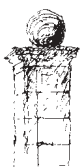


---

## STATED MEETING REPORT



### Education Reform: A Report Card

**Marshall S. Smith**, Program Director for Education, William and Flora Hewlett Foundation, and Professor of Education, Stanford University

Commentary: **Jerome Bruner**, University Professor, New York University

The following presentation was given at the 1858th Stated Meeting, held at the House of the Academy in Cambridge on April 10, 2002. At the event, the Academy honored Frederick Mosteller (Harvard University) and Howard Hiatt (Harvard Medical School) for their work in placing the health and welfare of children on the Academy's agenda. Richard Light (Harvard) spoke about Mosteller's distinguished career in educational research, and Jerome Kagan (Harvard) cited Hiatt's many accomplishments as director of the Academy's Initiatives for Children Program. A summary of the tribute appeared in the Summer 2002 *Bulletin* (pp. 9–13).

### Marshall S. Smith

Two years ago, I ended a seven-year stint as under-secretary of the US Department of Education. Tonight I would like to talk a little bit about quantitative studies—how I thought about them in the government and what I think might be done to improve them. I'll start with some history, going back forty years or so; then I'll talk a little bit about my sense of our progress. I will close with a brief report card on reform, as interpreted through the words of John Adams.

Forty years ago, in the 1960s, various activities in education were influenced by empirical studies. I will not argue that empirical studies drove such developments as the passage of Head Start and Title I. Lots of other things that went on in the sixties—including the civil rights movement and other social movements of those times—were far more impor-



Speaker Marshall S. Smith (William and Flora Hewlett Foundation; Stanford University)

tant than quantitative studies. Nevertheless, important quantitative studies were carried out, and they were part of the mix. A famous study of the effects of preschools, conducted in Ypsilanti, Michigan, contributed to (and was certainly cited during) the passage of Head Start. Among the researchers who were influential in that era, none exceeded Jerome Bruner. His landmark book *The Process of Education* (1960) was a crucial factor in the generation of a range of educational programs and experiments in the 1960s, including Title I and Head Start. During the early 1960s, I believe, Jerry was also a member of the President's Advisory Panel of Education.

Title I—the federally funded supplemental reading program for at-risk first-graders—changed the nature of evaluation in this country. A new federal provision—a Robert Kennedy amendment—required that every Title I project in the 14,000 local education agencies in the country had to be evaluated. The few words in that provision heightened thinking about evaluation in a major way.

In my own first research experience outside of the university, in the summer of 1965, I was on a team that helped to evaluate the Title I program in Boston. We spent most of that summer arguing about whether we should be measuring only outcomes—only student achievement—or whether

we should also be measuring some of the background variables and intervention processes that affected achievement. That argument continues, thirty-five years later. I think we know quite a bit more about it now than we did before, though people on both sides are still as passionate.

Many of you are familiar with James Coleman's report on *Equality of Educational Opportunity*, issued by the government in 1966. The findings of that report and subsequent reports building on Coleman's survey, especially *Racial Isolation in the Public Schools*, were instrumental in stimulating a large-scale social experiment: the widespread busing of students to achieve racial integration in US public schools. The consequences of that experiment, and of early evaluations of both Head Start and Title I—many of which were slightly negative—began to change a lot of people's thinking about what kinds of investments the country should be making in education, as well as in other areas.

I recall a phone call I got in 1969 or 1970 from Pat Moynihan, then domestic policy adviser for President Nixon. During his first stint at Harvard, from 1966 to 1969, Pat had been influenced by the Coleman report to believe that perhaps education didn't quite have the effect he once thought it had. Also, Head Start evaluations had led him to think that Head Start didn't quite have the intended effect. During the call, I was in my kitchen in Cambridge with two very young children, while he was in his office in the White House.

Pat had been advocating in the government for the negative income tax. He asked me whether I would rather put \$1,000 into a family to cover one year of Head Start for one of its children or put \$1,000 into that family to buy food, clothing, and shelter by means of a negative income tax. I conveniently ducked the question by saying I'd do both. But the question was an important one because it signaled an orientation toward thinking about what kinds of interventions would have the greatest effect—an orientation that was possible only because there had been empirical studies of at least some of the various domestic interventions.

