Does Anybody Really Want to Know the Consequences of High Stakes Testing?

Robert M. Hauser

The University of Wisconsin-Madison

August 2004

My graduate study in sociology did not begin with a moral commitment. I am no Marx, Weber, Durkheim – or Burawoy (2004). I wanted economic independence from my parents. I also wanted to do scientific research and to improve the lot of the urban poor, but I doubt that, even then, I would have called that desire a "moral commitment." Before entering graduate school, I had several years of research experience, first at the Operations Research Office of Johns Hopkins University, and then at a city planning firm in Chicago. A few years earlier, that firm had been instrumental in Hyde Park-Kenwood's urban renewal, and I was introduced to urban sociology through studies of the decline of the neighborhood movie theater in Chicago, the growing market for downtown housing with a lake view, and the miserable failure of family relocation for urban renewal.

By home training and consequent personal disposition, I was more inclined toward scientific research in the public interest than to social criticism or advocacy. Two male role models dominated the family horizon. My father, Julius Hauser, was an expert on regulatory law within the federal Food and Drug Administration. In that role, he first-drafted the 1962 Kefauver-Harris Amendments to the Food, Drug, and Cosmetic Act and concurrently introduced the first legislation requiring the informed consent of participants in clinical trials. His brother was the sociologist, demographer, and social statistician, Philip M. Hauser, whose career wandered freely between federal service and a professorship at the University of Chicago. Phil

was instrumental in a major social invention of the late depression years, the operational definition of unemployment as the activity of looking for work within a specific time interval. This invention led to one of the major tools of economic policy, the monthly unemployment rate. He also introduced the idea of a continuous or "rolling" Census (Hauser 1942), which is now at last coming to fruition. Phil was always a highly visible figure, willing to offer pronouncements about major public issues and capable of giving well-timed public talks with no visible signs of preparation.

There were negative as well as positive lessons from these two models. I learned very early that I did not share my Uncle Phil's gift of gab and that I wanted to spend more time with my family than he did. My father went through almost his entire federal career out of the public eye because of three disabilities: the political activities of his youth, his Jewish origin, and a severe hearing disability. These led to repeated security investigations, threats of job loss, and limited career opportunities. I had considerable appetite for social policy, whetted by exposure to the muckrackers and to writings of the Chicago school. However, at our home, dinner-table conversations often focused on the desirability of keeping opinions to oneself, on identifying the political agenda within seemingly innocent texts, or on the irrational element in strong identification with any political or religious cause.

Thus, as a young adult, and despite the political and social turmoil of the 1960s, I found myself free and possibly incapable of active social commitment. I rationalized this lack of conviction with the idea that, later on, with maturity, I would have sufficient knowledge, skills, and credentials to help change the world. Thus disarmed, I wandered toward research on social stratification.

In the mid-1980s, I began an association with the National Research Council (NRC). The NRC is a unique think-tank, the research unit of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The primary mode of operation of the NRC is to assemble expert panels. The panels are charged to respond to queries about important national issues that are posed by federal agencies or by congressional legislation. Unlike other think-tanks, the NRC does not compete for grants or contracts. It accepts only sole-source contracts, and it is reimbursed only for actual expenses. Expert panels are chosen to provide balance in studies of controversial subjects, and strict conflict-of-interest rules are enforced. The work of panels is supported by professional staff, but panel members themselves serve without pay. Panel reports are reviewed externally, and every panel must respond in writing and at length to each review. Sponsors have no control over the work of panels and are not informed about the contents and recommendations in a report until it is ready for release. When panels hold open meetings, they are subject to the rules of the Federal Advisory Committee Act (FACA), which require public notice and public access to materials used by the panel in open session. However, panels are allowed to meet privately as recommendations and final reports are fashioned in order to free panel members from commitment to their prior public stands. Authorship of panel reports is always collective, and they are properly cited as such, but often with named editors. Some 200 NRC reports are released to the public each year through the National Academies Press (http://www.nap.edu), both in hard copy and in PDF.

