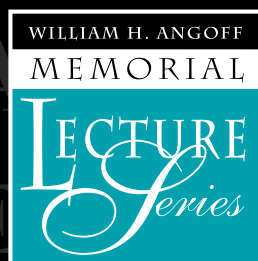


ND: RHETORIC AND REALITY OF
ALITY OF INTERNATIONAL LARG
AL LARGE-SCALE ASSESSMENT R
NT NO COUNTRY LEFT BEHIND: R
ND: RHETORIC AND REALITY OF
ALITY OF INTERNATIONAL LARG
AL LARGE-SCALE ASSESSMENT R

By Michael J. Feuer



William H. Angoff
1919–1993



William H. Angoff was a distinguished research scientist at ETS for more than 40 years. During that time, he made many major contributions to educational measurement and authored some of the classic publications on psychometrics, including the definitive text “Scales, Norms, and Equivalent Scores,” which appeared in Robert L. Thorndike’s Educational Measurement. Dr. Angoff was noted not only for his commitment to the highest technical standards but also for his rare ability to make complex issues widely accessible.

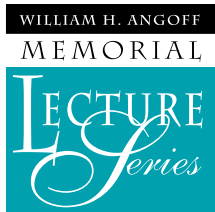
The Memorial Lecture Series established in his name in 1994 honors Dr. Angoff’s legacy by encouraging and supporting the discussion of public interest issues related to educational measurement. These lectures are jointly sponsored by ETS and an endowment fund that was established in Dr. Angoff’s memory.

The William H. Angoff Lecture Series reports are published by the Center for Research on Human Capital and Education, ETS Research and Development.

Copyright © 2012 by Educational Testing Service. All rights reserved. ETS, the ETS logo and LISTENING. LEARNING. LEADING. are registered trademarks of Educational Testing Service (ETS). 20032



NO COUNTRY LEFT BEHIND:
RHETORIC AND REALITY OF
INTERNATIONAL LARGE-SCALE ASSESSMENT



*The 13th William H. Angoff
Memorial Lecture was presented
at The National Press Club,
Washington, D.C.,
on February 18, 2011.*

Michael J. Feuer
The George Washington University
Graduate School of Education and Human Development

ETS
Research & Development
Center for Research on Human Capital and Education
Princeton, NJ 08541-0001

PREFACE

The 13th William H. Angoff Memorial Lecture was presented by Dr. Michael J. Feuer, Dean of the Graduate School of Education and Human Development (GSEHD) at The George Washington University. Feuer focuses on the nature and quality of comparative international assessments and outlines several caveats and cautions that policymakers should consider when interpreting the results. Because data from these assessments have a powerful effect on discussions about directions for education policy, Feuer helps us to understand the limitations of these data as well as their benefits.

Feuer focuses on three areas where policymakers need to take caution: inferring trends from snapshots, linking educational measures to national economic outcomes, and overrelying on international assessments for high-stakes national education policy decisions. While he agrees that the appropriate interpretation of comparative educational data can help educators and policymakers evaluate policy choices aimed at the improvement of teaching, learning and the governance of our schools, his lecture focuses on getting the right data — and getting the interpretation of those data right.

The William H. Angoff Memorial Lecture Series was established in 1994 to honor the life and work of Dr. Bill Angoff, who died in January 1993. For more than 50 years, Angoff made major contributions to educational and psychological measurement and was deservedly recognized by the major societies in the field. In line with Angoff's interests, this lecture series is devoted to relatively nontechnical discussions of important public interest issues related to educational measurement.

Ida Lawrence
Senior Vice President
ETS Research & Development
August 2012

ACKNOWLEDGMENTS

This paper is based on my William H. Angoff Memorial Lecture presented on February 18, 2011. Some sections of the paper draw on my presentation at the 2010 annual meeting of the American Educational Research Association. I am grateful to Randy Bennett and Ida Lawrence for the invitation to deliver the lecture and write the paper. Richard Atkinson, Avital Darmon, Steve Lerman, Iris Rotberg, Daryl Chubin, Cliff Adelman, Moshe Justman, Natalia Pane, Carol Kochar-Bryant, and Richard Rothstein provided helpful comments, and the thorough reviews by Richard Coley and Patrick Kyllonen were indispensable. Kim Fryer provided superb editorial assistance and Christine Frank stepped in at the final hour with excellent research assistance. I am grateful to these people and to my colleagues in and out of The George Washington University who attended the lectures and helped me shape my thinking as it has progressed. Remaining errors are, alas, my responsibility.

ABSTRACT

Few arguments about education are as effective at galvanizing public attention and motivating political action as those that compare the performance of students with their counterparts in other countries and that connect academic achievement to economic performance. Because data from international large-scale assessments (ILSA) have a powerful effect on policy discourse, it is important to understand the limitations of these data as well as their benefits. This paper concentrates on three sets of cautions and caveats in the interpretation of ILSA and proposes a modest research agenda aimed at reinforcing the validity and utility of cross-national educational comparisons. Although aimed primarily at American policy and research audiences, the arguments in the paper are relevant to other countries for which performance on ILSA influences education reform.

INTRODUCTION: NUMBER ONE IN THE WORLD?

In 1990, a political miracle occurred. For the first time in its 214-year history, the United States seemed poised and willing to write into federal law a set of national education goals. Friends from overseas may have not fully grasped the significance of this event and others surely wondered what had taken so long. How, indeed, did the United States accomplish its extraordinary human capital revolution (Goldin & Katz, 2008), adopt the earliest recorded policy of universal access in world history, and educate a larger share of its increasingly diverse population than in any other country *without* a coherent definition of its national educational ethos? The answer to that question — which involves the politics and economics of federalism (e.g., Inman & Rubinfeld, 1997; Manna, 2006; McDonnell, 2005), the culture of American *exceptionalism* (e.g., Schuck & Wilson, 2008), the meaning of *rationality* in public policy (Feuer, 2006), and the history of the *grand experiment* in American education (Cremin, 1990) — would obviously require more than a short paper to address adequately.

The effort to write acceptable national goals into law can be understood retrospectively as the natural continuation of a discourse that was launched by the landmark report *A Nation at Risk* (National Commission on Excellence in Education [NCEE], 1983). Following on that enormously effective wake-up call, which alerted Americans to impending calamity if action was not taken to shore up our educational foundations, it was easier — for a brief moment in time — to find receptor sites across much of the political spectrum for the notion that national goals could be the basis for needed reforms.

One of the best histories of this era is in the work of Maris Vinovskis. The process began in 1989 with a summit meeting of President George H. W. Bush and the National Governors Association, then under the leadership of Arkansas Governor Bill Clinton, in Charlottes-

ville. The National Educational Goals drafted there were announced by President Bush in his January 1990 State of the Union address, and within six months the National Education Goals Panel was established to monitor progress towards the goals. The six goals were then incorporated into the landmark “Goals 2000” legislation, passed in 1994, which added two additional goals (Vinovskis, 1999).

Granted, ideologues on the right and left continued to fear, for different reasons, the encroachment of federal bureaucracy on the local definition of education and the local governance of schools; the specter of a national *curriculum* remained a looming threat. Nonetheless, the national goals in the Goals 2000: Educate America Act were written in language simple enough to deter overwhelming opposition and yet were meaty enough to be more than apple pie and motherhood (see Swanson, 1991). They covered a number of key issues — early childhood readiness, high school completion, student achievement and citizenship, science and mathematics, adult literacy, school safety — and by implying a commitment to increased equity *and* higher standards, the goals echoed a theme that had been pervasive in American educational history for over a century (e.g., Cremin, 1990).

A complete analysis of the substance and politics of the goals is beyond my scope here. I would like to focus, rather, on the rhetoric of international competition and how it affects policy thinking. Consider, therefore, Goal 5, one of the more poignant and endearing examples in American history of the triumph of exuberance over empiricism: “By the year 2000, U.S. students will be first in the world in science and mathematics achievement” (Goals 2000: Educate America Act of 1994).

It is worth recalling that when this goal was first published some wry observers thought there had been a

printing error and that the intent was for the United States to be first in the world *in 2,000 years*. Humor aside, the rhetoric was not really strange or unique. Twenty years earlier, President Richard Nixon, in arguing for a set of dramatic new fiscal and monetary policies, warned that “... whether this Nation stays *number one in the world’s economy* or resigns itself to second, third, or fourth place ... depends on you, on your competitive spirit, your sense of personal destiny, your pride in your country and in yourself” (Nixon, 1971, italics added). In response, American folklorist Garrison Keillor proposed an alternative plan to ensure that the United States remained on top: “For the 28th straight year, the U.S. was named No. 1 country by a jury of more than 300 members of the Association of World Leaders. President Nixon made remarks to the press that were not for publication. He wore a necktie with the inscription ‘El Numero Uno.’ ... The U.S. has dominated the world scene in this century, though it still trails the Roman and British Empires and the Mongol Hordes in total wins ...” (Keillor, 1971, p. 35). Keillor may not have known the data, but he had a good intuition that President Nixon’s exuberance even about the past was off the mark: The United States had *never* really been “numero uno” on most indicators that matter.

Since then, American education policy and reform initiatives have benefited from a nearly steady stream of reports and headlines chronicling the erosion of U.S. competitive stature in the world and the decline in quality of the public school system. One of the most memorable and influential reports in the annals of government-sponsored policy documents — commissioned, ironically, by a president who hoped it would justify the *end* of the federal role in education — invoked the most compelling and chilling language: “If an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might

well have viewed it as an act of war ...” (NCEE, 1983, p. 9). Report writers ever since have suffered from a disorder known as *Nation-at-Risk envy*, with most of them failing to meet that high standard of rhetorical flourish.

Not that the basic theme was all that new. The possibility that our competitors in the world might catch up to or — perish the thought — surpass us has frequently hung over our political psyche like a dark and ominous cloud. The path-breaking work of Joseph Mayer Rice, for example, is understood to have been influenced by his time spent in Germany, and though he did not explicitly warn of an economic or military threat posed by superior European pedagogy, he was sufficiently moved by his experiences there to develop what may have been the first real comparative assessment methodology (Rice, 1893). More recently, the launch of Sputnik in 1957 shook the U.S. education policy world and spurred the rapid and robust investment of federal dollars toward science and mathematics education, a policy that dovetailed naturally with post-war thinking about American research (see, e.g., Atkinson & Blanpied, 2008; Atkinson & Pelfrey, 2010). In the 1970s, it was the threat from Japan, manifest in the flood of VCRs and Toyotas that washed up on American shores. And today, the emergence of the so-called BRIC economies — Brazil, Russia, India, China — has many American policy experts and journalists counting the days until our global economic presence fades into distant memory. Clouds gather, storms rise, the nation reacts ... and then, somehow, the sun comes out again.

