakly aligned) so they can be marketed to multiple states with minimal customizing.[13] These companies may seek, to a greater or lesser extent, to follow the *Test Standards*, but the threat of formal professional censure (which is rare) is far less salient than the desire to secure the largest possible client contracts and get the testing done and scored on time.

Similarly, education officials at the federal, state, district, and school levels — whether they are politicians, psychometricians, or anyone else connected with a particular jurisdiction's accountability systems — have strong incentives to demonstrate increases in the percent of students scoring proficient on their watch, while avoiding increases in taxes or budgets. As a result, they may be unlikely to push testing companies for improvements that could result in longer timelines, higher price-tags, or more challenging test questions.[14] In addition to such inaction, educators can take a more active approach — "gaming" the system by teaching to the test or focusing on students whose scores are just below the cut score, allowing students extra time, or even cheating outright.[15]

One of the easiest ways for states to avoid or postpone sanctions under NCLB is to set a low bar for proficiency. This approach is possible because states, districts, and schools face sanctions for failing to make "adequate yearly progress" (AYP) towards their state's proficiency standards under NCLB, while states are free to set proficiency standards high, low, or in between. States that set high standards risk having the most

11 Harris, "Challenges of Meeting the Standards," 43; Toch, *Margins of Error*, 9.

12 Toch, *Margins of Error*, 12.

13 Ibid, 14-16.

14 NCLB may also inhibit states from experimenting with innovations in accountability design. Note that although federal pressures and sharing of best practices across state borders has led to the dissemination of growth models, the original value-added model emerged in Tennessee during a period of independent state action.

15 Evidence of test misuse is abundant. See Wayne J. Camara and Suzanne Lane, "A Historical Perspective and Current Views on the *Standards for Educational and Psychological Testing*," *Educational Measurement: Issues and Practice* 25, no. 3 (Fall 2006), 38.

schools labeled "failing" under NCLB. Indeed, many states set very low proficiency standards (compared to the NAEP) to begin with, and others lowered their standards after they recognized the incentives to do so. This type of behavior is sometimes called a "race to the bottom."[16] Thus, NCLB may have in some cases exerted downward pressures on state proficiency standards.

What of legal incentives? How does the law deal with those who do not adhere to the *Test Standards* in their testing of public schoolchildren? The answer is that, for the most part, it does not. The *Test Standards* are not laws or regulations; they are professional guidelines. Some aspects of the *Test Standards* are paraphrased in the NCLB legislation and regulations, and states' adherence to these provisions is overseen through a peer review process (see the text box on page 119, "Current Federal Oversight of State Accountability Systems Under NCLB").

In general, however, the testing profession has a tradition of self-regulation, so psychometricians who do not follow the *Test Standards* may be subject to professional censure, but that sanction is imposed rarely and is a weak counter to the financial motives that test publishers face. Moreover, when tests are misused by the laypeople charged with implementing them, professional censure is not an available check. There is, in sum, no adequate mechanism for enforcing the *Test Standards*, much less the newer *Standards for Educational Accountability Systems* developed by the National Center for Research on Evaluation, Standards, and Student Testing and the Consortium for Policy Research in Education.[17]

As a result, educational accountability policies rarely are subjected to serious debate or scrutiny. Officials, educators, and test publishers face incentives to cut corners to save money, and to "game the system" to show short-run increases in the percent of students scoring proficient on their watch — i.e., during their term of office, during the term of their contract, or during their employee review period. The prime incentive facing the federal government, the states, and the testing companies is to do what it takes to make NCLB compliance work, at least for the time being, in order to keep the Title I money flowing.

2.3 Lack of Transparency

The variation in state standards and tests makes it difficult to interpret the test score data that states generate. Because each state defines its own curriculum standards, sets its own bar for proficiency, and uses its own tests, it is impossible to directly compare proficiency rates in one state with those in another. And because some states have changed their proficiency standards from one year to the next, it is also difficult to compare proficiency rates over time. Moreover, some states have systems for grading or ranking schools and districts that are at odds with the federal system, which adds to the public confusion.

The only reliable yardstick for comparing children's educational achievement across states and over time is the NAEP. Numerous studies and reports have sought to compare state standards with the NAEP, and

16 According to one new report, there has not actually been a "race to the bottom," with the majority of states dramatically lowering standards under pressure from NCLB, but rather a "walk to the middle," as some states with high standards dropped their expectations toward the middle of the pack. John Cronin et al., *The Proficiency Illusion* (Washington, D.C.: Thomas B. Fordham Institute and Northwest Evaluation Assoc., October 2007) 30.

17 Eva L. Baker et al., "Standards for Educational Accountability Systems," *CRESST Policy Brief* 5 (Winter 2002).

Current Federal Oversight of State Accountability Systems Under NCLB

The federal government conducts three types of oversight of state accountability systems under NCLB: peer review, studies by the U.S. Government Accountability Office, and reviews by the U.S. Department of Education's Office of the Inspector General. Although the federal government certainly has the authority to gain access to states' information and to enforce its rulings, its oversight is nevertheless limited to those aspects of state accountability systems that are subject to NCLB (or other federal laws such as the U.S. Constitution); other aspects of state accountability systems are generally free from independent oversight. Some analysts have noted that federal oversight has tended to focus on inputs and processes (e.g., how much are states spending on NCLB assessments?) rather than on the consequences of state systems (e.g., what is the impact on student learning? Are there unintended negative consequences, such as narrowing in the range of cognitive skills or subjects areas taught, or a "dumbing down" of tests?).

NCLB Peer Review — Under NCLB, states are required to prepare a report documenting aspects of their accountability systems and to submit it for peer review evaluation. The Education Department selects the peer review teams, which typically consist of a psychometrician, an educator who is an expert in working with special populations, and another testing professional with experience in large-scale assessment. Employees of testing companies are excluded from the teams. This system provides for expert, in-depth review of some aspects of state accountability systems. The teams' decisions are not "all or nothing," and states have an opportunity to request technical assistance. However, the reviews do not include an "opportunity to learn" criterion as would be required under the *Test Standards*, and review of consequential validity is weak. While the U.S. Department of Education offers technical assistance (largely through consultants), states generally have not used it.

GAO Studies — The U.S. Government Accountability Office (GAO) conducts frequent studies on selected aspects of states' NCLB accountability systems. These studies are usually one-time analyses based on a small sample of states and geared toward recommending improvements to the Department of Education's own systems and technical assistance efforts, rather than direct enforcement of states' adherence to federal law or the *Test Standards*. As the Department of Education does with peer review, the GAO has used teams of experts to inform at least some of its studies and has looked at whether states follow "generally accepted test development procedures." The GAO conducted an intergovernmental audit of NCLB assessment practices in coordination with selected state and local auditing agencies.