In NRC panel studies and their governing bodies, I found a venue for my interests in public policy and in the direction of the scientific enterprise that is a rewarding activity in itself, that permits me to work with talented scientists and professionals in diverse disciplines, that has enabled me to use my scientific knowledge and skills for the public good, and that occasionally

affects important matters in our society. Most of my NRC projects have involved national resources for social or economic measurement or social stratification. I have had a hand in *A Common Destiny: Blacks in American Society* (National Research Council 1989), *The Future of the Survey of Income and Program Participation* (National Research Council 1993), *Measuring Poverty: A New Approach* (National Research Council, Panel on Poverty and Family Assistance 1995), *Scientific Research in Education* (National Research Council 2003), *Protecting Participants and Facilitating Social and Behavioral Sciences Research* (National Research Council, Panel on Institutional Review Boards, Surveys, and Social Science Research 2003), *The 2000 Census: Counting Under Adversity* (National Research Council 2004), and *High Stakes: Testing for Tracking, Promotion, and Graduation* (National Research Council, Committee on Appropriate Test Use 1999).

Soon after I started working on NRC panels, I felt comfortable enough in the public arena to take on the issue of pay discrimination against professional and faculty women in the University of Wisconsin System. There, too, the theories, models, and methods of stratification research proved useful.

All of this probably leaves me rotating between the "professional" and "policy" cells in Michael Buroway's classification of "types of sociology" (2004:106). Perhaps I have even wandered into the "public" cell. I will leave that judgment to the reader, after I have described my experience with high stakes testing.

I want to start with a word or two about testing and tests. Contrary to one critic of my work with the National Research Council, I am not a test-hater (Phelps 2000; Phelps 2003:152-55). In my day job, one of my major preoccupations for the last 35 years has been to try to figure out what difference test scores make in people's lives. My wife, Tess, and I lead the Wisconsin

Longitudinal Study (WLS), which has followed Wisconsin's high school graduating class of 1957 for almost 50 years (Sewell, Hauser, Springer, and Hauser 2003). We have looked for long-term correlates of scores on tests that people took when they were adolescents. Of course, they're not the kinds of tests that are now in wide use. The main test taken by the WLS graduates is the Henmon-Nelson Test of Mental Ability, which was administered to almost all high school freshman and juniors in the State of Wisconsin in the 1950s by the Wisconsin State Testing Service (Henmon and Holt 1931; Henmon and Nelson 1946; Henmon and Nelson 1954). I don't think anybody uses the Henmon-Nelson Test anymore, except us. We are now, again, going back into records of the Wisconsin State Testing Service, and we are going to pick up a few Iowa test scores as well. We have a lot of information.

The Wisconsin State Testing Service was a cooperative program of the University of Wisconsin, other colleges in the state, and high schools throughout the state. As reported in the National Research Council's report on high stakes testing (1999:20):

> It is helpful to keep in mind that standardized tests have often been used historically to promote equal opportunity. In the early 1930s, the Wisconsin State Testing Service gave a standard test of academic ability to all graduating high school seniors and sent the names of high-scoring students to the state's colleges and universities, so they could identify academically promising recruits. In later years, the testing program was expanded to lower grades, to identify promising students who might need greater academic encouragement.

Thus, I thought I'd start by telling you what difference test scores make. We have been very successful, overall, in maintaining the participation of the WLS graduates across the

decades. We interviewed more than 85 percent of survivors in 1992-93, and we also expect high coverage in the current round of the survey, which is now in the field. For all of the things that we think are really important—what kinds of jobs or careers you have, your political participation, your health—test scores matter in only one way. Test scores affect how far you go in school, and everything else depends on how far you go in school. Period. That's what this is all about – and it took Herrnstein and Murray (1994) a lot of effort to miss this obvious point. To be sure, Wisconsin's high school seniors in the 1950s provide a narrow window on the world – and, among other omissions – there are no non-high school graduates and almost no minorities in the sample. However, serious comparisons of WLS findings with those in broader populations reveal few differences (Jencks, Crouse, and Mueser 1983). Survey response is the only thing that adolescent test scores (and high school grades) appear to affect directly, above and beyond the influence of educational attainment and other demographic characteristics; those with higher test scores are more likely to respond than those with lower test scores.

I'm a test-user, not a tester. So, how did I get into the testing business? I got a call in late in the summer of 1997 from Michael Feuer, who then directed the NRC's Board on Testing and Assessment. Feuer introduced me to the NRC's role in evaluating the development of the Voluntary National Tests (VNT) in 4th grade reading and 8th grade math (Wise, Hauser, Mitchell, and Feuer 1999), which had been proposed by the Clinton administration. He told me about the VNT, of which I was only vaguely aware. The idea was to create a test that could be administered in two 45-minute sessions, that would assess skills and knowledge relatively independent of curricular content, that could be administered year after year under secure conditions, whose items would be released after each test administration, that would yield reliable scores (and, possibly, subscores) at the individual, classroom, school, and system levels,

and that would yield interpretations of the scores that would be useful and intelligible to children, parents, teachers, administrators, and the general public. This would be no mean feat.