Later on, in the 1970s, methods were developed for synthesizing the results of quantitative studies. Richard Light and Paul Smith started that off with a little article in the *Harvard Educational Review*. Gene Glass came up with the concept of meta-analysis, which advanced research synthesis dramatically. During the 1970s, we in education began to look at qualitative studies more—and in some ways, qualitative studies began to drive out empirical studies for the next fifteen years.

This was a phenomenon of some importance. We lost some of the momentum around empirical studies, I believe—but at the same time, we gained some real insights into theories of intervention and into the ways and processes of classrooms, schools, and other organizations. So on the one hand, our field drew a sharp distinction between qualitative and quantitative that should never be drawn, in my view. This led to an almost ideological battle in the field of education. But if you look carefully down the middle on this one, you will find that those qualitative studies provided valuable insights that allowed people to begin to piece together the findings of research on how students learn and how teachers teach, and to apply those findings to situations that were more complex. Those insights gained us a great deal.

Then came the 1980s. Many of you will remember the 1983 government report titled *A Nation at Risk*, which relied on international quantitative data to assess education policy. Increasing attention also was focused on national assessments and test scores. Great growth occurred in cognitive science, yielding useful theories on how people learn. The eighties also brought the class-size experiment—a massive randomized field trial that has had an enormous effect on policy over time.

As we moved through the 1990s and into the new millennium, almost every state in the nation adopted a framework of standards-based school reform. The intent of the reforms is to bring resources, policies, and assessments into alignment with standards that specify clear and explicit goals for student learning. The assessments are used for accountability

purposes. This is a package that has bipartisan support. It started with strong support by the Clinton administration and is now being supported by the current administration.

An increased emphasis on accountability, reflected in the standards-based reforms, has also influenced other parts of our society. Accountability based on quantitatively measurable outcomes has moved both the government and the private sector to become much more sensitive to the kinds of effects that can be measured. In many areas, including education, that has sometimes led to a narrowing of the kinds of outcomes people worry about, which may be a negative byproduct of the policy. We seem to value what we measure, rather than rigorously measure what we value. Consequently, if we assess only things that are easy or inexpensive to measure, we may end up placing value on the wrong things. This happens too often in education. Nonetheless, measures focus people's attention. The emphasis on empirically based accountability has created coherence out of incoherence in many instances, not least in the government.

The positivist belief in the value of empirical and verifiable findings has also increased attention to the empirical evaluation of education policies and practices. This—unfortunately, in my view—has resulted in a rash of dramatic statements about randomized field trials being the “gold standard” of



*Left to right: Jerome Kagan and Henry Rosovsky (both, Harvard University)*

research. This form of rhetoric often implies that other forms of research are inferior, rather than that they provide different kinds of data and different insights. In fact, the fascination with randomized trials seems to have been elevated to an ideological level by some. The National Research Council addresses the issues of different methodologies for different purposes in an elegant new report. On the other hand, the interest in randomized trials may be seen as a counterbalance to an equally ideological perspective of many in the late 1980s and early 1990s who regarded qualitative research as the only path to truth.

The concerns about effectiveness have not only heightened attention to methodological issues; they have also resulted in increased attention to theory. The National Research Council, for instance, has issued some excellent books on theories of learning, including how children learn to read and do mathematics.

Program evaluation has benefited from this. We are beginning to marry good and appropriate methodologies with better theory, and our evaluations are becoming more and more powerful and useful for policy development. Anthony Petrosino's work at the Academy on theory and evaluation is becoming very influential. We better understand the challenges of implementation. We are also seeing improvements in synthesizing the results of prior research. The inception of the Campbell Collaboration in the late 1990s was a formal way of beginning to approach the synthesis problem.

Of course, technology is changing many of the rules right now. It is changing our ways of modeling and our ways of organizing data. It is changing our access to data in dramatic ways. In the humanities and arts areas, the opportunities for new forms of research and analysis are extraordinary. Through technology, we are now able to do things we couldn't even dream of doing before.

At least four of the major events or findings in these areas can be traced back to Fred Mosteller. There are surely many other links with which I am not

familiar; I will note just four of his important contributions in areas that I have mentioned. He played a very significant role in interpreting the Coleman report; gave extraordinary legitimacy to the class-size study; fostered strides in synthetic analysis, both as Richard Light's mentor and as a supporter of the Campbell Collaboration; and made major contributions through his work on the National Assessment of Educational Progress in the early days.