OIG Reviews — A third federal oversight mechanism is afforded by the U.S. Department of Education's Office of the Inspector General (OIG). OIG's 2006 report on NCLB found that Part A of the legislation (which covers standards and testing) contained 566 compliance requirements. The report questioned whether implementation of these requirements was being adequately monitored by the Office of Elementary and Secondary Education, and stated that it appeared that states were only monitoring a minimal number of the requirements. These conclusions were based on OIG's review of federal

Continued on the following page

and state "monitoring guides," documents stating what aspects of the law's implementation are being monitored. OIG does not appear to have studied how the monitoring guides were used in practice or to have reviewed the substance of state accountability systems.

Sources: Brian Gong, Testimony to the Commission on *No Child Left Behind,* www.aspeninstitute.org/atf/cf/%7BDEB6F227-659B-4EC8-8F84-8DF23CA704F5%7D/Brian%20Gong%20Testimony.pdf; Camara and Lane, "Current Views on the *Standards*," 39; Toch, *Margins of Error*, 20; Phoebe Winter, telephone conversation with author, September 19, 2007; Susan L. Davis and Chad W. Buckendahl, "Evaluating NCLB's Peer Review Process: A Comparison of State Compliance Decisions" (paper presented at 2007 annual meeting of National Council on Measurement in Education (NCME)); Susan L. Davis, e-mail message to author, Sept. 8, 2007; Harriet Ganson, telephone conversation with author, September 17, 2007; United States Government Accountability Office, *No Child Left Behind Act: Assistance from Education Could Help States Better Measure Progress of Students with Limited English Proficiency*, GAO-07-646T (Statement of Cornelia M. Ashby, Director, Education, Workforce, and Income Security Issues, Washington, D.C., March 23, 2007), www.gao.gov/new.items/d07646t.pdf; U.S. General Accounting Office, U.S. Department of Education, Office of Inspector General, State Auditor's Office, State of Texas, Department of Auditor General, Commonwealth of Pennsylvania, and City of Philadelphia, Office of the Controller, *A Joint Audit Report on the Status of State Student Assessment Systems and the Quality of Title I School Accountability Data (SAO Report No. 02-064)* (Austin: Texas State Auditor's Office, August 2002), www.ed.gov/about/offices/list/oig/auditreports/s14c0001.pdf; United States Department of Education Office of the Inspector General, *Compliance Requirements within Title I, Part A of the No Child Left Behind Act,* Final Management Information Report: State and Local No. 06-01 (March 29, 2006),

even to map state test scores onto the NAEP scale,[18] but such studies are not yet routine, nor have they reached the majority of ordinary citizens.[19]

2.4 Inefficiencies Resulting From Diversity of Standards

Variation in state standards and tests also poses challenges for teachers, schools of education, test developers, and textbook companies when they are deciding what material to cover, and for students when they move from one state to another.

As noted above, because state standards vary widely, publishers of tests and textbooks theoretically ought to custom-develop materials for each state. They save on development costs and increase profits, however, by developing generic materials that cover the common elements of multiple states' standards. In the case of tests, publishers may do some minimal customizing. It is questionable, however, whether such weakly aligned tests are effective at measuring how well students have learned what they were taught.[20]

Similarly, because of the variation in standards across states, teachers are not necessarily prepared to teach to the standards in the state in which they teach. They may attend a teacher education program in one state and move to another after graduation.

---

18 E.g., National Center for Education Statistics, *Mapping 2005 State Proficiency Standards Onto the NAEP Scales (NCES 2007-482)* (Washington, D.C.: U.S. Department of Education, June 2007).

19 With the possible exception of those who watched the June 11, 2007, episode of "The Colbert Report" on Comedy Central, which explained the methodology using Mississippi's 4th grade reading test as an example.

20 Toch, *Margins of Error*, 14-16.

2.5 More Research Needed on Validity of State Assessments

To some degree, the lack of scrutiny of state accountability policies is due to a scarcity of hard data on the direct and indirect impacts of large-scale assessment programs. Experts believe that the gap between the goals of educational accountability systems and the actual strength of the research base supporting policy reforms has led to problems.[21]

Obviously, when problematic test results are used to make high-stakes decisions about individual students, such as whether to promote them or grant them a high school diploma, the effects are harmful. But poorly designed or misused tests are also harmful when they lead education officials to base decisions about curriculum, resource allocation, personnel, or sanctions on flawed information. With the increased use of assessments for accountability purposes, there is a correspondingly greater need for routine evaluations of the validity of state and district testing policies. While a few states have years of usable data on their state assessments and have been open to researchers, independent validity studies are far from routine in most states. Ongoing evaluations and longitudinal studies would provide better information on which policymakers could base their decision making.[22]

What Is "Test Validity"?

The validity of a test is generally defined as the extent to which it measures the knowledge or skills that it is intended to measure. Tests themselves are not valid or invalid. Instead, experts validate the use of a test score in a particular context, for a particular purpose. Furthermore, validity is a matter of degree — not all or nothing. There are numerous aspects of validity. For example, "curricular validity" refers to the alignment between test questions and the objectives of the particular curriculum they are intended to assess. "Predictive validity" refers to the usefulness of test scores to predict future performance. "Consequential validity" refers to the social consequences of using a particular test for a particular purpose.

Source: College Board, "ACES: Validity Handbook," www.collegeboard.com/highered/apr/aces/vhandbook/testvalid.html and www.collegeboard.com/highered/apr/aces/vhandbook/evidence.html (accessed October 8, 2007).

3 Some Possible Functions of an Intergovernmental Entity

The foregoing problems have become apparent as states, districts, and schools have implemented the testing provisions of the federal No Child Left Behind Act. Even if NCLB were to be scrapped tomorrow, however, these problems would not disappear. Any system of educational accountability that is retained — whether at the federal, state, district, or school level — would have to contend with these problems. Since the 1980s, the public has come to accept the basic principle behind educational accountability systems in return

21 Eva L. Baker and Robert L. Linn, "Validity Issues for Accountability Systems," in *Redesigning Accountability Systems for Education*, eds. Susan H. Fuhrman and Richard F. Elmore (New York: Teachers College, 2004), 47.

22 Camara and Lane, "Views on the *Standards*," 39; Toch, *Margins of Error*, 14 (quoting Scott Marion); Daniel Koretz, telephone conversation with author, June 20, 2007.

for funding, : educational entities should demonstrate that they are achieving the desired outcomes (i.e., that students are learning appropriately). Whatever happens to NCLB, American voters are unlikely to revert to judging schools on their inputs, as was common prior to the standards-based reform movement.

This section will review some of the functions that intergovernmental mechanisms could perform to help states strengthen their accountability systems and address the problems described above.

3.1 Develop National Standards and Tests, Conduct Testing, and Report Results

National standards and testing have been proposed to address numerous issues in education, including several of the problems described above. The establishment of common standards and measures of proficiency would end the "race to the bottom" — the situation in which states lower their standards in order to increase the percentage of students labeled "proficient." National testing could put an end to the dumbing down of tests and could lead to transparency by ensuring that results are comparable across states and over time. Finally, the establishment of national standards would create a "national market" for curricular materials and tests; would enable teacher education programs to prepare candidates to teach the content and concepts that their students would be expected to attain;[23] and would facilitate the mobility of students, teachers, and school leaders across state borders.