And Feuer had a problem with the evaluation. Congress had asked the National Academies to evaluate the development of the VNT, but he was unable to find anyone to run the study who actually knew anything about testing. All of the leading psychometricians had either come out publicly against the VNT, had come out publicly in favor of the VNT, or had been hired to help develop the VNT. As a member of the NAS, would I be willing to collaborate with a psychometrician, Lauress Wise, in the evaluation of the first year of test development? Feuer was very persuasive, I was both curious and compliant, and I agreed.

That first step soon led to others. The Republicans in congress came out foursquare against national testing as a violation of the great American tradition of local control of education. (Compare what is happening now under No Child Left Behind – or what some call "No Child Left Untested.) Moreover, the Black congressional caucus was also then leery of national testing, fearing that it would be used against the interest of minority children. These two groups formed an odd coalition against the VNT. Ultimately, passage of the omnibus budget reconciliation bill was possible only after a meeting between President Clinton and Representative Goodling, the Republican chair of the house education and labor committee. They agreed to limit test development and to mandate two more NRC studies. Items could be developed for the VNT only if students were not actually tested, but talk-aloud laboratory sessions were later permitted. This agreement violated all of the rules for test construction, for that procedure involves writing items, followed by extensive pretesting, statistical analysis, pilot administration, item selection, and, finally, the creation of alternate, equivalent test forms. We later referred to this as Zen test development, after the old line, "What is the sound of one hand

clapping?" One of the two additional studies, designed by the Republican leadership, was to assess the possibility of placing students' scores on existing achievement tests into a single, comparable metric. The other study, designed by the black caucus, was to

> "study and make written recommendations on appropriate methods, practices, and safeguards to ensure that—(1) existing and new tests that are used to assess student performance are not used in a discriminatory manner or inappropriately for student promotion, tracking or graduation; and (2) existing and new tests adequately assess student reading and mathematics comprehension in the form most likely to yield accurate information regarding student achievement of reading and mathematics skills" (Public Law 105-78, enacted November 13, 1997).

At that point, before the evaluation project had scarcely begun, Michael Feuer asked me to chair the panel that wrote *High Stakes* between February and July of 1998 (National Research Council, Committee on Appropriate Test Use 1999). I am truly grateful for the experience that I had, not only with that project, but with two annual evaluations of the Voluntary National Test (VNT) (National Research Council, 1999; Wise et al. 1999), and, especially for the privilege of working with many of the leading figures in the education and testing communities.

Since the National Academy of Sciences has something of a reputation for coming up with scientific and technical fixes, I want to say a word about the report on test equivalence (National Research Council, Committee on Equivalency and Linkage of Educational Tests 1999). The major recommendation of that report is neatly summarized in its title, *Uncommon Measures*. In other words, there are no short cut ways to create comparable measures of academic achievement. Not satisfied with this response, there soon came a call for evaluation of

yet another makeshift comparability scheme: Would it be possible to establish comparability by embedding questions from a national assessment into independent state assessments? To that question, as well, a new NRC panel said, "No" (National Research Council, Committee on Embedding Common Test Items in State and District Assessments 1999).

I want to put the report on high stakes testing for tracking, promotion, and graduation in context. This is considered to be one of the academy's most effective reports and one which has been fairly widely used. I've been involved in several academy panels and they've been equally effective. For example, my first NRC experience was with the committee on the status of black Americans, which wrote *A Common Destiny* (National Research Council 1989), and as you know race/ethnic differences are no longer an issue in this country. I was also involved in the report that proposed a new measure of poverty for the United States (National Research Council, Panel on Poverty and Family Assistance 1995), and as you know our poverty standard now accurately ranks the population with respect to its access to economic resources. I feel the same way about *High Stakes*.

I wish NRC reports had more impact. Traditionally, the Academy has not been especially effective at delivering its messages. Occasionally there are great successes, and I think that there recently has been a much stronger effort within the academy to try to capture public attention when major reports are issued. One very positive recent example has been the IOM's report on medical errors, which, I think, really has had a salutary effect (Kohn, Corrigan, Donaldson, Institute of Medicine (U.S.), and Committee on Quality of Health Care in America 2000). Another was a report that reconciled the conflict between reading instruction using phonics and whole-word recognition (National Research Council. Committee on the Prevention of Reading Difficulties in Young Children 1998; Burns, Griffin, Snow, and Committee on the

Prevention of Reading Difficulties in Young Children 1999), but the contribution of that work has been blocked by the Bush administration's doctrinaire commitment to phonics.