Except that some clouds last longer and have a sustaining effect on political morale and policy discourse. It is not an exaggeration to say that the pillars of the contemporary reform movement in American education — goals, national standards, accountability, testing, performance measurement — were erected in the aftermath

of the publication of *Nation at Risk*, and that our current efforts to implement common core standards and new systems of assessment are a natural extension of the narrative that compares us to foreign competitors today and to ourselves in an earlier time. Golden-era thinking is pervasive (see, e.g., Gardner, 2011), even if, as the literary critic Randall Jarrell once observed, “The people who live in a golden age usually go around complaining how yellow everything looks” (Kirsch, 2011, p. 96).

Lest this begin to sound like an apology for the status quo, let me affirm now my belief that moving toward higher and common standards is a good thing for American education. And if the only way to move the country onto a path of genuine improvement in its education is to indulge occasionally in bombastic rhetoric, that may be a price worth paying. On the other hand, my preference would be for a sustained discourse based on empirically sound evidence; wild swings of despair and exuberance do not strike me as an enlightened or sustainable approach to public policy. Hot rhetoric engenders backlash or boredom, neither of which is likely to maintain the nation’s focus on the real problems that beset our schools.

My point is that precisely because the reform movement is so crucial to the nation’s future it deserves a system that privileges objective evidence over ideological rhetoric, that places facts ahead of advocacy, and that admits to the possibility that some of the problems have no instant or obvious solution (see also Feuer, 2006). Instead, it is becoming dangerously commonplace to hear even the most thoughtful and prestigious organizations joining the chorus of politicians and pundits in the cacophony of despair (e.g., Klein, 2011; National Academy of Sciences, 2007; Ripley, 2010). Even typically moderate and reasonable voices seem unable to resist the temptation to indulge in scary comparative rhetoric

to further political goals. The traditionally more temperate Secretary of Education Arne Duncan opined recently that “we live in a globally competitive knowledge-based economy and our children today are at a competitive disadvantage with children from other countries. That is absolutely unfair to our children and that *puts our country’s long-term economic prosperity absolutely at risk ...*” (Associated Press, 2010, para. 4, italics added).

Rumors of the death of the American economy brought on by the decaying system of public education are, to paraphrase Mark Twain, a bit exaggerated. But while the United States still probably ranks as number one in the world in the capacity for naïve or extravagant educational rhetoric, it is worth noting that other countries are catching up. In Germany, for example, a poor showing in the 2000 Programme for International Student Assessment (PISA) prompted this statement from the Federal Minister for Education and Research: “The findings ... are alarming. A country with the economic and political significance of Germany belongs *at the top of the league* and cannot be satisfied with an education system performing at the OECD average level — never mind below it” (Bulmahn, 2002, para. 1, italics added).

Similarly, the condition of education and the effectiveness of major reforms in other typically high-performing countries, such as Japan, causes tremors of anxiety: “In striking contrast to the international acclaim during the 1990s for Japanese schools’ instructional excellence and solid curricular contents, the Japanese media, scholars, politicians, and the public continued to perceive their country’s schooling as steeped in a dire crisis ... [and the public debate] framed education reform as having direct bearing on the nation’s rise and fall, evoking a sense of urgency for immediate intervention ...” (Takayama, 2007, p. 423).

The fact that so much international comparative rhetoric is strident and overwrought does not mean there is no reason for concern. On the contrary, changing global economic and demographic conditions, promising new theories of teaching and learning, and the evolving nature of work and its skill requirements impel modern societies to rethink their education systems. Sound international comparisons — of status and trends — can be powerful tools in understanding the challenges and in shaping sustainable reforms. My goal in this paper, therefore, is not to disparage the concept of international

large-scale assessment (ILSA), but rather to preserve its utility by focusing attention on sources of potential misunderstanding or misinterpretation. I focus on three sets of caveats and cautions: the problem of inferring trends from snapshots, the problem of linking educational measures to national economic outcomes, and the hazards of overreliance on international assessments for high-stakes national education policy decisions. In conclusion, I will offer a short list of questions for an ongoing research program.

SNAPSHOTS AND TRENDS

It is tempting to compare various countries' average educational performance at a given point in time, for their whole populations or disaggregated by socioeconomic and demographic criteria, and to make inferences about eroding (or improving) relative standing (e.g., Ripley, 2010). But other comparative exercises, especially in economics, provide cautionary lessons. Consider, for example, the productivity growth slowdown in the U.S. economy that began in the late 1960s and stretched through much of the 1970s. Because productivity growth is one of the most important indicators of the health and stability of an economy and is linked so crucially to measures of quality of life and social mobility (e.g., Baumol, Nelson, & Wolff, 1994), sluggish growth rates in the United States were legitimately a cause for concern. Some analysts and policymakers reacted perhaps too quickly, though, embracing management reforms aimed at mimicking the so-called Japanese miracle and — at times rather caustically — attributing the U.S. slowdown to performance of American students on international standardized tests.

Closer scrutiny led to more nuanced findings. For example, as indicated in Table 1, compared to other industrialized countries, U.S. productivity growth had in fact *never been the highest in the world*; its ranking among developed countries had always been well below the top. As argued forcefully in one of the most comprehensive analyses of the economic data, "... in no peacetime period since 1880 has the U.S. been outperformed by less than five countries..." (Baumol, Batey Blackman, & Wolff, 1989, p. 87). In terms of economic performance over time, although the aggregate productivity growth rate did slow between 1963 and 1980, trends for manufacturing productivity growth exhibited a marked upturn starting in the 1980s.

Table 1

	1870–1880	1900–1913	1929–1938	1950–1960	1970–1979
Canada	2.2	2.7	0	3.1	1.8
Finland	1.3	2.4	1.9	4	2.6
France	2.3	1.8	2.8	4.4	4.1
Germany	1.5	1.4	2.3	6.6	4.5
United Kingdom	1.6	0.9	0.9	2.2	2.8
United States	2.3	2	0.7	2.4	1.9
Superior to United States among this group	0	2	4	4	4
Superior to United States among 16 industrialized countries	1	6	11	14	13

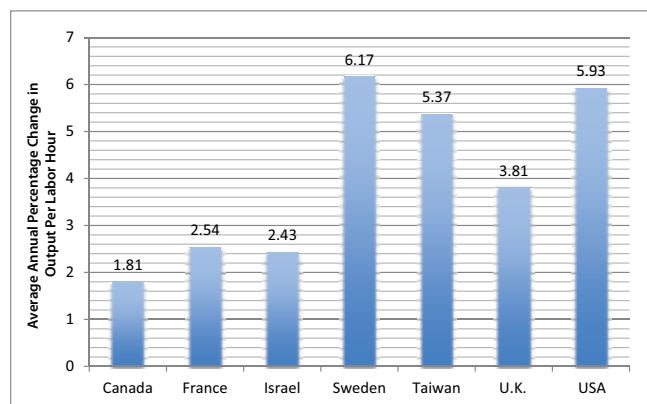
Note: Data from Baumol, Batey Blackman, & Wolff (1989).

More recent data echo the basic point. As shown in Figures 1 and 2, trends in productivity growth are complex, and there is little evidence of systemic failure of the U.S. economy over any significant historical period. Other countries have outperformed us at times, but they have also suffered traumatic jolts while the U.S. trend has been steady and positive. One of the best summaries of this complex situation is in a review of the book by Baumol, Batey Blackman, and Wolff that appeared in the flagship journal of the American Economics Association: "If there ever was a topic for which an understanding of the long run mattered, productivity performance is surely it. The experience of the United States with the productivity slowdown since the 1960s cannot be adequately understood without placing that experience in

the perspective of a century of productivity growth ... While the slowdown across the 1970s was big enough to warrant attention, it did not place America on a trajectory very different from that of the previous century” (Williamson, 1991, p. 54).

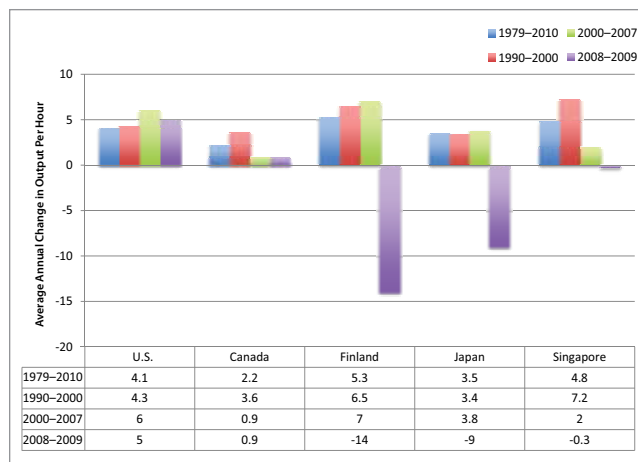
These findings suggest why it is important to place short-run data in their longer-term perspective. Attention to the time series and disaggregation by sector provided the needed perspective, although sadly most of the serious analysis came long after the scary headlines had already made their mark. The revised interpretation of trends is not a basis for naïve optimism, but rather provides policymakers with a stronger base of evidence to consider relevant corrective strategies.

Figure 1
Productivity Growth, Manufacturing Sector, 2000–2004, Selected Countries



Source: Data from Lach, Shiff, & Trajtenberg (2008).

Figure 2
Trends in Manufacturing Productivity, Selected Countries, 1979–2010



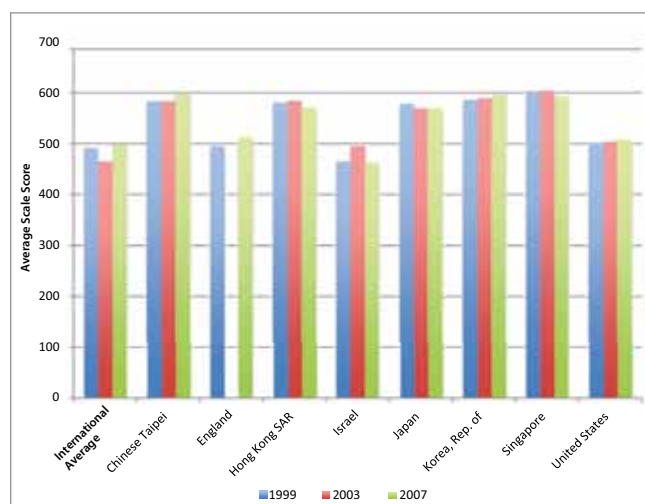
Source: U.S. Bureau of Labor Statistics, 2010.