The 2006 Thomas B. Fordham Institute report, *To Dream the Impossible Dream*, explores how national standards could be established without necessarily increasing the federal role in education: if other federal regulations are concomitantly eliminated; if states are left free to voluntarily adopt the standards; or if states actually develop the common standards and tests themselves.[24]

3.2 Audit or Accredit State and District Accountability Systems

An auditing or accreditation function could strengthen incentives for states and districts to scrutinize their accountability systems and work to ensure those systems are designed and implemented in accordance with the *Test Standards* or other jointly agreed standards.

Accreditation is a mechanism that is widely used in areas such as higher education, child care, and health care. It is a voluntary process that involves self-study and validation by professionals outside the program to verify that standards are met. In the child care field, for example, research has demonstrated that accreditation positively impacts program quality.[25] An independent accreditation process for state and district educational accountability systems could:

- increase public confidence in accountability systems;
- involve agency officials in evaluation and planning;

---

23 Chester E. Finn, Jr., et al., *To Dream the Impossible Dream: Four Approaches to National Standards and Tests for America's Schools* (Washington, D.C.: Thomas B. Fordham Foundation, August 2006), 10-11.

24 Finn, et al., *Impossible Dream*, 12-13, passim.

25 National Association for the Education of Young Children, "What research tells us about NAEYC accreditation," www.naeyc.org/ece/1996/16.asp (accessed September 30, 2007).

- spread awareness of good practices, create self-improvement goals for weaker programs, and stimulate a general raising of standards for educational accountability systems;
- shift the focus away from compliance and gaming, and insulate education agencies against harmful internal and external pressures; and
- help to identify programs for the investment of public and private funds, and provide one of several considerations used as a basis for determining eligibility for funding.

Audits could also lead to improved practice and transparency, for example by ensuring the accuracy and comparability of state and district test results from year to year[26] (see the text box on page 124, "Current Federal Oversight of State Accountability Systems Under NCLB," for a description of federal audits of NCLB assessment practices). In comparison with accreditation, government audits are generally mandatory rather than voluntary;[27] rely primarily on external review rather than self-study; and tend to shine a spotlight on weaknesses rather than highlighting success stories. Thus, they might reinforce rather than counter the current compliance mindset.

3.3 Conduct Validity Studies

There is some consensus in the testing community that longitudinal studies evaluating the effects of accountability programs and ongoing evaluation of the validity of test-based inferences are essential to ensure that accountability systems are serving the best interests of children and the public.[28] Indeed, the *Test Standards* provide that "[i]t is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences."[29]

Researchers might begin by determining what issues are of greatest concern to stakeholders, and then choose those research questions that are most critical for achieving consensus regarding appropriate use of the test in question. The research could aim to provide information for decision-makers and point future researchers toward the most perplexing issues.[30]

---

26 Michael Petrilli has proposed national audits to ensure transparency in test scores (e-mail message to author, June 25, 2007; phone conversation with author, July 23, 2007). Diane Ravitch has proposed that New York State establish such a function, though she views this alternative as a second-best solution in the absence of national testing (Diane Ravitch, "Reflections on the Math Scores in New York City and State," New York City Public School Parents blog, June 17, 2007, http://nycpublicschoolparents.blogspot.com/2007/06/diance-ravitch-reflections-on-math.html; e-mail message to author, July 13, 2007).

27 Voluntary audits conducted by private organizations such as the Educational Testing Service and the Buros Institute are quite limited in their potential to enforce the *Test Standards* (Linn, "Following the *Standards*," 55).

28 Baker and Linn, "Validity Issues," 68-69; Daniel Koretz, e-mail message to author, June 15, 2007.

29 Camara and Lane, "Views on the *Standards*," 39.

30 Lee J. Cronbach, "Construct Validation After Thirty Years," in *Intelligence: Measurement, Theory, and Public Policy*, ed. Robert L. Linn (Urbana: University of Illinois Press, 1989), 164-65.

Model Criteria for Accreditation or Audits

Various models exist for accreditation or audit criteria:

- the *Test Standards*, which include broadly accepted psychometric criteria for test instruments and their use;
- the CRESST/CPRE standards for accountability systems, which may be most on point;
- the Middle States Commission on Higher Education standards for Assessment of Institutional Effectiveness and Assessment of Student Learning — which specify, for example, that assessment processes must be useful; cost-effective; reasonably accurate and truthful; carefully planned; and organized, systematic, and sustained;
- the Joint Commission on Accreditation of Healthcare Organizations' standards for accreditation of medical testing laboratories, which are divided into categories such as: Leadership, Management of the Laboratory Environment, Management of Human Resources, Management of Information, and Quality Control.

Sources: American Education Research Association, American Psychological Association, and National Council on Measurement in Education, Standards for Educational and Psychological Testing, (Washington, D.C.: Authors, 1999); Eva L. Baker, Robert L. Linn, Joan L. Herman, and Daniel Koretz, "Standards for Educational Accountability Systems"; Middle States Commission on Higher Education, *Assessing Student Learning and Institutional Effectiveness: Understanding Middle States Expectations* (Philadelphia, 2005); Joint Commission, "Comprehensive Accreditation Manual for Laboratory and Point-of-Care Testing," www.jointcommission.org/AccreditationPrograms/LaboratoryServices/Standards/FAQs/default.htm.

4 Institutional Alternatives

There exists a rich array of possible mechanisms to perform the functions and address the problems outlined above. This section will raise some of the pros and cons of various options.

We might ask the following questions about any proposed entity:

- How would it balance federal, state, and local interests? To what degree would it preserve state flexibility to experiment with innovative accountability schemes? If it is a voluntary arrangement, how might governments be encouraged to participate? How would the right degree of political accountability be achieved?
- What enforcement role might it have? Would it have the legitimacy and authority to gain access to information and to enforce adoption of its recommendations? Would it rely on the *Test Standards*, or would new standards or regulations be needed?
- What role would the private sector play (i.e., employers, testing organizations, foundations)?
- How would it be funded and staffed?

Possible structures include a nongovernmental research agency, a federal agency, a state-led collaborative, and a state-federal-private hybrid. The first three of these are discussed next.

4.1 Research Entity with Government Support

When policymakers need a steady supply of research on which to base their decisions, they often create research entities to conduct studies on their behalf.[31] Such entities may be nonprofit or university-based, and they typically receive government funding to support their policy work.

Plans are currently underway for an international consortium to conduct comparative studies of testing programs in Europe, where government agencies have tended to be more open to such research and have granted access to the necessary data.[32] If more state governments were similarly willing to provide data, such consortia could be established in the U.S. as well.