Sometimes it just takes a while for the messages to be heeded. I was on a panel evaluating the design of the Census Bureau's Survey of Income and Program Participation in which we made a number of strong recommendations for sample redesign (National Research Council 1993; Weinberg 2002). The Bureau was so resistant to the recommendations we made that the then chief of demographic programs came over and told the Committee on National Statistics that he wanted the report buried and that we were not to discuss it in public. Of course, we ignored that injunction. Around 2000 the Census Bureau adopted essentially the design that we had recommended, after four years of disastrous failure with a different design.

I have wondered whether a reissue of *High Stakes*, perhaps revised to eliminate all references to the voluntary national tests, would help get these issues back in the limelight. Sometimes issues are revisited, the policy climate changes, and things improve. The immediate effect is not the only one we seek. *Measuring Poverty* is a case in point (National Research Council, Panel on Poverty and Family Assistance 1995; Citro, Michael, and Hauser 1996). Although we are still stuck with the manifold inadequacies of Mollie Orshansky's measure as the official index of poverty in America, by the end of the Clinton administration, the new concept and measure was used in the Economic Report of the President, and experimental series based on the report are still in development.

I can think of no better way to talk about *High Stakes* than to review what it had to say. What are high-stakes tests? They are tests used for decisions that have major impact on the test taker, such as tracking, promotion, or high school graduation. The questions our committee began with were:

- What are appropriate, nondiscriminatory uses of these tests?
- How can the participation of students with disabilities and English language learners in these kinds of tests be maximized at the same time that we ensure that their test results are comparable to the results for all students?
- How can we make sure that test makers and test users will abide by norms of appropriate, nondiscriminatory test use?

In thinking about these questions we needed to consider the goals for this kind of testing. We want to set high standards. We want to raise student achievement. We want to ensure equal educational opportunity. We want to foster parental involvement, and we want to increase public support for the schools.

But testing can also have negative consequences for individuals, so policy makers should be sensitive to the balance between individual and collective benefits and costs. We had a framework for our analysis of testing and its consequences that was largely based on the efforts of a now deceased member of our panel, the brilliant Samuel Messick, who was a major contributor to the report, not only through his previous, highly influential work, but in making major substantive contributions to the volume. The framework had to do with measurement validity, that is to say, how well a test covers the knowledge and skills it was intended to cover. It also addressed the attribution of cause for performance, that is, whether performance really reflects knowledge and skills based on proper instruction, as one hopes it will, or whether it reflects poor instruction, or whether it reflects irrelevant factors such as language barriers or unrelated disabilities. Finally, we considered whether the consequences or potential consequences of decisions based on tests would be more beneficial educationally than other available treatments.

We also stated some principles of test use. First of all, tests have validity only in relation to specific purposes. Second, they're not perfect, but neither are the alternatives to tests. Third, no high-stakes educational decision about a test taker should be made solely or automatically on the basis of a single test score; other relevant information should be taken into account. Think about Chicago in the recent past and New York in the very near future (Steinhauer 2004; Nagaoka and Roderick 2004; Allensworth 2004). Fourth, neither test scores nor any other kind of information can justify educational decisions that are not beneficial for students. Last, tests should be used for high-stakes decisions only after students have been taught the knowledge and skills on which they will be tested.

In order to place those principles in context, we considered some of the dimensions of tests: they may legitimately lead or follow instruction; they may be useful as indicators, as individual diagnostics, or as determining factors in educational decisions. The information in tests may be useful at the individual level, that of the school, that of school system, or that of the state or nation. Tests may be given to samples or to entire populations. They vary in the levels of knowledge or skill that they tap, in the type of performance that they demand, and in the schemes for scoring test performance that are used.

The panel worried a lot about the imperfections of tests: that tests are built from samples of items, that test scores are themselves samples of students' knowledge and skills, and that test scores vary relative to true knowledge and skill. In other words, there's really a standard error of measurement in individual test scores. For example, there Rogosa (1999) found that the chances that two students with identical "real" achievement will score more than 10 percentile points apart on the same test may range from 42 to 57 percent in one widely used assessment.