Although interesting in its own right, the example of productivity growth in the short and long term holds lessons for education policy generally and the interpretation of ILSA specifically. Consider, for example, the performance of American students on international comparisons of mathematics achievement, about which the conventional rhetoric emphasizes stagnation and decline (e.g., Hanushek, Peterson, & Woessmann, 2010). It is perhaps disheartening to see where U.S. eighth-graders rank among the countries participating in Trends in International Mathematics and Science Study (TIMSS) — ninth among 36 — but the fact is that U.S. math achievement has actually been improving over time and, as the data show, we were never at the top (e.g., Lovelless, 2011). As shown in Figure 3, U.S. performance has been steady while other countries, even Singapore and Japan, have experienced dips. Granted, the gap that separates the United States from some of the highest performing countries is significant (i.e., we could be doing better). But the rhetoric of precipitous decline is

simply unfounded: In the First International Mathematics Study (FIMS) in the mid-1960s, the United States ranked at or near the bottom (Medrich & Griffith, 1992; for a contemporary critique of the rhetoric of decline, see also Kilpatrick, 2011).

Figure 3

Average Performance of Eighth-Graders, TIMSS, 1999–2007, Selected Countries



Source: Data from NCES (1999, 2003, 2007b).

The most recent data from the National Assessment of Educational Progress Trial Urban District Assessment (TUDA) provide further evidence that the rhetoric of despair is exaggerated. At grade eight, average mathematics scores were higher in 2011 than in 2009 for public school students in the nation, large cities, and six of the 18 urban districts that participated in both years. In comparison to 2003, scores were higher in 2011 for nine of the 10 districts that participated in both years, as well as for large cities and the nation (National Center for Education Statistics [NCES], 2012). Again, does this mean that the United States should now become

complacent about its commitment to improved mathematics education? Certainly not, but the rhetoric of impending catastrophe in the popular and professional discourse is overstated, at least in terms of available trend data.

If the outcomes of education, as measured by international assessments, are a major source of public consternation and political handwringing, participation and attainment data are no less alarming by the ways they are typically reported. Here, too, careful analysis of comparative data is both necessary and helpful, but getting the numbers right — and using the right numbers — are minimal prerequisites. For example, there is some confusion about where the United States ranks and how its standing has shifted compared to other countries on key indicators such as college attainment. A commonly held view is that U.S. attainment rates are flat or declining compared to other countries: “The percent of the American population with a postsecondary credential or degree has remained flat for 40 years, in spite of the dramatic economic and social changes during that period. Meantime, higher education attainment in the rest of the world has increased — in some cases at dramatic rates” (Lumina Foundation, 2009, p. 3).

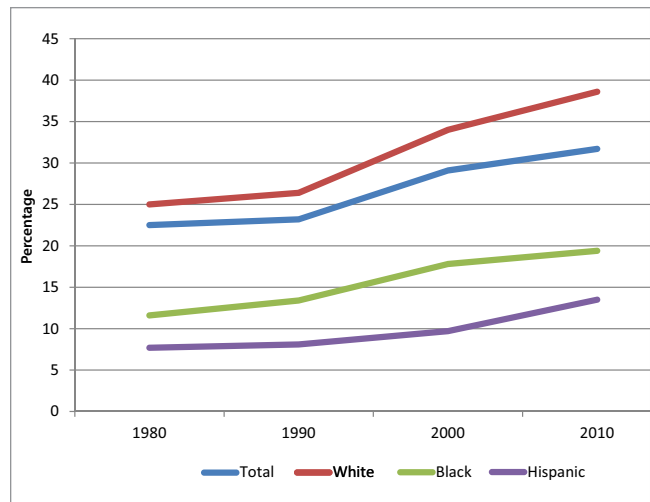
But the data are more complicated, and in fact tell a different and at times confusing story. For example, even the compilations by the Organisation for Economic Co-operation and Development (OECD) are confounded by fundamental problems in defining variables and are therefore the target of considerable methodological critique (e.g., Adelman, 2009; Hauptman, 2011). In the original printed version of its 2008 report, for example, OECD ranked the United States first among 30 OECD

countries in attainment of bachelor’s degrees for all age groups except the 25–34 bracket, where it ranked second (see Adelman, 2009, Table 1, p. 15). In subsequent revisions online, however, the numbers changed, and for the 25–34 year old age bracket U.S. performance was surpassed by Denmark, Korea, Netherlands, Norway, and Sweden, while the United States remained at the top of the ranking in all other age categories (see OECD, 2008). What is usually missing in these displays, though, is any reference to the *denominator effect* on attainment and completion rates — the possibility that the estimates are biased upwards in countries experiencing population declines in the relevant age groups (C. Adelman, personal communication, March 3, 2012).

The historical trend in degree completion within the United States has also been mischaracterized: Whereas the Lumina Foundation reported that “attainment rates are rising in almost every industrialized or post-industrial country in the world, *except for the U.S. ...*” (Lumina Foundation, 2009, p. 2, italics added), data from the U.S. Census Bureau as compiled by NCES show that between 1980 and 2010 the percentage of Americans aged 25–29 who had received a bachelor’s degree or higher rose from roughly 23% to roughly 32%. Furthermore, among White students, the rate increased from about 25% to 39%; among Black students, from 12% to 19%; and among Hispanic students, from about 8% to 14% (NCES, 2011a). These trends are shown in Figure 4. It may still be true that some other countries have experienced a sharper increase, especially in recent years, but the U.S. trend is by no means alarming or embarrassing.

Figure 4

Percentage of 25–29 Year-Olds with Bachelor’s Degree or Higher, 1980–2010

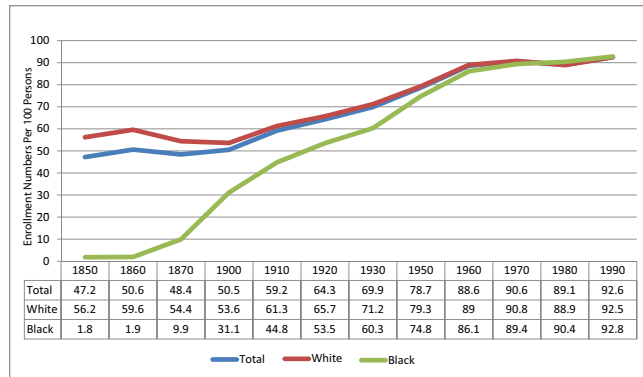


Source: Data from NCES (2011a).

Indeed, perhaps the most significant caution about these types of data relates to demographic trends. Many industrialized countries have made impressive strides in recent years to include increasingly large and diverse segments of their population in their education systems. In the United States, though, this policy, or philosophy really, has been in effect for much longer. Compulsory attendance began in the United States in the late 19th century, decades before the idea took hold in any of our economic competitors, which is one factor explaining the remarkable differences in enrollments and years of schooling across countries over time (Goldin & Katz, 2008). This history, for the United States, is captured in Figure 5.

Figure 5

U.S. School Enrollment of 5–19 Year-Olds Per 100 Persons, 1850–1990

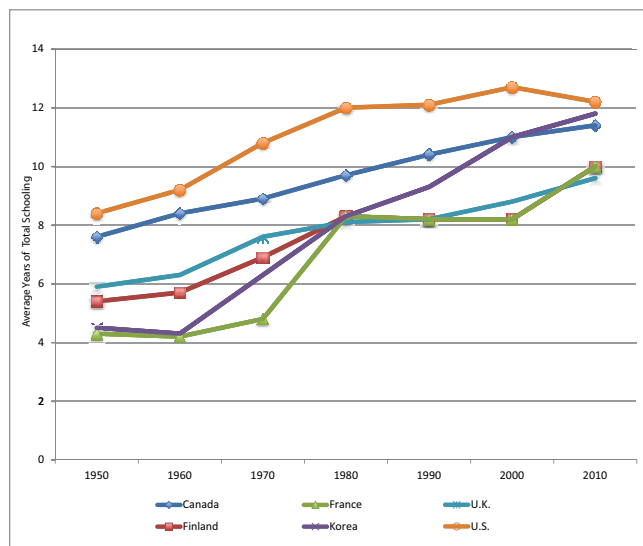


Source: Data from Snyder (1991).

And as shown in Figures 6 and 7, the United States has led the way both in terms of average time in school and in participation rates, a fact that can easily be overlooked if one focuses only on the most recent slice of time.

Figure 6

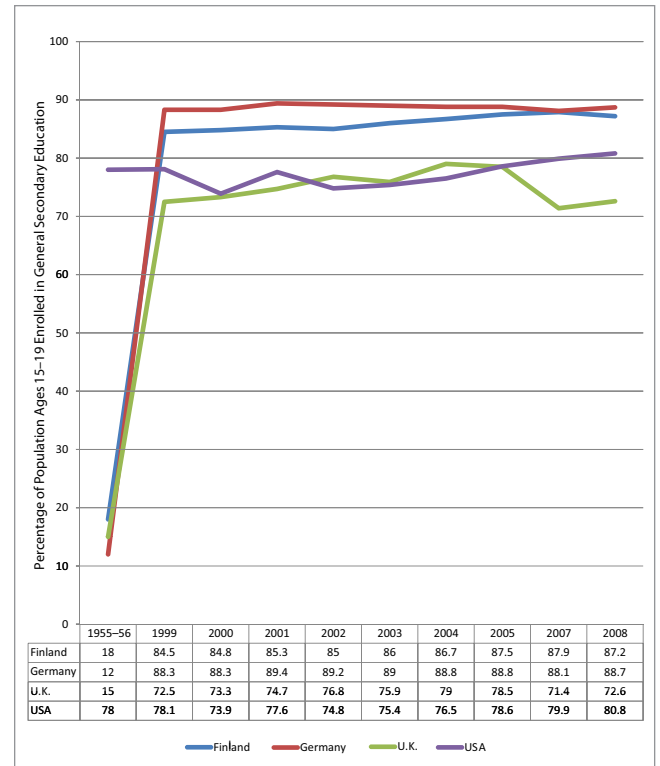
Average Years of Total Schooling, Selected Countries, 1950–2010



Source: Data from Barro & Lee (2010).