To encourage state participation, an intergovernmental research entity would, at a minimum, need to ensure data confidentiality. It would likely also offer states a role in identifying issues of concern and nominating ideas for studies. Researchers would have to be free of unfavorable biases.

Beyond these basics, such an entity could offer several advantages to participating states. For example, it could generate outside funding from private or federal sources, so that participating states would not have to finance the research themselves. It could facilitate multi-state studies, which might reduce the burdens for each participating state while simultaneously generating useful comparative information and possibly even providing political cover. Finally, it could also serve as a central clearinghouse for data, freeing states from having to respond to many requests for information.[33]

4.2 Possible Federal Models

While past federal attempts to set curriculum standards have failed (see Appendix A: Lessons from Previous Federal Efforts), the federal government has been quite successful in working with private industry to standardize technical specifications and processes. The National Institute of Standards and Technology (NIST) is a federal agency whose mission is to "promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life." It conducts research, co-funds public-private research and development (R&D) partnerships, provides technical assistance to small manufacturers, and gives an annual award recognizing excellent performance and quality in the private sector. It also coordinates federal, state, and local technical standards with those developed voluntarily in the private sector, with the goal of eliminating duplicative or unnecessarily complex regulations.

Increased awareness among education stakeholders of the ways in which NIST has promoted innovation and opened up markets in the technical sector might generate interest in using NIST as a model.[34] A federal educational standards agency modeled after NIST could:

---

31 Some examples include the National Research Council, a congressionally chartered nonprofit that provides policy advice to the federal government; MDRC, a nonprofit created by the Ford Foundation and a group of federal agencies to study social policy; and the Education Finance Research Consortium, a university-based research and policy analysis venture established by the New York State Education Department.

32 Daniel Koretz, e-mail message to author, June 15, 2007, and telephone conversation with author, June 20, 2007.

33 Many of these ideas were suggested by Thomas Gais.

34 This idea was first suggested to me by Richard Nathan.

- promulgate voluntarily developed educational standards;
- develop improvements in educational measurement science;
- provide technical assistance to states and districts; and
- recognize promising innovations in educational accountability.

Another interesting model has been proposed by Thomas Toch of Education Sector. He has called for the establishment of an independent national oversight agency to audit state testing programs and test publishers, "in the spirit of the Consumer Product Safety Commission" (CPSC).[35] The mission of this agency, which Toch dubbed the "National Testing Quality Commission" (NTQC), would be to protect students and society from harms caused by poorly designed tests and improper test use, in the same way that CPSC protects the public from risks of injury or death due to consumer products.

Its functions could include all of those outlined in the previous section and more. Additional powers, analogous to the powers of the CPSC, might include:

- promulgating standards for tests and test use;
- regulating the tests and testing services produced and sold by private companies;
- banning tests or testing practices;
- collecting and disseminating information about the quality of tests and testing programs; and
- maintaining a hotline, online forms, and other mechanisms for the public to report suspected problems with tests or test use.

To target resources and enforcement efforts most effectively, a federal (or intergovernmental) agency charged with regulating state accountability systems could use an approach known as "differential oversight." In such a system, states that have a good record of compliance (in terms of their commitment to federal goals or their actual educational achievement levels/progress) could be given more flexibility in using federal funding and in dealing with problem districts, while those states that are recalcitrant are subject to more intrusive oversight and sanctions.[36] To be legitimate, however, a system of differential oversight must be insulated from political pressures; otherwise, it is vulnerable to manipulation. So, for example, the enforcement body might rely on input from an independent board (such as the National Research Council's Board on Testing and Assessment, or the NCLB peer review panels) regarding which states deserve flexibility and which require greater oversight.[37]

---

35 Toch, *Margins of Error*, 20.

36 Thomas Gais suggested including this idea, which was also recently proposed by Frederick M. Hess and Chester E. Finn, Jr., in "Crash Course: NCLB is Driven by Education Politics," *Education Next* (Fall 2007), 45. See also Thomas Gais and James Fossett, "Federalism and the Executive Branch," in eds. Joel D. Aberbach and Mark A. Peterson, *The Executive Branch* (Oxford: Oxford University Press, 2005), 510.

37 See Richard P. Nathan, "Reinventing Government: What Does It Mean?" *Public Administration Review* 55, No. 2 (March/April 1995), 213-15, for a discussion of political insulation.

4.3 State-Led Collaborative

Rather than go it alone, many states have found that they can save money and make progress faster and more effectively by collaborating on standards and testing. There are several factors motivating these collaborations. States that work together can learn from one another's experiences, policies, and programs. By collaborating, they may be able to achieve compliance with NCLB more quickly (or at least demonstrate good faith in trying to comply). Finally, collaborating states can share the cost of retaining experts and achieve economies of scale.[38]

Examples include nonprofit-based collaboratives such as Achieve's American Diploma Project and the Council of Chief State School Officers' State Collaboratives on Assessment and Student Standards (SCASS) program (notably, the Technical Issues in Large-Scale Assessment and Assessing Limited English Proficient Students projects), and interstate compacts such as the New England Compact's Common Assessment Program (NECAP).

Such collaboratives typically:

- begin with a goal shared by a few states, then involve more states incrementally;
- gain buy-in by building on the substantive strengths and political support of states' existing systems, rather than seeking to replace them;
- provide resources and capacity building to help states implement and integrate reforms into their existing systems, to ensure that all the components work together to form a coherent instructional system;
- set realistic timelines for implementation that do not depend on election cycles;
- rely on independent consultants to do much of the technical work; and
- leverage financial support from the private sector — business, foundations, and universities.

Note that although the private sector has provided expertise, frameworks, and financing, these collaboratives are fundamentally intergovernmental and rely on the commitments made by participating governments for much of their power and legitimacy. By contrast, strictly private organizations — such as the National Board on Educational Testing and Public Policy (NBETPP) and FairTest — have in the past been unable to conduct meaningful oversight of educational testing, in part because they lack access to testing information and authority to enforce the practices they endorse.

Achieve's president, Michael Cohen, believes that there are two reasons why some large states such as California, Florida, and New York have not participated in interstate education collaboratives such as the ADP Network and the New England Compact. First, they have more internal capacity and do not need to rely on the access to resources that such collaboratives afford. Second, because of their size, these states face some very different challenges than smaller states, and they may believe there is little they could learn from those states through collaboration.[39] Thus, unless federal carrots and sticks are used, many states may never join such collaboratives.

---

38 To the extent that the national shortage of psychometricians has motivated interstate collaborations, that has been a fortuitous side effect. See footnote 8.

39 Michael Cohen, telephone conversation with author, June 21, 2007.

The roles of federal and local governments in these state-led collaboratives have been limited. Federal grants have provided some of the funding for NECAP, and federal representatives have worked closely with SCASS. Recently, the Secretary of Education has publicly praised the American Diploma Project, and the administration is reportedly considering providing funding to encourage more states to collaborate in developing assessments that comply with NCLB.