In 1957 I took a course with Fred. Ever since, I've carried a Mosteller quote in the back of my mind, and I looked it up the other day in the 1953 edition of the *Handbook of Social Psychology*. What I found just goes to show that Fred hasn't changed his beliefs about the importance of carefully planned, theoretically driven research designs. Mosteller and Bush wrote, "In no circumstances do we think that sophisticated analytical devices should replace clean design and careful execution, unless very unusual economic considerations arise." Clear thinking should prevail.

Now, let me ask a rhetorical question: If we know so much about all of this, why don't we have better policy? Other countries appear to have strong linkages between improved knowledge and improvement in their schools. Back in the fifties, the National Science Foundation developed a set of very exciting and rigorous math and science courses in response to the challenge represented by *Sputnik*. For a good while during the sixties, lots of schools in the United States adopted those courses, and some actually still use them. In general, though, they began to die out around 1969 or 1970. Yet they were used in other countries for far longer. Materials based on US research are picked up and used by other countries fairly regularly. Yet in the United States, the curriculum materials developed through NSF investments in the 1950s and 1960s lasted only a while, and materials developed in the 1990s have been largely unused—some having been bought and then shelved by publishers that did not want them competing with their own textbooks.



Ellen Lagemann (Harvard Graduate School of Education) and Thomas Payzant (Boston Public Schools)

But the publishers are not the only culprits. The governance system can also be part of the problem. In the United States, we have an amazingly complex policy environment. California alone, for example, has seven different state agencies that influence the development and implementation of education policy. The elected state school officer and state governor are both Democrats, but they don't talk to each other, because they're battling over the turf. A variety of other groups out there are also in the fray. California has term limits, so there is almost no legislative memory. And the legislators seem to evaluate the quality of their term on the basis of the amount and number of legislative items passed rather than the effectiveness and coherence of the laws. This is not just a problem in California; state and federal legislators have the same disease. California also has government by public proposition, which means that anybody with a lot of money can put anything they please on the ballot. Consequently, a cacophony of chaotic provisions is placed into law, and that makes effective governance almost impossible.

On the other hand, as I learned during my years with the government, policymakers actually do listen. I was in the Clinton administration for seven years, in a policymaking role, and I don't think there was any major issue where quantitative re-

search didn't enter into the picture. There's no reason to think that it made a telling contribution, but people thought about it, worried about it, and looked at it. In some instances, research—for example, the Tennessee study of class size—really tipped the balance because it changed people's views in the Office of Management and Budget, the president's office, and Congress.

Generally, however, the effect sizes in research studies are small. If effect sizes are small, and if multiple studies are done, we are likely to get a distribution of effects that covers zero and goes into negative territory. As a consequence, anybody who wants to argue any position can base the argument on empirical research.

Let me spend a couple of minutes on a report card on education reform, just to give you some sense of where I think we stand today. I'm not going to relate it back too much to empirical research—just a little bit. It's a complicated picture. We have a set of standards-based reforms now that are in their early adolescence—nine, ten, eleven years old at best. In California, they're only three or four years old. So nationwide, these reforms are going through tremendous growing pains.

Although there are still many debates over the reforms, I believe they have begun to have some effect over time. Math scores on the National Assessment of Educational Progress have risen significantly in the fourth and eighth grades—by over a grade level—in the past six or seven years. That's quite a bit of progress. And that's not just for white students; it's also for African American, Hispanic, and Asian students.

We have individual states that do very well in the international studies. It is a difficult thing for us, as a country, to be compared with Singapore, or even with Holland, or Denmark, or Norway. One might think that Minnesota, for example, would compare more closely with Norway or Sweden than would the entire United States—or that some fairly small, well-off area of the United States might compare more closely with Singapore than would the whole

United States. When we do look at places that are well off and compare them with Singapore, our students do pretty well. They don't quite reach the level that the students in Singapore do, but they are competitive. When we look at how Minnesota does, compared with the Scandinavian countries, it actually does very well.

Some states have shown significant gains in many regards over the past few years. Texas, North Carolina, and Connecticut—all states that have pushed these standards-based reforms hard—show good gains in reading and mathematics. As for Massachusetts, we'll see—there's a big debate here. Virginia, Maryland, and other states have shown substantial gains. Nonetheless, we have a long way to go, especially for our least advantaged.

Many think that US education reform is taking us much too far in the direction of testing and assessment. Others think that perhaps we are not pushing hard enough. I was pleased to find support for my own views on the reforms in David McCullough's book *John Adams*, which I read on my plane trip here. I was struck by two quotes from Adams on education because they fit with my assessment of where we stand right now.