I want to mention a few of our findings and recommendations. First of all, we agreed that accountability for educational outcomes should be shared among states, school districts, public officials, educators, parents, and students—not borne by students alone. If we think about the No Child Left Behind Act of 2001 (NCLB), it's clear that there is some sharing involved, although the way in which this sharing is done at the schoolhouse level is highly problematic. In fairness, I do not think one can say that responsibility is just devolving on the students. Whether the system works effectively across all the players is another question. We also found that, although tests should be used for high-stakes decisions about individual mastery only after students have been taught the knowledge and skills at which they will be tested, it is appropriate for tests to lead instruction when high stakes are not attached to individual student performance. Of course, one of the problems that we continue to face is that once a test is given, it tends to be used for a variety of purposes, no matter what the original intent, plan, or design was. Given the massive increase in testing required by NCLB, that is truly a scary prospect.

We noted that the consequences of high stakes testing are often "either/or"—that is, pass or fail, be promoted or not—but that doesn't have to be the case. Tests and other information can lead to early diagnosis and effective intervention when students have learning problems. Another important point for us was that some educational practices are typically bad for students. These include placement in typical low track classes and simple retention in grade, which I will say more about in a bit. Neither tests nor any other type of information should be used to make such decisions. Unfortunately, no one is paying attention to that.

We argued that all students are entitled to sufficient test preparation—familiarity with item format, appropriate test taking strategies, etc. On the other hand, we thought it was important that educators avoid narrowly teaching to the test. Finally, we said that high-stakes

testing programs should include a well designed evaluation component, and that the consequences of high stakes assessment should be measured for all students and major subgroups of students. There is at least one exemplary case of that, which is the recent use of tests to retain students in the city of Chicago, and the most important findings have just come out. There were no surprises (Allensworth 2004; Nagaoka and Roderick 2004).

We also tried to say something about strategies for promoting appropriate test use. We noted that the current mechanisms are inadequate; there are standards in the testing profession, and they are widely ignored. Consider, for example, the statement of the publishers of the Iowa Test of Basic Skills – used to retain Chicago students – that the test is not valid for that purpose. There are legal avenues to promote proper test use -- either administrative mechanisms or litigation – and they are not very effective. One egregious example is the litigation by the Hispanic community in Texas, where a federal judge found evidence of disparate impact of the state's high school exit exam, but ruled for the state on the simple ground that the intent of the test requirement was to improve educational outcomes. We considered a number of additional policy mechanisms, including deliberative forums, independent oversight bodies, labeling, and perhaps most important, the proposal to use federal regulation to enforce existing professional standards. There was a federal resource guide, which was released in December of 2000 (U.S. Department of Education,Office for Civil Rights 2000), but by April of the following year, it had been relegated to archival status on the web site of the Department of Education (http://www.ed.gov/offices/OCR/archives/testing/index1.html).

Our overarching conclusions were as follows: When used appropriately, high-stakes tests can help promote student learning and equal opportunity in the classroom by defining standards of student achievement and by helping school officials to identify areas in which students need

additional or different instruction. When used inappropriately, they can undermine the quality of education and reduce opportunities for some students, especially if results are misinterpreted or misused, or students are relegated to a low-quality educational experience as a result of their scores.

We also laid out the ingredients in an appropriate high stakes testing plan, though I'm not sure that anybody has tried such a plan, and we can't be sure it would work. These include curricular and performance standards, alignment between curriculum and standards, teachers trained to teach to the standards, tests built to assess performance relative to the standards, a phase-in period of several years to lead instruction, early diagnosis and remediation for students with difficulties, provisions for repeating tests and using other information to evaluate students, and evaluation of short- and long-term consequences.

My participation in the study piqued my interest in promotion *per se*. That was because it was related both to the demography of schooling and testing. Promotion was something that I could study for the panel in ways that nobody else had done up to that time, and I was inspired by President Clinton's State of the Union address in 1998 in which – despite his desperate political situation at the time – he got quite a large round of applause when he mentioned that we should "end social promotion." Promotion and retention practices have become publicly visible since then, and they are likely to become even more so in the near future.