Appendix A: Lessons From Previous Federal Efforts

Following is an abridged retelling of the origins of the standards-based reform movement 25 years ago and the federal government's subsequent attempts to establish a national system of educational accountability. The focus is to draw some lessons about the federal politics of educational standards-setting and testing oversight.

The modern movement to develop national standards and tests began with the publication in 1983 of **A N*ation at Risk*,** a report by the National Commission on Excellence in Education, whose members had been appointed by President Reagan's Secretary of Education. This "galvanizing event" focused public attention on the need to improve educational standards.[40] The response was immediate, but it was not coordinated at the national level. Individual states created task forces and commissions to appraise their curricula and graduation standards.[41] The report did not put an end to the crisis, however, nor did it resolve ongoing debates over education policy.

The federal government didn't step in again until 1989, when President George H.W. Bush (Bush I) invited the nation's governors to an **education summit** in Charlottesville, VA. At the summit, the president and the governors agreed on the idea of establishing national goals for education. By the early to mid 1990s, that idea had evolved into efforts — either inspired or financed by the federal government — to develop voluntary national standards. The ensuing efforts to establish national curriculum standards and tests fell victim to political strife, one after the other:

The **America 2000** Act, which Bush I introduced in April 1991, called for creating voluntary American Achievement Tests pegged to curriculum standards in five subjects (English, math, science, history, and geography) in grades 4, 8, and 12. But the legislation never passed, amid wariness over the federal government's role in prescribing curricula and debates over whether curriculum standards should be accompanied by "opportunity-to-learn standards" (which liberals argued were necessary to define the programs, staffing, and other resources needed to help students achieve at high levels).

While Congress was still debating the parts of America 2000 that had been submitted as legislation, the Bush I administration used discretionary funds to award grants to organizations of scholars and teachers[42] — chosen noncompetitively — to develop **voluntary national standards** in seven academic subjects (not including math, because the National Council of Teachers of Mathematics had independently promulgated its own standards in 1989). Unfortunately, there was no provision for independent quality review of the standards these organizations developed; the plan was for the standards to be judged by the Education Department staff monitoring the grants. Instead, however, the standards were judged largely in the political arena.[43] In late 1994, a tremendous controversy erupted over the U.S. History Standards, which critics deemed revi-

---

40 Diane Ravitch, *Left Back: A Century of Failed School Reforms* (New York: Simon & Schuster, 2000), 411.

41 Ibid., 413.

42 Kevin R. Kosar, *Failing Grades: The Federal Politics of Education Standards* (Boulder: Lynne Rienner Publishers, 2005), 32.

43 Michael Cohen, "Do We Need National Standards?" in "No Child Left Behind: A Five Year Review," ed. Dick Clark, special issue, *Aspen Institute Congressional Program* 22, no. 1 (February 20-25, 2007), 46.

sionist and negative towards Western culture. The Secretary of Education issued a statement disowning the standards, and the U.S. Senate voted 99 to 1 to condemn them.[44]

The Clinton Administration's **Goals 2000** legislation, enacted in early 1994, provided funds to states to develop academic standards, and almost every state began developing them. The legislation also authorized the creation of a National Education Standards and Improvement Council to review the quality of the national standards that were already being developed with the Bush administration's grants. However, fears of a "national school board" that would dictate local curriculum from Washington — fueled by the ideologically charged debate over the U.S. History Standards — rendered the council so controversial that the provision was repealed by the new Republican Congress before the members of the council could be appointed. Thus, the voluntary national standards funded by the Bush administration were published at the same time that most states were developing their own curriculum standards, with the result that the national standards were very influential on state standards even though their quality had never been independently reviewed.[45]

Moreover, although Clinton's 1994 Title I reauthorization, the **Improving America's Schools Act**, ostensibly tied future Title I funds to states' adoption of challenging standards and aligned assessments, those provisions were never enforced due to the administration's fears of creating in the new Republican Congress a backlash against the federal role in education.[46]

At the beginning of his second term, Clinton proposed **Voluntary National Tests** in 4th grade reading and 8th grade math. The Council of Chief State School Officers was chosen to co-draft specifications for the tests, which were to be based on the highly regarded National Assessment of Educational Progress (NAEP) frameworks. Unlike the NAEP, however, which is overseen by the National Assessment Governing Board — an independent, bipartisan organization — Clinton's proposed tests were to be overseen by a board appointed by the secretary of education; they were thus seen to be inherently political. Predictably, conservatives in Congress objected to the expansion of the federal role in education; in addition, liberals revived their demands for opportunity-to-learn standards. The administration tried to bypass Congress by using discretionary funds to begin creating the exams (as Bush had done a few years earlier), but Congress struck back by voting to require explicit congressional authorization of the development of the tests, and by late 1998 the project was dead.[47]

In leading Congress to pass the **No Child Left Behind Act**, President George W. Bush insisted on reaching a compromise that was acceptable to both parties. Thus, rather than establishing national standards and testing, the law allows states to define their own curriculum and proficiency standards and to establish their own tests. To provide a common yardstick for gauging the stringency of state proficiency standards, NCLB requires all states to participate in the National Assessment of Educational Progress (participation had been voluntary). Considering the unprecedented amount of testing mandated by the law — annual testing of all children, including those with disabilities and limited English proficiency in reading and math in grades 3 through 8 — the timeline for implementation was very fast: NCLB was signed in 2002, and the first year of testing was 2003.

---

44 Ravitch, *Left Back*, 435.

45 Ibid., 433.

46 Kosar, *Failing Grades*, 166-67.

47 Ibid., 171-74.

The Bush I and Clinton efforts to establish national standards and tests provoked two kinds of debates: "[D]ebates over which level of government ought to be responsible for education standards and curriculum, and [debates] over the ideological persuasion of the individuals or organizations involved in the process. These fights prevented almost all of the federal initiatives from enactment and/or implementation."[48] Understanding what motivated them may help present-day reformers anticipate and prepare for — or prevent — similar battles. As Michael Cohen notes, the heated political debates over the 1990s initiatives "often eroded public support for standards-based reform led by any level of government, started ideological and often partisan battles that spilled over to the state and local levels, and distracted attention that could otherwise have helped standards based reform move forward."[49]

Federal efforts have also been criticized for particular design flaws:

- Bush I funded the development of national standards documents whose quality was never independently reviewed, leading to the promulgation and adoption by states of potentially flawed standards by states. Contemporary state-led efforts have sought to develop standards that hold up to independent scrutiny.
- The Clinton administration failed to provide meaningful oversight of the implementation of its Title I accountability provisions. NCLB has been enforced more stringently, with some degree of professional oversight afforded by its peer review process.
- Federal efforts have failed to promote a coherent approach to standards-based reform. In other words, they did not help states translate standards into a system that could improve student learning. To create such a system requires aligning curricula with standards, designing assessments to gauge progress towards the standards, developing data systems to track results, and training educators to implement it all. This shortcoming is highlighted by the debates over opportunity-to-learn standards, in which activists continue to decry the imposition of isolated accountability demands without assurances that states can align their educational efforts to meet those demands, much less that districts and schools have the capacity to ensure all children receive the education envisioned by policymakers. This is another area in which contemporary state-led efforts have sought to improve upon the federal government.
- NCLB has exacerbated some of the structural problems in educational accountability and institutionalized some perverse incentives.