Here's one of them, written about 220 years ago: "A memorable change must be made in the system of education, and knowledge must become so general as to raise the lower ranks of society nearer to the higher."

I would venture to say that the lower ranks of society today are almost as low on the education totem pole as they were 220 years ago. We haven't changed that particular phenomenon in our society. We still have people at the bottom, and we can predict who they are, by and large. We know where they live. We know what the problems of their schools are—and we haven't done enough about it. So our reforms haven't done very well on that particular dimension.

The second quote that impressed me was from a letter John Adams wrote to John Quincy Adams at

around the same time. John Quincy had just been denied admission to Harvard, despite having demonstrated his extraordinary abilities. He'd been told that he would have to complete several months of tutoring in Greek with the Reverend Shaw in Haverhill in order to go to Harvard.

Apparently, Adams was a bit concerned that John Quincy would study too hard and get too involved in his Greek. He wrote to him, "The smell of the midnight lamp is very unwholesome. Never defraud yourself of sleep, nor your walk. You need not be in a hurry." What was essential, Adams advised, was an inquisitive mind. John Quincy must get to know the most exceptional scholars and question them closely: "Ask them about their tutors, manner of teaching. Observe what books lie on their tables. Ask them about the late War, or fall into questions of Literature, Science, or what will you."

There is a message of caution for us in Adams's prophetic words. We may be losing, in our passion for increasing achievement test scores in mathematics, reading, and science, the breadth of knowledge and understanding that needs to be developed in all students if they are to be productive citizens of our increasingly complex society.

## Jerome Bruner

I want to comment first on what I see as some of the deep wisdom in Mike's analyses, emphasizing some things that he didn't have a chance to discuss in detail. Then, after that, I want to offer a slightly different perspective with regard to where we Americans stand internationally in the World Education League. In doing so, I want to use a lesson I learned from Fred Mosteller, who has been my friend and mentor for many, many years, starting back at Princeton in another century. Fred likes to say, "In comparing performance scores, don't just pay attention to the means. Look at the variance too." Well, that's what I want to do: look at variability. I'll turn to that presently.



*Left to right: Commentator Jerome Bruner (New York University), Marshall S. Smith, and Vice President Louis Cabot (Cabot-Wellington LLC)*

But let me look first at some of the lessons that Mike set forth in his talk. The first was that there has to be a good fit between what a program for educational improvement is seeking to improve, and how it goes about assessing its results. In assessing a program, to put his point briefly, you can't just use any old standardized test. The assessment test needs to fit the objectives of your attempted intervention. There are no all-purpose assessment procedures that fit all needs. Adequate assessment has to be relevant to the theory behind the intervention program you are evaluating. You can't fly blind—but that, in effect, is what you end up doing if you don't design your assessment to fit the objectives of your intervention.

I remember this classic problem from the early days of the Physical Science Study Committee (PSSC), one of the first curriculum reform efforts of the 1960s, directed by Jerrold Zacharias and Franny Friedman at MIT. A lot of people urged them to evaluate the PSSC curriculum effort with the standardized physics tests available at the time. Zacharias replied boldly, "Hell no, we're not teaching that kind of physics." So PSSC developed new assessment procedures (with the help of the Educational Testing Service) geared to their own instructional objectives and to their own ideas about what it meant to understand physics. It was a real step forward.

Indeed, every educational intervention program has some underlying theory that shapes it, implic-

itly or explicitly, and the more explicit it is, the better the evaluation will be. Even when the theory is “simply” that small classes get better results than large ones, as in Fred Mosteller’s now famously successful Tennessee Study, there is an underlying theory that is not as simple as it seems. If you mindlessly attempt to replicate it, as they did in the state of California, the chaos is unbelievable. First of all, the way in which you set up small classes has to have some mind for who’s teaching. Teaching small classes requires skills in communicating.

So, what of California’s replication? They didn’t have enough teachers available, so they began hiring teachers willy-nilly—and got more than the usual proportion of weak and inexperienced ones. Small classes also require more classrooms, not just corridors or hastily remodeled closets and bathrooms. It’s not surprising that “reduced class size” didn’t bear fruit in California.