Figure 7

Participation in Secondary Education, Selected Countries, 1955–2008

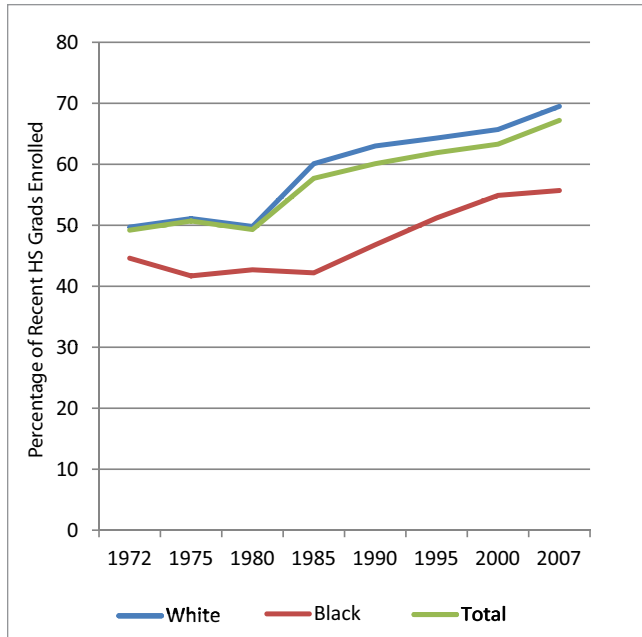


Source: Data for 1955–1956 from Figure 1.7 in Goldin & Katz (2008) and data for 2008 from OECD (2008).

Similar cautions are in order when considering post-secondary enrollment data: NCES reported that “between 2000 and 2009, undergraduate enrollment in degree-granting postsecondary institutions increased by 34%, from 13.2 to 17.6 million students. Projections indicate that it will continue to increase, reaching 19.6 million students in 2020” (NCES, 2010b, para. 1). The trend in enrollment has been steadily positive, since 1976, for all population groups, as shown in Figure 8 (see also Feuer, 2011).

Figure 8

College Enrollment: Recent Trends



Source: Data from NCES (2010a).

If today the percentage of young people enrolled in traditional secondary education in the United States is slightly lower than in Germany and Finland, that should be taken as a good sign that those countries are shedding their historically elitist traditions in favor of greater inclusion and heterogeneity. It will certainly be interesting to watch how they manage their newly diverse educational environments, and the prospects are not necessarily rosy: “... as the European Union falls into economic disarray, older dreams of tolerance and social inclusion have lost ground ...” (Gordon, 2011, para. 2). In any event, why U.S. policymakers and politicians should be anxious, rather than relieved, to see greater numbers of people around the world benefiting from education is something of a mystery.

EDUCATION AND THE ECONOMY

Implicit in the preceding discussion of problems in the interpretation of comparative data are links between educational and economic measures. The literatures of economics and sociology are rich in theoretical and empirical evidence of the relationship between education, earnings, social mobility, and other indicators of quality of life (e.g., Becker, 1964; Goldin & Katz, 2008; Sewell, Hauser, & Featherman, 1976). These studies are primarily focused at the micro level (i.e., they provide evidence for the individual and social returns to investments in human capital). The preponderance of evidence on the effects of investments in education on lifetime earnings, social mobility, and longevity is incontestable.

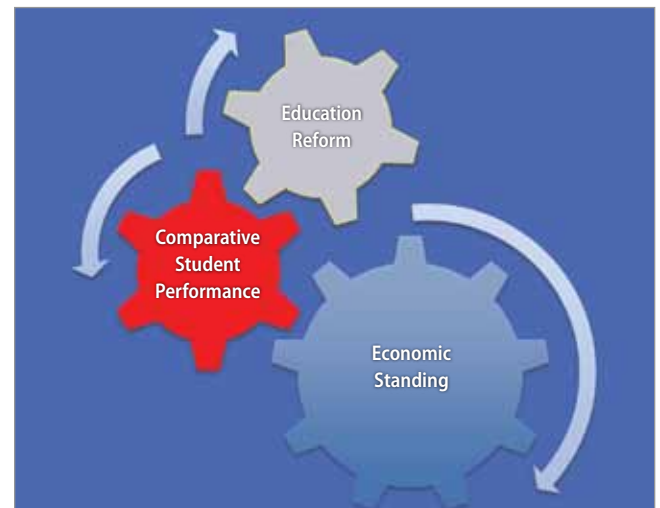
But international educational comparisons based on ILSA programs are often the basis for claims about *macro*-level economic performance (e.g., productivity, unemployment, and global competitiveness), which depend on a host of variables outside the realm of academic achievement and educational attainment. (Even the most ardent proponents of education reform as the foundation for economic growth do not attribute the 2008 meltdown in the housing and finance markets to poor performance of high school students on standardized tests.) Indeed, much of the ILSA rhetoric hinges on the assumption that educational achievement is a principal *cause* of economic standing and a nation's international competitiveness (e.g., Hanushek & Woessmann, 2011). As compelling as this logic may be, a number of cautions and caveats are warranted.

A pictorial representation of the implied relations between individual educational achievement and economic performance is in Figure 9, which suggests that education policy initiatives (labeled *education reform*) lead to changes in educational output (labeled *comparative student performance*), which in turn bring about measur-

able improvements in national economic indicators such as productivity or competitiveness. A body of evidence suggests why increased proficiency in academic subjects is a necessary condition for improved labor market opportunities or, in the language of human capital theory, why both private and social returns to educational investments are positive (e.g., Levy & Murnane, 2005). But the idea that educational performance as measured by ILSA or other standardized academic tests is the sole or even the principal determinant of national economic performance stretches credulity and flies in the face of counterfactual historical evidence.

Figure 9

Implied Relationship Between Individual Educational Achievement and Economic Performance



A striking example comes from the recent history of Japan. When U.S. productivity growth slipped in the 1970s (down to 1.9% per year from its long-term historical average of roughly 2.3%), Japan experienced a 5% average annual increase, and it became tempting to attribute the difference to the relative superiority of Japanese education. When results from one of the early versions of ILSA, the Second International Mathematics Study (SIMS), were reported starting in the early 1980s,

the finding that Japan was one of the highest-achieving countries at the final year of secondary school (International Association for the Evaluation of Educational Achievement [IEA], 2011) came as no surprise, especially to those who had already concluded that the sluggishness of the American economy must be a function of our lousy schools.

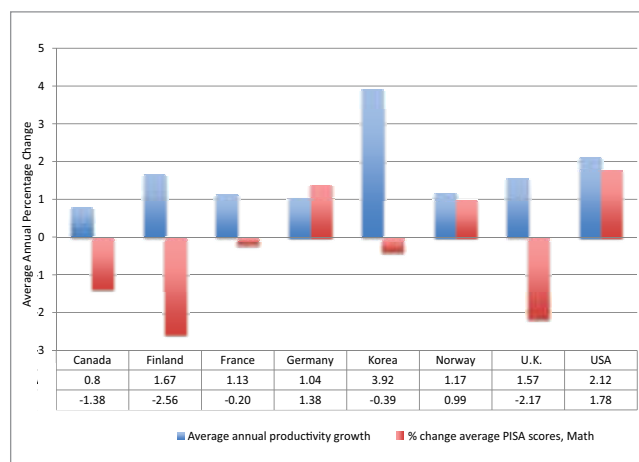
Counterfactual evidence was available (e.g., that as impressive as the Japanese growth rate was in the 1970s, it was significantly down from its even more remarkable 9.96% per year during the preceding decade [Williamson, 1991]). But it was hard for such questions to be heard against the drumbeat of media and political attention to the possibility that the United States was about to lose its global economic hegemony. The rhetoric in *Nation at Risk* (NCEE, 1983) was obviously influenced by — and fueled — this casual empiricism, notwithstanding a host of known methodological and conceptual problems (e.g., Cremin, 1990; Koretz, 1987; Stedman & Smith, 1983).

Let us remember, too, that when the Japanese economy experienced its substantial downturn in the 1990s and into the current decade (Bureau of Labor Statistics (BLS), 2010; Gardner & Ivancevich, 1994), there was an eerie silence about whether this might have been caused by the Japanese education system. If U.S. economic indicators are pinned on test scores of American students, shouldn't the same logic apply elsewhere?

Clearly, then, timing is key to untangling connections between comparative educational and economic performance. Most eighth-graders in 1980, when the SIMS data were collected, entered the labor market four or more years later, which means their impact on aggregate productivity statistics would only be measurable starting

in the mid-1980s at the earliest. During that period, however, average annual productivity growth in the United States began to recover, ultimately reaching a rather strong 3% per-year average for the decade (BLS, 2010), which was hardly an indicator of chronic economic malaise. Japan was still higher, at 3.8% per year (albeit considerably lower than in the previous period), while Germany, Norway, Canada, Denmark, and Sweden — countries that typically outperform the United States on ILSA — all recorded productivity growth rates lower than the United States in that time period. Figure 10 provides another glimpse into why the popular rhetoric that connects scores on ILSA to economic standing can be misleading: In Finland, for example, the improvement in PISA seems to have been accompanied by a sharp decline in output per hour worked, at least during the most recent decade.

Figure 10
Comparing PISA Scores and Productivity Growth, Selected Countries, 2000–2009



Source: Data from author's computations of OECD (2009).

A final and extremely important caveat concerning the presumed causal link between academic and economic performance relates to the effects of poverty on

academic performance. On the one hand, there is abundant evidence that poor children can learn and advance academically, if they are afforded decent opportunities at school and at home: See, for example, Figure 11, which shows steady gains in math achievement even for poor children.