Nevertheless, it has been argued that these federal efforts succeeded in focusing diverse actors (including policymakers, analysts, researchers, and advocates) on the problem, the proposed solution (standards-based reform), and the question of how best to structure educational accountability systems. In addition, the federal government spurred states to establish standards and assessments through rhetoric, financial aid, and nominal financial incentives.[50] Finally, these efforts increased awareness of the resources and technical know-how needed to develop sound accountability systems.

---

48 Cohen, "Do We Need," 46.

49 Ibid.," 46.

50 John F. Jennings, *Why National Standards and Tests? Politics and the Quest for Better Schools* (Thousand Oaks, CA: Sage, 1998), 183.

Appendix B: Previous Efforts to Establish Oversight of Testing

Over the past century, testing experts have put forward a range of proposals for more effective oversight:[51]

- In 1925, a well-known author of standardized tests called for an independent organization to evaluate tests and provide information about them to buyers. He compared the issue with food labeling and held up Consumers' Research, Inc. (forerunner of the Consumers Union) as a model.
- In the 1930s, Oscar K. Buros tried and failed to start a test consumers' research organization. According to Buros, test publishers were marketing tests that failed to meet professional standards, using exaggerated, false, or unsubstantiated claims. While he believed that test users were becoming somewhat more savvy, tests that were nicely packaged and made impossible promises nevertheless found many gullible buyers.
- When the American Psychological Association formed its original committee on test standards in 1950, it also considered establishing a Bureau of Test Standards to enforce them. Tests that satisfied the Test Standards would have been granted a Seal of Approval. The proposal went nowhere.
- In the 1970s, a national commission formed under the former Department of Health, Education, and Welfare recommended the establishment of a federal agency to oversee testing. The commission report stated that well-designed standardized tests could have value when used appropriately, but that tests were too often of poor quality and misused, and that the efforts of professional organizations and reputable test publishers did not prevent "widespread abuse." Comparing the effects of poor tests and testing on children's opportunities to the effects of tainted food and drugs on health, the commission recommended the establishment of a National Bureau of Standards for Psychological Tests and Testing. The proposed Bureau would have set standards for tests; test uses and test users; acted on complaints; conducted research; and disseminated its findings. The recommendation was never followed up.
- In 1991, FairTest proposed requiring an Educational Impact Statement (modeled on Environmental Impact Statements) before adoption of any new national testing system.
- In 2006, Education Sector proposed the establishment of a National Testing Quality Commission modeled on the Consumer Product Safety Commission.

The exact rationales and proposed solutions vary, but the common theme is that independent oversight — with teeth — is necessary because poorly designed and misused tests continue to harm children, despite the existence of professional guidelines. Without exception, however, these proposals have failed.

Watchdog organizations and university-based consortia have also sought to carve out a role overseeing testing. For example, the National Board on Educational Testing and Public Policy (NBETPP), based at Boston College, monitors tests for appropriate use and technical quality. The National Center for Fair & Open Testing, better known as FairTest, advocates for an end to what it calls "the misuses and flaws of standardized testing." However, both of these organizations have come to be seen as opposing testing rather than

51 Most of these examples are drawn from George F. Madaus, "A Brief History of Attempts to Monitor Testing." *NBETPP Statements* 2, no. 2 (February 2001).

ensuring it is done right. The National Center for Research on Evaluation, Standards, and Student Testing (CRESST), based at the University of California at Los Angeles, seems to be striving for a broader consensus. In cooperation with the Consortium for Policy Research in Education, CRESST has published a set of *Standards for Educational Accountability Systems*.[52] Yet none of these nongovernmental organizations has the authority to demand access to testing information or to enforce the practices they endorse.

Why has effective oversight of testing never been established? As noted above, politicians generally have little appreciation of the technical issues around educational testing. Rather, many politicians have a stake in persuading the public that the accountability systems they have put in place are working effectively to demonstrate growth in educational achievement during their tenure in office, as discussed in the following subsection. By contrast, while members of the testing profession understand the problems with testing, they have not been particularly active in lobbying for policy changes — at least until recently. Even those who believe current policies need to be changed may be reluctant to come forward — because if they document the problems with testing, they risk a collective loss of credibility and, if they lobby for external oversight, they risk a loss of professional autonomy.

52 Baker, et.al., "Standards for Educational Accountability Systems."

Appendix C: Achieve

The organization that is leading much of the intergovernmental collaboration in educational accountability is not a creature of government, but a nonprofit organization. Conceived during discussions that were held at the 1996 National Governors Association education summit, Achieve, Inc. was created later that year by a bipartisan coalition of governors and business leaders. Its mission is to help states strengthen accountability by raising academic standards and improving assessments.

Achieve President Michael Cohen, who acknowledges being a veteran of more failed attempts to establish national standards than anyone else in the country,[53] argues that states are the *de facto* leaders in establishing standards-based reforms and that they are currently in a better position than the federal government to carry on that leadership. "Over more than a decade," he notes, "every state has implemented systems of standards and assessments." Moreover, they have invested "significant … political capital and financial resources [in] developing and building support for these initiatives."[54] In the wake of the controversies around the federal government's national standards initiatives, "state officials were particularly cautious about advancing standards based reform, and were often forced to spend more time and attention protecting their efforts from political attack than ensuring their educational quality."[55] Because state accountability systems have managed nevertheless to earn political support and buy-in, he argues, future efforts to improve standards-based accountability should seek to build on and improve, rather than replace, state systems — thereby avoiding (as much as possible) the political fallout that hampered federal reforms in the 1990s.

Achieve works to strengthen and build on states' existing accountability systems, thereby capitalizing on the existing political support and buy-in that those systems enjoy. It has avoided political battles because it works only with those states who want to participate, it works with them to build on their existing systems, it provides financial and technical resources, and its standards are anchored with the legitimacy of real-world demands. For all these reasons, Achieve has been successful in working towards "increasingly consistent state standards" without creating a national political battle.

Anchoring Standards in Real-World Demands

Achieve gained national attention with its American Diploma Project (ADP), which it sponsored in partnership with the Education Trust and the Thomas B. Fordham Foundation. Rather than relying on scholars and teachers to define content standards by group consensus and compromise, as the voluntary national standards of the 1990s did, ADP started from the premise that the standards high school graduates need to attain are defined by the real-world demands they face when they apply to college, seek employment, or enlist in the military.[56] ADP conducted a two-year research study with postsecondary institutions and employers in five states to define the knowledge and skills in math and English that high school graduates must have in order to succeed. Based on this study, they produced a set of external benchmark expectations for the end of high school. Unlike standards developed by academic consensus, external benchmarks have their legitimacy

53 Michael Cohen, telephone conversation with author, March 28, 2007.

54 Cohen, "Do We Need," 47.

55 Ibid., 46.

56 Ibid., 46.

in real-world demands, and that legitimacy is reinforced as Achieve continually updates the standards with new information from employers and higher education institutions across the nation.