Ironically, the basic issues have been clear for almost a century (Ayres 1909; Hauser 2004). The problem is not social promotion; it is low academic achievement. We have to keep that in mind. We already retain students in very large numbers, and the bottom line – the bulk of the evidence shows that retention leads to lower achievement and higher dropout (Hauser 2001; Hauser 2004). How much retention is there? There are still no good national statistics. There

are a few statistics available from states. Roughly speaking, these are the major points: about 7 – 11% of kids are retained in kindergarten and first grade; about 15% more are retained from ages 6 – 8 through 15 – 17. There is a great deal of retention in high school, but it's a little bit hard to measure because of the fact that it's really credits that matter at that stage, rather than years. The population group that is most behind at entry to the first grade is white males. Boys fall further behind girls as time passes; minorities fall further behind majority students; and there has been growing age/grade retardation since 1970, largely because age at school entry has increased. Promotion and retention practices vary widely across states. For example, in Texas, 44% are retained at least once from K to 12; 55% if they're black. In Louisiana, 66%; in Wisconsin, a relatively low retention state, 30%. But the statistics are really not very good. Figure 1 is my effort to squeeze out some information – updated with data through 2002, from the October Current Population Survey. It shows age/grade relationships for cohorts. The time scale at the bottom is the age at which a group of children were 6 – 8 years old. You can see that being behind the modal grade for age increases as you read up the diagram. The red line is for 6 – 8 year olds, the yellow line is for 15 – 17 year olds. We count dropout as a form of age/grade retardation, which it is, but I've also shown the trend in dropout, which is actually downward, so that's not producing the increase that we observed through about 1990 in age/grade retardation among 15 – 17 year olds. The problem here is that you'll notice the red line is running upward through much of this period, and then it kind of flattens out, and may even have declined a little bit. That has to do with increases in age in school in entry to the first grade. Some of that reflects changes in laws governing age at school entry, and some of it is kindergarten retention. It's very difficult to untangle those two factors – law and retention practice so I have redisplayed the cohort data relative to the share of the population that was not age/grade retarded at school

entry in Figure 2. This is, again, for the whole population, age/grade retardation in percentages conditional on timely entry to the first grade. We know that – aside from artifacts of testing practice – test scores are very slow to move. Data from the Current Population Survey are also very slow to move. You can see that for cohorts who entered school after the mid-1980s for a period of time – to the early 1990s – there was a decline in age/grade retention after entry to school. But the interesting thing to me in the present policy context is that for the past three or four years it looks like there's the beginning of an uptick again. And I think that is a very serious matter. Some people may applaud it; I don't. I think we know a lot about the consequences of grade retention.

Figure 2 was for the total population; Figure 3 is the same graph for whites – again you see an uptick. Figure 4 is the same graph for blacks – you see an uptick at the end, primarily at the older ages. It's not so clear what's happening at ages 9 – 11, but at ages 12 - 14 and 15 – 17 there appears to be an increase in age-grade retention. Figure 5 is the same thing for Hispanics. I think that there may have been a major turnaround, in the extent of age/grade retardation that one could only attribute to changing promotional practices, some of which are test based.

So, what does retention do for children? Retention almost always lowers later achievement of students assessed at the same age, and, at best, it has no effect on achievement among students assessed at the same grade level (Holmes 1989; Hauser 2001; Jimerson 2001; Shepard 2004; Jimerson 2004). While most assessments of retention policy focus mainly on same-grade comparisons, as Shepard (2004) observes, the unequivocally negative evidence from same-age comparisons is highly relevant to what students know when they drop out of school. By dint of legal school-attendance requirements as well as the availability of non-school activities, dropout is largely determined by age, and retention multiplies the risk of school

dropout (Roderick 1993; Temple, Reynolds, and Miedel 2000; Alexander, Entwisle, and Kabbani 2001; Temple, Reynolds, and Ou 2004; Alexander, Entwisle, Dauber, and Kabbani 2004; Hauser, Simmons, and Pager Forthcoming). The long term costs of retention are high to students, and they're also very high to school systems, although typical school system accounting does not assess the cost of retention. School systems just deal with the number of students at each grade level, and they don't get disaggregated by years in grade. Very frequently, when retention is done in early grades, the negative effects occur years later, and thus are invisible to decision makers.

You should be skeptical of claims that retention works, such as in Karl Alexander, Doris Entwisle, and Susan Dauber's Baltimore studies (Alexander, Entwisle, and Dauber 2003), or in Texas (Dworkin 1999; Lorence, Dworkin, Toenjes, and Hill 2002), or in Chicago where we now know the sad story; and now, forthcoming, in New York. This is a lesson that people simply refuse to learn as each experience reiterates what we knew already. It's very difficult to understand. On the collective level, it is akin to Einstein's definition of insanity: "Doing the same thing over and over again and expecting different results."