Figure 11

NAEP Mathematics, Eighth Grade, by Eligibility for Free or Reduced-Price Lunch, 2003–2011



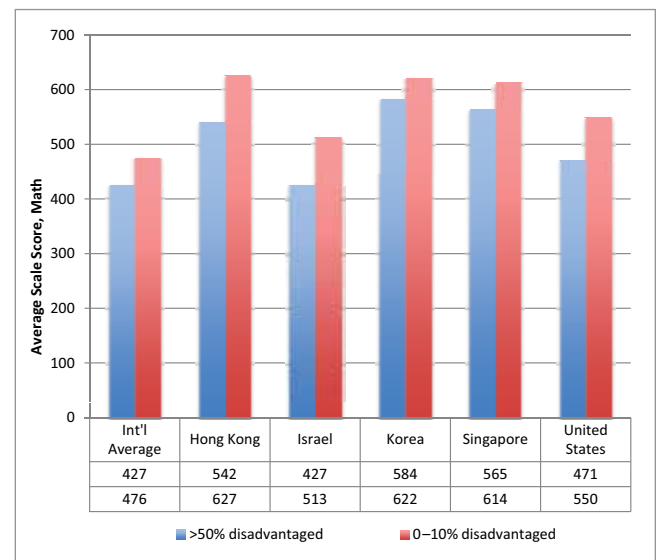
Source: Data from NCES (2011b, Figure 28).

On the other hand, as shown in Figure 12, there is equally compelling evidence — worldwide — that poverty (as measured by concentration of poor students in schools) suppresses average academic performance. Indeed, there are significant externality effects of poverty as suggested by research conducted in the 1990s: “School poverty depresses scores of all students in schools where at least half the children are eligible

for subsidized lunch and seriously depresses the scores when more than 75 percent of students live in low income households” (Puma, Jones, Rock, & Fernandez, 1993). At the postsecondary level, as well, there are clear patterns regarding enrollment, persistence, and indicators of demographic and economic status, as shown in Figure 13. These data suggest that persistence (i.e., whether students who enroll in undergraduate education complete their degree requirements) is correlated with demographic background variables such as race, ethnicity, socioeconomic status, and educational level of parents. The direction of correlation is consistent with everything that is known about the environmental factors that influence educational achievement and attainment.

Figure 12

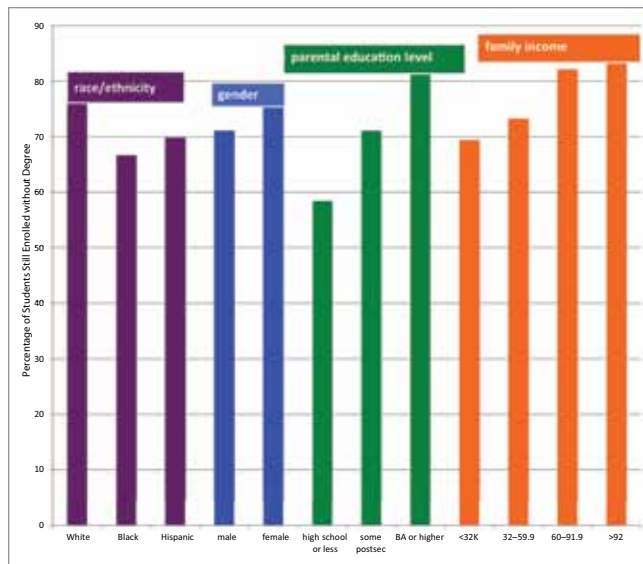
Poverty and TIMSS, Selected Countries, 2007



Source: Data from NCES (2007b).

Figure 13

Persistence in College: Students Who Enrolled in 2003–2004 Still Enrolled in 2006 Who Do Not Yet Have Degree



Source: Data from NCES (2007a).

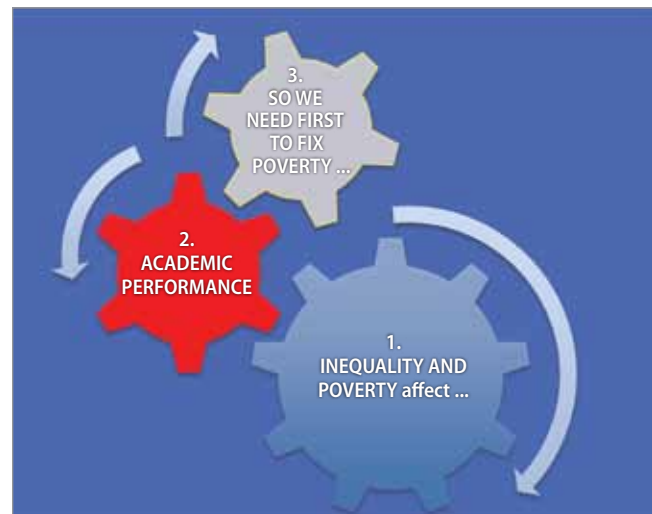
Taking account of the effects of poverty on academic achievement suggests a variation on Figure 9, shown as Figure 14: In simple terms, this diagram reverses the implied direction of causality by suggesting that economic conditions affect academic performance. But one needs to be careful not to allow the needed focus on poverty and socioeconomic status to be misconstrued as excusing, rather than explaining, achievement gaps between children of different racial/ethnic and economic groups. In other words, the reform movement has at times become trapped in an argument about the efficacy of holding schools accountable for improvement of student learning, given the measurable effects of external conditions (rampant inequality and poverty) on achievement.

This type of “either-or” logic — *either* we fix poverty first, *or* efforts to raise achievement of the poorest-performing students will essentially be futile —

confounds policy thinking both with respect to internal data (e.g., persistent gaps between minority and White students on the National Assessment of Educational Progress [NAEP] as well as state and other assessment programs) and with respect to unraveling the meaning of relative performance of the United States compared to other countries. In both cases, evidence of the effects of poverty, and especially in a period characterized by the deplorable rise in economic inequality (e.g., Saez, 2010), are key to understanding and shaping the prospects for education reform.

Figure 14

Education and the Economy: An Alternate View



The fact that poverty and inequality constrain the pace of growth in student achievement should not become an excuse for abandoning or even postponing reforms and insisting on holding schools to high standards of management and teaching. With respect to the uses of ILSA in reaching judgments about the quality of schools and schooling, two overlapping caveats are in order that relate to poverty and other so-called “background” variables: First, it is imprudent to adopt the stance that socioeconomic and demographic diversity are sufficient to

explain performance differences, especially in the light of evidence of increasing population heterogeneity and inclusion in countries that are, somehow, sustaining high aggregate levels of academic achievement. Second, diversity in U.S. school enrollment, a reflection of the nation's demographic mosaic, is often cited as a factor to explain low mean performance compared to more homogenous societies. But this argument falls short if it ignores considerable within-country variance that persists even in relatively high-achieving countries (Koretz, McCaffrey, & Sullivan, 2001).

Within the United States, we have many examples of substantial gains in academic performance even in communities racked with poverty and social disarray: Massachusetts, with about 12% of children in poverty

(lower than the U.S. national level of 22% but four times higher than in high-scoring Finland) is now one of our top-performing states (Feuer, 2011). Looking at a single slice of time camouflages significant cumulative effects: Today roughly five million English-language learners attend U.S. schools, twice the number from a decade ago and half the number projected for 2015. It is important to probe the meaning and significance of cross-national demographic differences and in particular to account for historical trends in the inclusion of children with widely diverse cultural, linguistic, and economic backgrounds. For these comparisons to have meaning, however, the socioeconomic measures that are typically used — such as eligibility for free or reduced-price lunch — may be inadequate (Hauser, 2009).

INTERNATIONAL LARGE-SCALE ASSESSMENTS AND CONSEQUENCES

Although it has become fashionable to demonize the excessive use of testing and assessment as a peculiarly late 20th- and early 21st-century phenomenon, the history is longer and more complicated. Standardized measures of teaching and learning did not suddenly make their appearance in the 2002 version of the Elementary and Secondary Education Act (No Child Left Behind), and from the inception of uniform written examinations in the mid-19th century (see, e.g., Office of Technology Assessment, 1992), there have always been tensions and controversies (Cronbach, 1975; Linn, 2001; Tyack, 1974). The well-known predicaments in assessment policy (e.g., Feuer, 2008; Hout & Elliott, 2011; Koretz, 2008), for which the American experience is in many ways unique for its historical reliance on tests and for its enormous contributions to the science of measurement, provide cautionary lessons for the design and uses of ILSA. I will focus here on just two.

First, there is the question of content. The fact that ILSA data often derive from significantly different concepts of what should be taught, to whom, and when is often overlooked in the public discourse that focuses on a country's ranking on international tables. How many policymakers have taken the time to understand the fundamental differences between the most prominent comparative frameworks, PISA and TIMSS? At the most rudimentary level, these tests differ in their emphasis on curriculum-specific versus more generic knowledge, and they therefore reflect different norms about what should and can be taught (e.g., Schneider, 2009).

Given the differences between various assessment programs, there is a natural tendency to look for linkages, or equivalencies, when results suggest discrepancies that are not easily understood. This problem is all the more significant when international results are tied to domestic trends at the state, local, and national

levels. After 30 years of seeing NAEP results reported in varying innovative ways, including the reporting of state-to-state comparisons and the positioning of results by achievement levels, the appearance of discrepancies between U.S. and foreign students on TIMSS and PISA naturally evokes questions about what it all means. Advances in the general theory of test linkage and equivalency (Dorans, Pommerich, & Holland, 2007), the prospects for linking among diverse state tests (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999), and in efforts to explain convergence and divergence between NAEP and various international assessments (Phillips, 2009) are central to the ongoing improvement of ILSA technology and interpretation.

Nevertheless, the normative question — *What should we be teaching?* — is not answerable through even the most advanced methods of linkage, equating, and validation. It is fundamentally a values question, not a statistical puzzle to be solved. Surrounding the question of content and linkage is the bigger problem of sequence: Should assessments be used to judge performance — as if the included constructs are the agreed-upon and fixed goals of the education system — or rather as a tool for policymakers and the general public to use as they probe the values and purposes that their societies place on education? Experience with NAEP may be instructive: The first step (at least in the original conceptualization of NAEP) is a complex process of public participation in the development of frameworks and items, which in its idealized model informs the drafting and refinement of assessment items. The items are then field-tested and validated and only then are included in the actual assessments. Results are intended both to measure progress on specific domains of skill and knowledge and to inform a broader and ongoing discussion of norms, values, and goals of education.