Achieve is currently backmapping its end-of-high-school expectations through 4th grade for English, and all the way down to kindergarten for math. They may possibly develop benchmarks for science, but they do not plan to expand their work to any other subjects.[57] In recognition that students must be prepared to compete in a global economy, a new initiative undertaken by Achieve in partnership with the National Governors Association and the Council of Chief State School Officers will examine how states can benchmark their education systems against high-performing countries around the globe.

A Coherent Approach to Supporting States

Cohen believes that simply changing standards is not enough. To be effective, improvement efforts "must *help states* integrate new or revised standards into a coherent instructional system that aligns professional development, curriculum, formative assessments and data systems and other tools to promote continuous instructional improvement."[58] Accordingly, Achieve founded the ADP Network, a group of states that have committed to:

- align their high school standards and assessments with real-world expectations;
- require all high school graduates to take challenging courses that prepare them for life after high school;
- streamline their assessment system so that the tests students take in high school also can serve as readiness tests for college and work; and
- hold high schools accountable for graduating students who are ready for college or careers, and hold postsecondary institutions accountable for students' success once enrolled.[59]

The ADP Network currently includes 29 states that collectively educate almost 60 percent of American high school students, and Cohen expects it to grow even larger.[60] Although all ADP Network states have committed to the common policy priorities listed above, each state has developed an individualized action plan for carrying out the agenda. Thus, while alignment of each state's standards with external benchmarks should lead to increasing consistency in academic standards across states, differences will remain — in other words, the benchmarking work will not proceed neatly to the development of national standards. Once a critical mass of states has gone through the benchmarking process, Achieve will conduct a comparative study to see how the revised state standards differ from the original ADP benchmarks.

Achieve provides the states in its ADP Network with an extensive array of support services, including:

- helping participating states share strategies for policy design, implementation, and advocacy, by convening meetings and publishing reports;

---

57 Michael Cohen, telephone conversation with author, June 21, 2007.

58 Cohen, "Do We Need," 47 (emphasis added).

59 Achieve, Inc., "What Is the American Diploma Project Network?" website, www.achieve.org/node/604.

60 Michael Cohen, telephone conversation with author, June 21, 2007.

- providing expert advice and tools on assessment, curricula, and the design of state educational data systems;
- working with states to help them replicate the ADP benchmarking study in order to align their standards and assessments with external benchmarks; and
- providing tools and resources to help states develop and mobilize teams of advocates — including business leaders, teachers, principals, parents, and community groups — to support the ADP Network agenda at state and local levels.[61]

End-of-Course Exams

Achieve's agenda and projects are determined, to a large degree, by the ideas and needs of participating states. Moreover, its initiatives are piloted by small groups of interested states, so other states can learn from their experience before deciding whether to participate.

Thus, a group of nine ADP Network states have gone beyond benchmarking and alignment to develop a common end-of-course exam for Algebra II. While this is a logical extension of Achieve's coherent approach, Cohen says the Algebra II exam consortium was not part of some grand strategic plan. Rather, a few ADP Network states thought that having an Algebra II test would help them improve math achievement, and they decided that they could develop the test faster, more effectively, and for less money if they collaborated.

Achieve helped the states develop a Request for Proposals, which Ohio issued on behalf of the partnership. Ohio established a contract with Pearson, a large test publishing company, and other states could buy the test through that contract if they agreed to the terms of the partnership — which included agreeing to share the test results so that Achieve can issue comparative reports. States that do not wish to participate in the partnership are free to approach Pearson independently if they wish to purchase the test on a stand-alone basis.

Leveraging Financial and Technical Resources

Foundations and for-profit contributors provide much of the organization's funding. In addition, Achieve offers some programs to states on a fee-for-service basis.

Achieve works with a team of accountability experts, some of whom are hired as consultants on a project basis.

Federal Role?

Achieve is a state- and private-sector-led collaboration in which federal and local governments currently play little or no role. In a recent paper, Achieve President Michael Cohen argues that the federal government's role should be to provide states with technical support and financial incentives to:

1. align high school standards, assessments, curriculum, and graduation requirements with the demands of postsecondary education and work, and align standards and tests in grades 3-8 with these high school standards;

---

61 Achieve, Inc., "American Diploma Project (ADP) Network Services," website, www.achieve.org/node/303.

2. work together to develop and use common assessments aligned with common standards; and

3. improve the quality and rigor of state tests.

At the same time, Cohen points out; the federal government must ensure that NCLB's AYP requirements do not get in the way of these goals. States will need time and flexibility to align their tests with real-world demands, so those tests should not be tied to AYP under NCLB. Likewise, if states develop more rigorous assessments, they should not be "locked into using existing tests by AYP requirements."[62]

ADP has recently received favorable comment from U.S. Secretary of Education Margaret Spellings,[63] and it is possible that the feds will carve out a role in financially supporting ADP and similar collaborations.

---

62 Cohen, "Do We Need," 49.

63 Margaret Spellings, "A National Test We Don't Need," *Washington Post* (June 9, 2007), www.washingtonpost.com/wp-dyn/content/article/2007/06/08/AR2007060802259.html.

References

Achieve, Inc., "American Diploma Project (ADP) Network Services." www.achieve.org/node/303

Achieve, Inc., "What Is the American Diploma Project Network?" www.achiev_Hlt179858577_Hlt179858578eBM_1_BM_2_.org/node/604

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing.* Washington, D.C.: Authors, 1999.

American Psychological Association. "The Standards for Educational and Psychological Testing." www.apa.org/science/standards.html

Baker, Eva L., Robert L. Linn, Joan L. Herman, and Daniel Koretz, "Standards for Educational Accountability Systems." *CRESST Policy Brief 5* (Winter 2002).

Baker, Eva L., and Robert L. Linn, "Validity Issues for Accountability Systems." In *Redesigning Accountability Systems for Education*, edited by Susan H. Fuhrman and Richard F. Elmore, 47-72. New York: Teachers College, 2004.

Camara, Wayne J., and Suzanne Lane. "A Historical Perspective and Current Views on the *Standards for Educational and Psychological Testing*." *Educational Measurement: Issues and Practice* 25, no. 3 (Fall 2006): 35-41.

Cohen, Michael. "Do We Need National Standards?" Special issue, "No Child Left Behind: A Five Year Review," edited by Dick Clark, *Aspen Institute Congressional Program* 22, no. 1 (February 20-25, 2007): 45-50.