My experience with high stakes educational tests did not end with the release of the NRC report, but we have won neither battles nor wars in consequence. I list these activities because I think they are a form of public sociology. I testified about our findings and recommendations before the Senate Committee on Health, Education, Labor, and Pensions, and there I met the late Senator Paul Wellstone of Minnesota. Wellstone invited me to give a staff briefing. He introduced legislation to write our recommendations into law, which was voted down overwhelmingly in the senate. I was told that this was not so much a matter of opposition to the appropriate use of tests, but of unwillingness to support unfunded federal mandates on the states.

Yet no less a light than California's Senator Feinstein introduced an amendment to Wellstone's bill that would have prohibited social promotion (United States and Congress 2000:S1074):

> "That is the practice of passing children on to the next grade regardless of whether they make passing grades. It is called social promotion. While this practice may be politically correct, it has, I believe, become the single most important factor leading to the decline in quality of public education in America. Under our amendment, in order to receive Federal funds, States would be required to prohibit the practice of social promotion and adopt achievement standards in the core academic subjects."

I had brief contact with a lawyer who was hoping to sue the State of Louisiana for its use of high stakes tests to retain elementary students, but that effort barely got off the ground.

I prepared statements for a local advocacy group that had hoped to block the plan of the Chicago Public Schools to retain students at several grade levels if they did not pass a single administration of the Iowa Test of Basic Skills or, failing that, successfully retake the test after a cram course in summer school. On the basis of the first evaluation, in late 1999 (Roderick, Bryk, Jacob, Easton, and Allensworth 1999), I predicted the ultimate failure of the project, and I failed to see the claimed evidence of gains reported in a second evaluation in 2000 (Roderick, Nagaoka, Bacon, and Easton 2000). The failure of the Chicago retention plan has at last been recognized in major reports on achievement test performance and high school dropout that were released in spring 2004 (Nagaoka and Roderick 2004; Allensworth 2004).

I debated social promotion on the editorial page of *USA Today*, a Gannett newspaper that was unique among major media outlets in touting Texas's supposed success in retaining students. It was not, in my opinion, at all coincidental that this report was released as George W. Bush was

gearing up for his move from the state house to the White House. I also wrote an editorial on the subject for *Education Week*, and spoke at conferences held by the NAACP and by the Urban League.

I prepared a brief for the New York Performance Standards Consortium, which was attempting to preserve its successful efforts to keep minority high school students on track to graduation and college entry. The state-imposed requirement for universal administration of the Regent's Examinations for high school graduation was opposed by schools in the Consortium, and my brief pointed to numerous flaws in the design and proposed operation of the exit exams. The Consortium was opposed by the Regents, who won in court, even though the major contentions of the Consortium had been supported by a technical advisory group convened by the Regents.

This is not a happy story, but I will keep on doing what I can when I can. I believe that the current frenzy of test-based accountability fostered by *No Child Left Behind* will lead to much worse educational abuses of our children in the near future. What should we be asking about the future? What will be the consequences of efforts to raise educational standards both in the immediate and long term? Do we know that reforms will work before we put them in place on a large scale? I think there's a clear answer to that. Are we measuring reforms and their consequences as they take place? Well, sometimes and in some places. Are we balancing the costs and benefits among all parties? Let's remember the problem – the problem is low academic achievement.

Reference List

Alexander, Karl L., Doris Entwisle, and Nader S. Kabbani. 2001. "The Dropout Process in Life Course Perspective: Early Risk Factors at Home and School." *Teachers College Record* 103(5):760-822.

Alexander, Karl L., Doris R. Entwisle, and Susan L. Dauber. 2003. *On the Success of Failure: A Reassessment of the Effects of Retention in the Primary Grades*. 2nd ed. Cambridge: Cambridge University Press.

Alexander, Karl L., Doris R. Entwisle, Susan L. Dauber, and Nader Kabbani. 2004. "Dropout in Relation to Grade Retention: An Accounting of the Beginning School Study." Pp. 5-34 in *Can Unlike Students Learn Together? Grade Retention, Tracking, and Grouping,* Edited by Herbert J. Walberg, Arthur J. Reynolds, and Margaret C. Wang. Greenwich, Connecticut: Information Age Publishing.

Allensworth, Elaine. 2004. *Ending Social Promotion: Dropout Rates in Chicago After Implementation of the Eighth-Grade Promotion Gate*. Chicago, Illinois: Consortium for Chicago School Research.