Although in recent years NAEP has taken on more of an accountability role than in its original design — a trend not without its own complications (Linn, 1998) — the basic idea that the program can promote dialogue, rather than issue summative comparative judgments of quality of teaching or schooling in various locales, remains one of its distinguishing characteristics. By analogy, then, rather than view the results of ILSA programs as *prima facie* evidence of comparative success or failure and by extension as the clinching argument for reforms that imitate characteristics of school systems where students seem to perform better, it might make more sense to explore how different types of assessments reflect values and expectations of schooling and to use the results as catalyst for public conversation and debate (E. Olshstein, personal communication, May 15, 2009).

The extent to which tests primarily designed to *assess* student performance on limited but important

domains of skill and knowledge are used to *drive* decisions about content, curriculum, and the governance of schools is a flash point in the debate over accountability and reform. Because neither accountability nor testing is a new phenomenon, it is not a new debate, even if the rhetoric for and against has become more heated (e.g., Ravitch, 2010). A core question is whether limited measures of academic achievement should be used to define a society's educational values, as emphasized eloquently in a recent report of the Israel Academy of Sciences and Humanities: "The purpose of education is ... also meant to shape character ... instill ... love of learning, the ability for independent study, self-confidence, risk-taking, imagination, creativity, leadership, respect and consideration for others, openness, inner contentment, familiarity with culture, skepticism, self-discipline ..." (Justman & Bukobza, 2010, p. 15).

NEXT STEPS: RESEARCH FOR BETTER COMPARATIVE ASSESSMENT

Visitors to the banks of the Rio Grande, or to the military demarcation line between North and South Korea, or to the fence separating Israel from neighborhoods of the West Bank, might quibble with the assertion that “we are living in a world without borders” (Feuer, 2010b, p. 1). Nonetheless, that opening line in a recent blue-ribbon report (National Governors Association, Council of Chief State School Officers, & Achieve, Inc. [NGA, CCSSO, & Achieve], 2008) captures a mood that makes international comparisons of student achievement especially prominent in debates about education and school reform. As the report noted,

“To meet the realities of the 21st-century global economy and maintain America’s competitive edge into the future, we need students who are prepared to compete not only with their

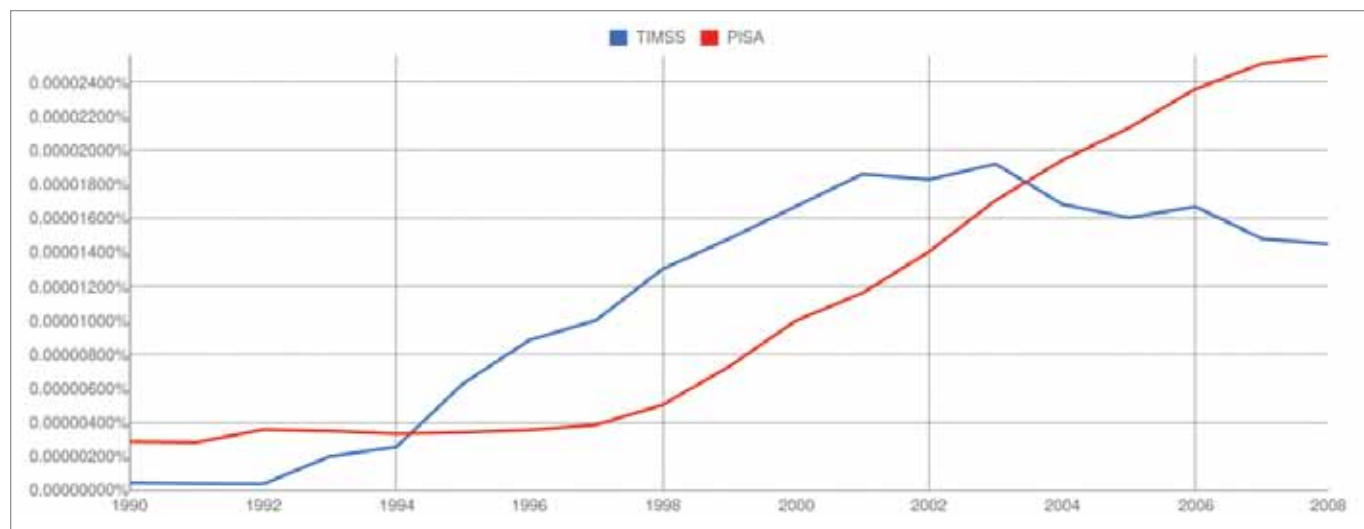
American peers, but with students from all across the globe for the jobs of tomorrow.” (NGA, CCSSO, & Achieve, 2000, p. 1)

With this zeitgeist, it is clear why data that produces rankings of countries on measures of educational performance have become so popular. Figures 15–17 capture the trend using the wonders of Internet search technology: The overall pattern clearly shows rapidly increasing interest in both major programs (TIMSS and PISA), with some subtle differences between countries. For these systems to continue to provide valid and useful information to policymakers, planners, and educators worldwide, a number of issues warrant ongoing (and improved) research. Here I focus on three suggested areas for the emerging research agenda.

Figure 15

Google *n*-gram of Growing Popularity of International Large-Scale Assessment:

TIMSS and PISA, 1990–2008, based on scan of books written predominantly in English, published anywhere. The y-axis shows the percentage of unigrams (mentions of the single word) that are TIMSS or PISA in the Google sample of books written in English. Numbers on the y-axis are in reference to roughly 5.2 million books scanned, or about 4% of the total books published. For more detail, see <http://books.google.com/ngrams/info>.

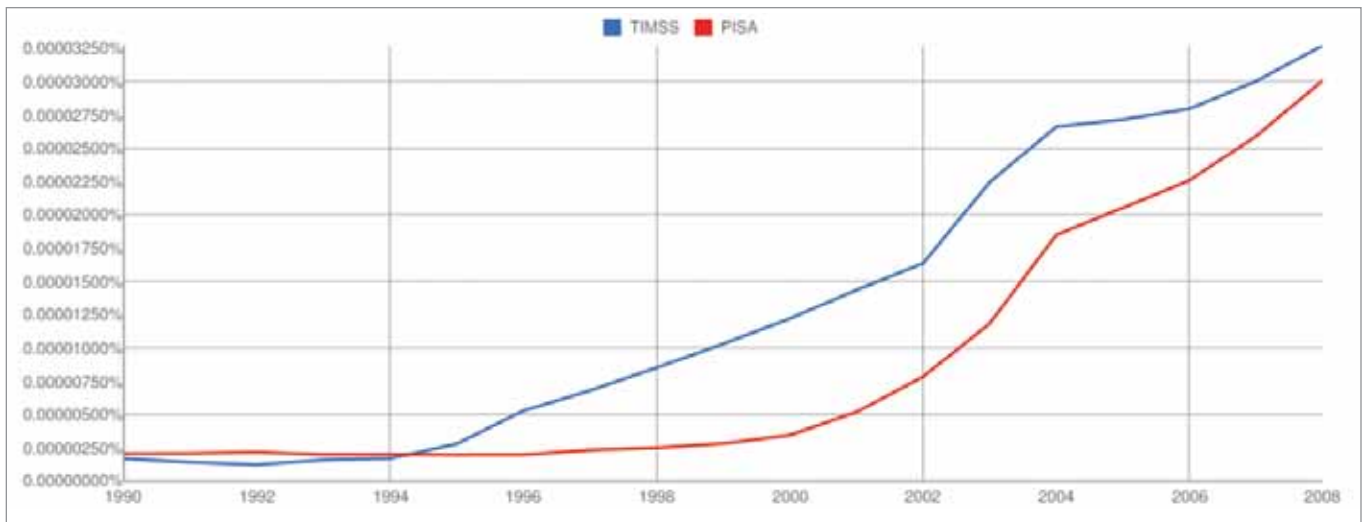


Source: Michel et al. (2010).

Figure 16

Google *n*-gram of Growing Popularity of International Large-Scale Assessment:

TIMSS and PISA, 1990–2008, based on scans of books written predominately in American English, published in the United States.

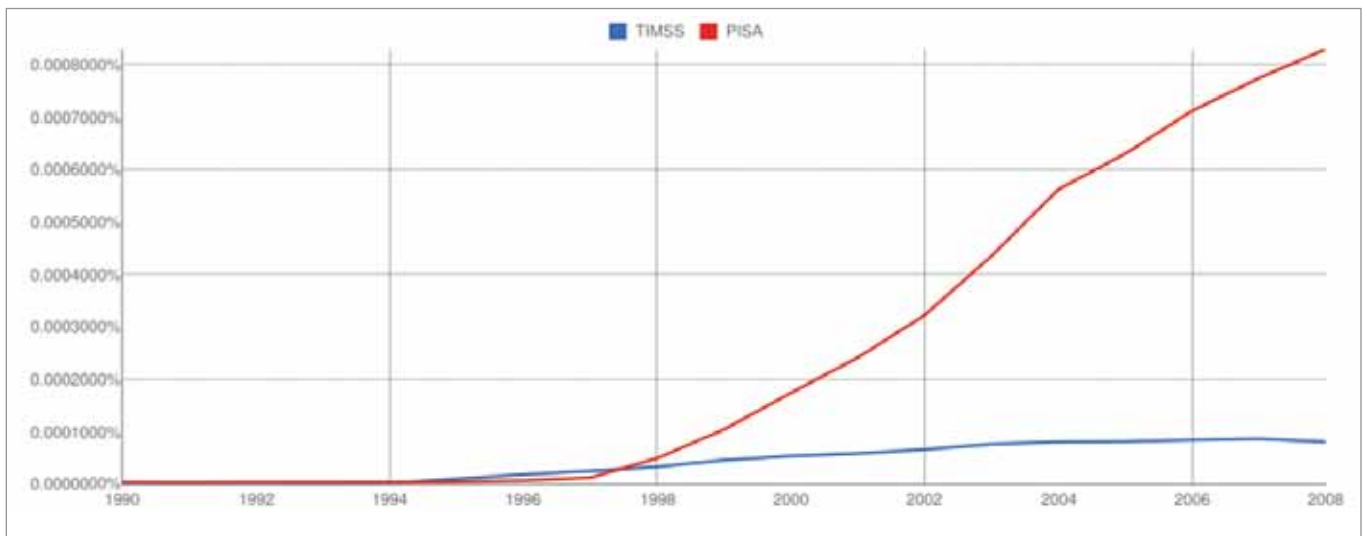


Source: Michel et al. (2010).

Figure 17

Google *n*-gram of Growing Popularity of International Large-Scale Assessment:

TIMSS and PISA, 1990–2008, based on scans of books written predominately in German.