But there’s more to it than that. We don’t fully know why smaller classes work better, given the right conditions; we haven’t thought through the question. Is it that smaller classes lead to a different strategy on the part of the teacher, to different discourse patterns? Do they change the teacher-pupil authority relationship? We need a lot more theory to proceed wisely.

Let me give an example. I have been studying the famous preschools in Reggio Emilia in Italy. Here’s a surprising finding: when a teacher asks a child something, she waits for an answer. If the child has some difficulty answering, the teacher typically asks the other children in the class to help little Giovanna or Giuseppe figure out an answer, and a discussion starts. The context changes: knowledge seeking becomes communal. I’ve seen some astonishing scenes there. I’ve even started using this approach teaching graduate students. I’m still trying to think through the theory behind it, and even making a little progress. As Mike has been trying to tell us, people need to think about what they have in mind with their interventions. Then they’ll be able to evaluate properly.

Now I want to move on to Fred Mosteller's admonition about attending to variance. I'd like to look at it from the point of view of American performance on the tests now being widely used for comparing adult "literacy" in the nineteen most well-off countries in the world, including top-ranking America. These tests, devised by the Office of Economic Cooperation and Development, are thoughtfully designed and carefully translated into the different languages required. There are three subtests: one for ability to recognize prose, as in news stories and the like; another for "document literacy," or the ability to understand order forms, tables, and so on; and a third for "quantitative literacy," or knowing how to perform such tasks as balancing a checkbook and figuring a tip. Let's take a look at some findings from these tests.

First of all, as everybody knows, America doesn't do well on international tests. For example, among those nineteen well-off countries, we're ninth on the prose score, fourteenth on the document score, and thirteenth on quantitative—twelfth among nineteen on the composite score. You'd think, given our riches, we'd do better than that.

But where we undoubtedly lead the world is in variability, or dispersion. American standard deviations on all the tests are just about at the top. For example, on the prose test, we rank first in the size of our standard deviation; on the document test and on the quantitative test, we rank second. We lead the world in the standard deviation of composite scores—the most diverse country in the well-off world.

If you look at the test-score difference between the top tenth percentile and the lowest tenth percentile in each country, again we lead the pack. Our lowest percentile is way, way down; our top tenth is way, way up. America seems to have a gift for fostering maldistribution or inequality. No country in the civilized world can match us in terms of the maldistribution of wealth, the gap between rich and poor. And it seems, too, that none can match the gap we create between our most literate and

least literate countrymen. Ours is a diversity of inequality.

What about the history of all this? Are we getting better or worse in literacy, in comparison with other well-to-do nations? We can estimate this by looking at different age groups, and what comes out is not encouraging. Our youngest Americans—ages sixteen to twenty-five—rank fourteenth out of nineteen in the world on the composite literacy score. The age group twenty-six to thirty-five ranks eleventh. With the group that is thirty-six to forty-five years old, we go to fifth place. And the two oldest groups, ages forty-six to fifty-five and fifty-six to sixty-five, are second and third in the world ranking. So either America is falling behind, or the rest of the world is surging ahead, in literacy.

How much of this has to do with immigration? Our native-born Americans ranked tenth out of the seventeen countries on which there were immigration figures. Our foreign-born ranked sixteenth out of those seventeen countries. Our own past history suggests that when immigrants get segregated in caste conditions, as in our inner-city slums, second-generation “immigrants” continue to lag behind or even get pushed down further. So immigration is an issue, alright, though not an enormous one numerically.

I suspect, though, that the ones who are falling furthest behind world standards are poor blacks and poor second-generation Latinos. Yet there is an irony in this decline, for we know from intensive studies that with improved teacher expertise and classroom conditions, these groups can be greatly helped. If we in America are willing to do something about it, plenty can be done. But not much is being done. So our world position remains parlous—not to mention the conditions that such inequalities produce here in the United States.

If we follow Mike’s wisdom, we can begin to turn the tide, though we will have to take measures beyond the usual educational ones—for instance, assuring a more equitable distribution of wealth. After all, we know that the sense of helplessness

and despair produced by poverty is the worst block against improved school performance. On that basis, school reform without concomitant economic reform is simply not sufficient.

So, to return to Mike's message, we should indeed look more deeply and more theoretically at the causes of good and poor school performance, and propose reforms that take into account what it is that makes American society so prone to inequality—what it is that puts us in top position for variability in national literacy.

---

Remarks © 2002 by Marshall S. Smith and Jerome Bruner, respectively.

Photos © 2002 by Martha Stewart.