College Board, "ACES: Validity Handbook," www.colleg_Hlt179860296_Hlt179860297eBM_3_BM_4_board.com/highered/apr/aces/vhandbook/testvalid.html and www.collegeboard.com/highered/apr/aces/v_Hlt179860312_Hlt179860313hBM_5_BM_6_andbook/evidence.html

Cronbach, Lee J. "Construct Validation After Thirty Years." In *Intelligence: Measurement, Theory, and Public Policy,* edited by Robert L. Linn, 147-171. Urbana: University of Illinois Press, 1989.

Cronin, John, Michael Dahlin, Deborah Adkins, and G. Gage Kingsbury. *The Proficiency Illusion.* Washington, D.C.: Thomas B. Fordham Institute and Northwest Evaluation Association, October 2007.

Davis, Susan L., and Chad W. Buckendahl. "Evaluating NCLB's Peer Review Process: A Comparison of State Compliance Decisions." Paper presented at 2007 annual meeting of National Council on Measurement in Education, Chicago: April 12, 2007.

Finn, Chester E., Jr., Liam Julian, and Michael J. Petrilli, *To Dream the Impossible Dream: Four Approaches to National Standards and Tests for America's Schools*. Washington, D.C.: Thomas B. Fordham Foundation, August 2006.

Gais, Thomas and James Fossett. "Federalism and the Executive Branch." In *The Executive Branch,* edited by Joel D. Aberbach and Mark A. Peterson, 486-522. Oxford: Oxford University Press, 2005.

Gong, Brian, Testimony to the Commission on *No Child Left Behind* at its hearing on "State Standards: Assessing Differences in Quality and Rigor and How They Impact NCLB" held at Lesley University, Cambridge, MA on August 31, 2006. www.aspeninst_Hlt179958775_Hlt179958776iBM_7_BM_8_tute.org/atf/cf/_Hlt179948266_Hlt179948267BM_9_BM_10_%7BDEB6F227-659B-4EC8-8F84-8DF23CA704F5%7D/Brian%20Gong%20Testimony.pdf

Harris, William G. "The Challenges of Meeting the Standards: A Perspective from the Test Publishing Community." *Educational Measurement: Issues and Practice* 25, no. 3 (Fall 2006): 42-45.

Hess, Frederick M. and Chester E. Finn, Jr. "Crash Course: NCLB is Driven by Education Politics." *Education Next* 7, no.4 (Fall 2007): 40-45.

Jennings, John F. *Why National Standards and Tests? Politics and the Quest for Better Schools.* Thousand Oaks, CA: Sage, 1998.

Joint Commission, "Comprehensive Accreditation Manual for Laboratory and Point-of-Care Testing," www.jointcommission.org/AccreditationPrograms/LaboratoryServices/Standards/FAQs/default.htm.

Kosar, Kevin R. *Failing Grades: The Federal Politics of Education Standards.* Boulder: Lynne Rienner Publishers, 2005.

Linn, Robert L. "Following the *Standards*: Is it Time for Another Revision?" *Educational Measurement: Issues and Practice* 25, no. 3 (Fall 2006): 54-56.

Madaus, George F. "A Brief History of Attempts to Monitor Testing." *NBETPP Statements* 2, no. 2 (February 2001).

Middle States Commission on Higher Education, *Assessing Student Learning and Institutional Effectiveness: Understanding Middle States Expectations*. Philadelphia: Author, 2005.

Nathan, Richard P. "Reinventing Government: What Does It Mean?" *Public Administration Review* 55, No. 2 (March/April 1995): 213-215.

National Association for the Education of Young Children, "What research tells us about NAEYC accreditation," www.n_Hlt179862459_Hlt179862460aBM_11_BM_12_eyc.org/ece/1996/16.asp.

National Center for Education Statistics. *Mapping 2005 State Proficiency Standards Onto the NAEP Scales (NCES 2007-482)*. Washington, D.C.: U.S. Department of Education, June 2007.

Ravitch, Diane. *Left Back: A Century of Failed School Reforms.* New York: Simon & Schuster, 2000.

Diane Ravitch, "Reflections on the Math Scores in New York City and State," NYC Public School Parents blog, June 17, 2007," http://nycpublicschoolparents.blogspot.com/2007/06/diance-ravitch-reflections-on-math.html.

Spellings, Margaret. "A National Test We Don't Need." *Washington Post*. June 9, 2007. www.washingtonpost.com/wp-dyn/content/article/2007/06/08/AR2007060802259.html.

Toch, Thomas. *Margins of Error: The Education Testing Industry in the No Child Left Behind Era*. Washington, D.C.: Education Sector, 2006.

United States Department of Education Office of the Inspector General. *Compliance Requirements within Title I, Part A of the No Child Left Behind Act.* Final Management Information Report: State and Local No. 06-01 (March 29, 2006). www.ed.gov/about/offices/list/oig/auditreports/s06e0027.pdf.

United States Government Accountability Office. *No Child Left Behind Act: Education Assistance Could Help States Better Measure Progress of Students with Limited English Proficiency.* GAO-07-646T. Statement of Cornelia M. Ashby, Director, Education, Workforce, and Income Security Issues. Washington, D.C., March 23, 2007. www.gao.gov/new.items/d07646t.pdf.

U.S. General Accounting Office, U.S. Department of Education, Office of Inspector General, State Auditor's Office, State of Texas, Department of Auditor General, Commonwealth of Pennsylvania, and City of Philadelphia, Office of the Controller. *A Joint Audit Report on The Status of State Student Assessment Systems and the Quality of Title I School Accountability Data (SAO Report No. 02-064).* Austin: Texas State Auditor's Office, August 2002. www.ed.gov/about/offices/list/oig/auditreports/s14c0001.pdf.

Acknowledgments

I am indebted to Richard Nathan for his ideas, his support, and his determination to develop this project. Thomas Gais provided valuable suggestions and insights every step of the way. Thanks are also due to Caleb Offley for helping to develop the proposal and to both Caleb and David Shaffer for their comments on drafts of this paper.

Many others — including Paul Barton, John H. Bishop, Wayne Camara, Michael Cohen, Susan Davis, Martha Derthick, Harriet Ganson, Jeanne Hubelbank, Daniel Koretz, Holly Kuzmich, Suzanne Lane, Dane Linn, Robert Linn, Paul Manna, Townsend McNitt, Michael Petrilli, Paul Posner, Diane Ravitch, Judith Rizzo, Ray Scheppach, Hanna Skandera, Patty Sullivan, Mike Usdan, Phoebe Winter, Lauress Wise, and Joseph F. Zimmerman — were generous with their time and helped tremendously by responding to my queries by e-mail, over the phone, and in person.

Thanks to Barbara Stubblebine for managing the Symposium logistics, and to Brad Armour-Garb and Karen Armour for listening and making good suggestions.

Finally, this paper and Symposium would not have been possible without the funding provided by the Spencer Foundation and the Joyce Foundation.

Of course, any errors, omissions, mischaracterizations, or gross oversimplifications are entirely my own!

Allison Armour-Garb
October 9, 2007