Source: Michel et al. (2010).

LOGIC IN POLICY RHETORIC

First, there are basic puzzles of logic that would be worth unraveling. In the United States in particular, advocates of reform who worry about our current and future standing in the changing global economy make good use of ILSA results that show American students doing poorly compared to their peers elsewhere. Yet, many of these reformers, despite their willingness to rely on standardized measures from TIMSS or PISA, are among the first and loudest to advocate for radical overhaul in the design and uses of tests (e.g., Tucker, 2012). It seems that the indictment of our poor educational performance rests on results from the types of tests that the advocates of reform tend to distrust the most. In other words, claims that the United States is performing poorly compared to other countries are defended by reference to scores on international testing programs ... but at the same time, it is argued that one of the main reasons we perform poorly is because we rely too much on essentially the same basic types of tests.

At the risk of belaboring what may be an obvious point, the question is this: How is it that the same basic technology of assessment can be so trusted as the evidence source for inferences about our relative educational standing, and so distrusted as a tool for holding our schools and teachers accountable? Granted, TIMSS and PISA are credited with important innovations in item content and format, but neither of them comes close to the idealized models considered by many reform advocates and measurement theorists to be the path to educational improvement. ILSA results often are the prelude to a lamentation about the overwrought reliance on certain kinds of testing for accountability, followed by a plea to import allegedly superior assessment systems from Britain, Australia, New Zealand, and elsewhere. Embedded

in these calls, which are sometimes remarkably explicit in their celebration of European and Asian assessment systems and in their disdain for the American preference for so-called curriculum-free tests (e.g., Schneider, 2009), is the message that the assessment system itself is responsible for our poor educational performance.

Leaving aside for a moment this riddle — why ILSA is respected as the basis for assertions about the quality of American education while similarly-designed assessments are viewed as a major source of our problems — the bigger question is whether nations are willing and able to borrow from each other's successes (and avoid each other's failures). Here, too, there is a logical conundrum, perhaps more acute in the United States but observable elsewhere, too. If it is true, or at least plausible, that the United States is doing poorly compared to other countries, then one would think the logical remedy would be to import from those countries (and adapt to the U.S. scene) the policies and practices that seem most likely to contribute to these other countries' superiority.

Instead, there is evidence of a peculiarly inverted strategy: The current approach to reform in the United States rests on assumptions about the benefits of high-stakes testing in judging — and influencing — teacher quality and student learning. But this is clearly not the approach that can be credited with the high performance observed in countries such as Finland, Korea, and Canada. Put differently, the question is whether countries with high (average) performance on international assessments employ educational practices that are in vogue in the United States — and the answer seems to be no (Engel, Williams, & Feuer, 2011). A robust research program on the uses and consequences of ILSA should incorporate data and analyses that might narrow the chasm between assessment results and policy choices.

THE USES AND MISUSES OF TESTS

A second and related research priority pertains to the general problem of using assessments for purposes beyond those for which they are designed and validated. This is, of course, one of the reasons that test use in the United States is frequently in crisis. Though it is quite possible that new forms of assessment, including those that rely on more authentic representations of performance than are possible from conventional multiple-choice exams, will spur improvements in teaching and learning, it is still unclear whether any tool of measurement should be relied upon for double (or triple) duty: to assess learning, to reshape it, and to hold teachers (and schools) accountable for their performance. The future value of ILSA will depend, to a great degree, on the ability of designers and users to understand the principal purposes of their programs — and on the willingness of policymakers to curb their enthusiasm for tests as the solution to all educational problems.

BENEFITS AND COSTS

The third research priority follows directly: How can we narrow the gap between decision-makers' realities and the imprecision of test-based information? On the one hand, the impetus to use rigorous measurement technology in pursuit of valid and reliable data is beyond reproach. But because no assessment system is likely to perfectly or even adequately satisfy the many purposes for which it is used (Feuer, 2010a), a challenge for the policy community is figuring out the criteria by which to accept or reject findings from even the most advanced comparative designs. In other words, knowing that the inferences from ILSA — as from any large-scale testing program — are subject to varying degrees and types of error, the challenge is to decide if the anticipated benefits of administering the assessments and using the

results for policy outweigh the potential risks. And one of the more obvious risks relates to a theme of this paper — namely, the tendency to ignore or underplay the role of culture, context, history, political norms, and other factors that impinge on the effectiveness of importing educational policies across national boundaries based solely on results of ILSA.

The benefits side of this equation would include indicators such as the capacity of policymakers to learn from best (or at least reasonably good) practices elsewhere, the likelihood that lessons from abroad will be adaptable to a specific country's values and needs, and the incentives for improved effort and output that comparative results tend to create. The risks include narrowing of curricula and teaching in the chase for competitive advantage in rankings, errors in projecting economic and social outcomes that may distort resource allocation, and continued erosion of morale among the educators whose performance is judged as subpar. Placing these issues on the research agenda would open the world of assessment design and practice to methods of benefit-cost analysis that have proven useful and effective in other fields (Feuer, 2008).

Given the challenges of an increasingly interconnected world, the changing nature of work and its implications for the kinds of skills people will need to remain productive citizens, the continued pressures of accountability in democratic systems of education, and the desire to capitalize on advances that have been made in understanding the cognitive, social, and behavioral foundations of teaching and learning, international comparative assessment of student performance can be a vital source of knowledge to inform school reform and policy — but only if its limits are understood and its results are kept in proper perspective.

REFERENCES

- Adelman, C. (2009). *The spaces between numbers: Getting international data on higher education straight*. Washington, DC: The Institute for Higher Education Policy.
- Associated Press. (2010, December 7). In ranking, U.S. students trail global leaders. *USA Today*. Retrieved from http://www.usatoday.com/news/education/2010-12-07-us-students-international-ranking_N.htm
- Atkinson, R., & Blanpied, W. (2008). Research universities: Core of the US science and technology system. *Technology in Society*, 30, 30–48.
- Atkinson, R., & Pelfrey, P. (2010, March). *Science and the entrepreneurial university*. Paper presented at the Conference of Presidents of Academies and Scholarly Societies in honor of the 50th anniversary of the establishment of the Israel Academy of Sciences and Humanities, Jerusalem, Israel.
- Barro, R., & Lee, J.-W. (2010.) *Barro-Lee education attainment data set* [Data file]. <http://www.barrolee.com/>
- Baumol, W., Batey Blackman, S. A., & Wolff, E. N. (1989). *Productivity and American leadership: The long view*. Cambridge, MA: MIT Press.
- Baumol, W. J., Nelson, R. R., & Wolff, E. N. (1994). *Convergence of productivity: Cross-national studies and historical evidence*. New York, NY: Oxford University Press.
- Becker, G. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. New York, NY: Columbia University Press.
- Bulmahn, E. (2002, May). PISA: The consequences for Germany. *OECD Observer*. Retrieved from http://www.oecdobserver.org/news/fullstory.php/aid/702/PISA:_The_consequences_for_Germany.html.
- Bureau of Labor Statistics. (2010). *Economic news release: Table 1, Output per hour, output and hours*. Retrieved from <http://bls.gov/news.release/prod4.t01.htm>
- Cremin, L. (1990). *Popular education and its discontents*. New York, NY: Norton.
- Cronbach, L. (1975). Five decades of public controversy over mental testing. *American Psychologist*, 30(1), 1–14.
- Dorans, N., Pommerich, N., & Holland, P. (Eds.). (2007). *Linking and aligning scores and scales*. New York, NY: Springer.
- Engel, L., Williams, J., & Feuer, M. (2011). *The global context of practice and preaching: Do high-scoring countries practice what US discourse preaches?* Working Paper 2.3, Graduate School of Education and Human Development, The George Washington University. Presented at the annual meeting of the World Education Research Association, Taiwan. Retrieved from http://gsehd.gwu.edu/documents/gsehd/research/Working%20Paper%20Series/WPS2.3_Engel%20Williams%20Feuer_web.pdf
- Feuer, M., (2006). *Moderating the debate: Rationality and the promise of American education*. Cambridge, MA: Harvard Education Press.
- Feuer, M. (2008). Future directions for educational accountability: Notes for a political economy of measurement. In L. Shepard & K. Ryan (Eds.), *The future of test-based educational accountability*. New York, NY: Routledge.
- Feuer, M., (2010a). Externalities of testing: Lessons from the blizzard of 2010. *Measurement: Interdisciplinary Research and Perspectives*, 8, 59–69.
- Feuer, M. (2010b, April). *International benchmarking and the rhetoric of reform: Caveats and cautions*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.
- Feuer, M. (2011, August). The real S and T crisis. *Huffington Post*. Retrieved from http://www.huffingtonpost.com/dr-michael-j-feuer/the-real-s-and-t-crisis_b_922606.html
- Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Linkage and equivalence of educational tests*. Washington, DC: National Academy Press.
- Gardner, W. (2011, February 7). The golden age of education never was. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/walt_gardners_reality_check/2011/02/the_golden_age_of_education_never_was.html
- Gardner, E., & Ivancevich, J. (1994). Productivity in the US and Japan: A reexamination. *Interfaces*, 24(6), 66–73.
- Goals 2000: Educate America Act of 1994, 20 USC 5811 *et seq.* (West, 1994).
- Goldin, C., & Katz, L. (2008). *The race between education and technology*. Cambridge, MA: Belknap Press.

- Gordon, P. (2011, December 29). What hope remains? *The New Republic*. Retrieved from <http://www.tnr.com/article/books-and-arts/magazine/98567/jurgen-habermas-religion-philosophy>
- Hanushek, E., Peterson, P., & Woessmann, L. (2010, November 9). *U.S. math performance in global perspective: How well does each state do at producing high-achieving students?* Retrieved from <http://media.hoover.org/documents/TT-Report-global-perspective.pdf>
- Hanushek, E., & Woessmann, L. (2011). The economics of international differences in educational achievement. In E. Hanushek, S. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (Vol 3., pp. 89–200). Amsterdam, Netherlands: North Holland.
- Hauptman, A. (2011, February). *Increasing higher education attainment in the United States: Challenges and opportunities*. Paper presented at Degrees of difficulty: Can American higher education regain its edge? Washington, DC. Retrieved from <http://www.aei.org/events/2011/02/15/degrees-of-difficulty-can-american-higher-education-regain-its-edge-event/>
- Hauser, R. (2009). *On quality and equity in the performance of students and schools*. Unpublished manuscript.
- Hout, M., & Elliott, S. W. (Eds.). (2011) *Incentives and test-based accountability in education*. Washington, DC: National Academies Press.
- Inman, R. P., & Rubinfeld, D. L. (1997, Autumn). Rethinking federalism. *The Journal of Economic Perspectives*, 11(4), 43–64.
- International Association for the Evaluation of Educational Achievement. (2011). *Second international mathematics study*. Retrieved from <http://www.iea.nl/sims.html>
- Justman, M., & Bukobza, G. (2010). *Guidelines for revising the system of education indicators in Israel*. Retrieved from the Israel Academy of Sciences and Humanities, Initiative for Applied Educational Research in Israel website: <http://education.academy.ac.il/english/PublicationsList.aspx?AreaID=11&FromHomepage=false>
- Keillor, G. (1971, October 2). U.S. still on top, says rest of world. *The New Yorker*, 35.
- Kilpatrick, J. (2011, January 11). *Review of U.S. math performance in global perspective: How well does each state do at producing high-achieving students?* Retrieved from the National Education Policy Center website: <http://nepc.colorado.edu/thinktank/review-us-math>
- Kirsch, A. (2011). *Why trilling matters*. New Haven, CT: Yale University Press.
- Klein, J. (2011, June). The failure of American schools. *The Atlantic*. Retrieved from <http://www.theatlantic.com/magazine/archive/2011/06/the-failure-of-american-schools/8497/>
- Koretz, D. (1987). *Educational achievement: Explanations and implications of recent trends*. Washington, DC: Congressional Budget Office.
- Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.
- Koretz, D., McCaffrey, D., & Sullivan, T. (2001, September 9). Predicting variations in math performance in four countries using TIMSS. *Education Policy Analysis Archives*. Retrieved from <http://epaa.asu.edu/ojs/article/view/363>
- Lach, S., Shiff, G., & Trajtenberg, M. (2008, August). *Together but apart: ICT and productivity growth in Israel* (Working Paper Series STE-WP-2008). Haifa, Israel: Samuel Neaman Institute – STE program.
- Levy, F., & Murnane, R. (2005). *The new division of labor: How computers are creating the next job market*. Princeton, NJ: Princeton University Press.
- Linn, R. L. (1998). Validating inferences from National Assessment of Educational Progress achievement-level reporting. *Applied Measurement in Education*, 11, 23–47.
- Linn, R. L. (2001). A century of standardized testing: Controversies and pendulum swings. *Educational Assessment*, 7(1), 29–38.
- Loveless, T. (2011, February). *The 2010 Brown Center report on American education*. Washington, DC: The Brookings Institution. Retrieved from http://www.brookings.edu/reports/2011/0207_education_loveless.aspx
- Lumina Foundation. (2009). *Lumina Foundation's strategic plan, goal 2025*. Retrieved from http://www.luminafoundation.org/wp-content/uploads/2011/02/Lumina_Strategic_Plan.pdf

- Manna, P. (2006). *School's in: Federalism and the national education agenda*. Washington, DC: Georgetown University Press.
- McDonnell, L. (2005). No Child Left Behind and the federal role in education: Evolution or revolution? *Peabody Journal of Education*, 80(2), 19–38.
- Medrich, E., & Griffith, J. (1992, February). *International mathematics and science assessment: What have we learned?* (NCES 92-011). Washington, DC: U.S. Department of Education Office of Educational Research and Improvement.
- Michel, J.-B. , Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Brockman, W., ... Aiden, E. L. (2010, December 16). Quantitative analysis of culture using millions of digitized books. *Science Express*. Retrieved from <http://www.sciencemag.org/content/early/2010/12/15/science.1199644>
- National Academy of Sciences. (2007). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. Washington, DC: National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=11463
- National Center for Education Statistics. (1999). *Trends in International Mathematics and Science Study (TIMMS): Mathematics and science achievement of eighth-graders in 1999* [Data file]. Retrieved from http://nces.ed.gov/timss/results99_1.asp
- National Center for Education Statistics. (2003). *Trends in International Mathematics and Science Study (TIMMS): Mathematics and science achievement of eighth-graders in 2003* [Data file]. Retrieved from http://nces.ed.gov/timss/results03_eighth03.asp
- National Center for Education Statistics. (2007a). *Persistence and attainment of 2003–04 beginning postsecondary students: After three years, first look* (NCES 2007-169). Retrieved from <http://nces.ed.gov/pubs2007/2007169.pdf>
- National Center for Education Statistics. (2007b). *Trends in International Mathematics and Science Study (TIMMS): Mathematics achievement of fourth- and eighth-graders in 2007* [Data file]. Retrieved from http://nces.ed.gov/timss/results07_math07.asp
- National Center for Education Statistics. (2010a). *Digest of education statistics, 2010: Table 235* [Data file]. Retrieved from http://nces.ed.gov/programs/digest/d10/tables/dt10_235.asp
- National Center for Education Statistics. (2010b). *The condition of education*. Retrieved from http://nces.ed.gov/programs/coe/indicator_hep.asp
- National Center for Education Statistics. (2011a). *Digest of education statistics, 2010, Table 8* [Data file]. Retrieved from http://nces.ed.gov/programs/digest/d10/tables/dt10_008.asp
- National Center for Education Statistics. (2011b). *The nation's report card: Mathematics 2011*. Retrieved from <http://nces.ed.gov/nationsreportcard/pdf/main2011/2012458.pdf>
- National Center for Education Statistics. (2012). *Mathematics 2011: Trial urban district assessment results at grades 4 and 8*. Retrieved from <http://nces.ed.gov/nationsreportcard/pdf/dst2011/2012452.pdf>
- National Commission on Excellence in Education. (1983, April). *A nation at risk: The imperative for educational reform*. Washington, DC: Government Printing Office.
- National Governors Association, Council of Chief State School Officers, & Achieve, Inc. (2008). *Benchmarking for success: Ensuring U.S. students receive a world-class education*. Washington, DC: National Governors Association. Retrieved from <http://www.achieve.org/files/BenchmarkingforSuccess.pdf>
- Nixon, R. (1971, August 15). *Address to the nation outlining a new economic policy: The challenge of peace*. Retrieved from The American Presidency Project website: <http://www.presidency.ucsb.edu/ws/?pid=3115>.
- Organisation for Economic Co-operation and Development. (2008). *Education at a glance 2008, Table A1.3a* [Data file]. Retrieved from http://www.oecd.org/document/9/0,3746,en_2649_39263238_41266761_1_1_1_1,00.html
- Organisation for Economic Co-operation and Development. (2009). *Database – PISA 2009* [Data file]. Retrieved from http://pisa2009.acer.edu.au/interactive_results.php

- Organisation for Economic Co-operation and Development. (n.d.). *StatExtracts* [Data file]. Retrieved from <http://stats.oecd.org/Index.aspx?DatasetCode=STAN08BIS&lang=en>
- Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions*. Washington, DC: Government Printing Office.
- Phillips, G. (2009). *The second derivative: International benchmarks in mathematics for U.S. states and school districts*. Washington, DC: American Institutes for Research.
- Puma, M. J., Jones, C., Rock, D., & Fernandez, R. (1993). *Prospects: The congressionally mandated study of educational growth and opportunity* (interim report). Cambridge, MA: Abt Associates.
- Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York, NY: Basic Books.
- Rice, J. M. (1893). *The public-school system of the United States*. New York, NY: Century.
- Ripley, A. (2010, December). Your child left behind. *The Atlantic Magazine*. Retrieved from <http://www.theatlantic.com/magazine/archive/2010/12/your-child-left-behind/8310/>
- Saez, E. (2010, July 17). *Striking it richer: The evolution of top incomes in the United States (Updated with 2008 estimates)*. Retrieved from the University of California, Department of Economics website: <http://elsa.berkeley.edu/~saez/saez-UStopincomes-2008.pdf>
- Schneider, M. (2009, Fall). The international PISA test: A risky investment for states. *Education Next*, 9(4). Retrieved from <http://educationnext.org/the-international-pisa-test/>
- Schuck, P., & Wilson, J. Q. (Eds.). (2008). *Understanding America: The anatomy of an exceptional nation*. New York, NY: PublicAffairs.
- Sewell, W. H., Hauser, R. M., & Featherman, D. L. (Eds.). (1976). *Schooling and achievement in American society*. New York, NY: Academic Press.
- Snyder, T. (Ed.). (1993). *120 years of American education: A statistical portrait*. Washington, DC: U.S. Dept of Education.
- Stedman, L. C., & Smith, M. (1983, Fall). Recent reform proposals for American education. *Contemporary Education Review*, 2, 85–104.
- Swanson, B. (1991). An overview of the six national education goals. *ERIC Digest*. Retrieved from <http://www.ericdigests.org/pre-9220/six.htm>
- Takayama, K. (2007). A nation at risk crosses the Pacific: Transnational borrowing of the U.S. crisis discourse in the debate on education reform in Japan. *Comparative Education Review*, 51, 423–446.
- Tucker, M. (2012, January). On curriculum, assessment and (especially) writing in the modern age [Blog]. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/top_performers/2012/01/on_curriculum_assessment_and_especially_writing_in_the_modern_age.html
- Tyack, D. (1974). *The one best system: A history of American urban education*. Cambridge, MA: Harvard University Press.
- U.S. Bureau of Labor Statistics. (2010). *Table 1. Output per hour, output and hours* [Economic news release]. Retrieved from <http://www.bls.gov/news.release/prod4.t01.htm>
- Vinovskis, M. (1999). *The road to Charlottesville: The 1989 education summit*. Washington, DC: The National Education Goals Panel.
- Williamson, J. (1991, March). Productivity and American leadership: A review article. *Journal of Economic Literature*, 29, 51–68.

About ETS

At ETS, we advance quality and equity in education for people worldwide by creating assessments based on rigorous research. ETS serves individuals, educational institutions and government agencies by providing customized solutions for teacher certification, English language learning, and elementary, secondary and post-secondary education, as well as conducting education research, analysis and policy studies. Founded as a nonprofit in 1947, ETS develops, administers and scores more than 50 million tests annually — including the *TOEFL*® and *TOEIC*® tests, the *GRE*® tests and *The Praxis Series*™ assessments — in more than 180 countries, at over 9,000 locations worldwide.