Ayres, Leonard P. 1909. *Laggards in Our Schools a Study of Retardation and Elimination in City School Systems*. New York: Charities Publication Committee.

Burawoy, Michael. 2004. "Introduction to Public Sociologies: A Symposium From Boston College." *Social Problems* 51(1):103-6.

Burns, M. S., Peg Griffin, Catherine E. Snow, and Committee on the Prevention of Reading

Difficulties in Young Children. 1999. *Starting Out Right : a Guide to Promoting Children's Reading Success*. Washington, DC : National Academy Press.

Citro, Constance, Robert Michael, and Robert Hauser. 1996. "Measuring Poverty: A New Approach." Pp. 32-41 in *Proceedings of the Section on Government Statistics, American Statistical Association* (Orlando, Florida. Alexandria, VA: American Statistical Association.

Dworkin, Anthony G. 1999. "Elementary School Retention and Social Promotion in Texas: An Assessment of Students Who Failed the Reading Section of the TAAS." Report to the Texas Education Agency Houston, Texas: Sociology of Education Research Group, University of Houston.

Hauser, Philip M. 1942. "Proposed Annual Sample Census of Population." *Journal of the American Statistical Association* 37(217):81-88.

Hauser, Robert M. 2001. "Should We End Social Promotion? Truth and Consequences." Pp. 151-78 in *Raising Standards or Raising Barriers? Inequality and High-Stakes Testing in Public Education*, edited by Gary Orfield and Mindy L. Kornhaber. New York: The Century Foundation Press.

———. 2004. "Progress in Schooling." Pp. 271-318 in *Social Inequality*, edited by Kathryn M. Neckerman. New York: Russell Sage Foundation.

Hauser, Robert M., Solon J. Simmons, and Devah I. Pager. Forthcoming. "High School Dropout, Race-Ethnicity, and Social Background From the 1970s to the 1990s." *Dropouts in America: Confronting the Graduation Rate Crisis*, Edited by Gary Orfield. Cambridge,

Massachusetts: Harvard Educational Publishing Group.

Henmon, V. A. C. and Frank O. Holt. 1931. *A Report on the Administration of Scholastic Aptitude Tests to 34,000 High School Seniors in Wisconsin in 1929 and 1930 Prepared for the Committee on Cooperation, Wisconsin Secondary Schools and Colleges*. Madison: Bureau of Guidance and Records of the University of Wisconsin.

Henmon, V. A. C. and M. J. Nelson. 1946. *Henmon-Nelson Tests of Mental Ability, High School Examination - Grades 7 to 12 - Forms A, B, and C. Teacher's Manual*. Boston: Houghton-Mifflin Company.

———. 1954. *The Henmon-Nelson Tests of Mental Ability. Manual for Administration*. Boston: Houghton-Mifflin Company.

Herrnstein, Richard J. and Charles Murray. 1994. *The Bell Curve: Intelligence and Class Structure in American Life*. New York: The Free Press.

Holmes, C. T. 1989. "Grade Level Retention Effects: A Meta-Analysis of Research Studies." Pp. 16-33 in *Flunking Grades: Research and Policies on Retention*, edited by Lorrie A. Shepard and Mary L. Smith. London: The Falmer Press.

Jencks, Christopher, James Crouse, and Peter Mueser. 1983. "The Wisconsin Model of Status Attainment: A National Replication With Improved Measures of Ability and Aspiration." *Sociology of Education* 56(1):3-19.

Jimerson, Shane R. 2001. "Meta-Analysis of Grade Retention Research: Implications for Practice in the 21st Century." *School Psychology Review* 30(3):420-437.

———. 2004. "Is Grade Retention Educational Malpractice? Empirical Evidence From Meta-Analyses Examining the Efficacy of Grade Retention." Pp. 71-96 in *Can Unlike Students Learn Together? Grade Retention, Tracking, and Grouping*, Edited by Herbert J. Walberg, Arthur J. Reynolds, and Margaret C. Wang. Greenwich, Connecticut: Information Age Publishing.

Kohn, Linda T., Janet Corrigan, Molla S. Donaldson, Institute of Medicine (U.S.), and Committee on Quality of Health Care in America. 2000. *To Err Is Human : Building a Safer Health System*. Washington, D.C. : National Academy Press.

Lorence, Jon, Anthony G. Dworkin, Laurence A. Toenjes, and Antwanette N. Hill. 2002. "Grade Retention and Social Promotion in Texas: Academic Achievement Among Elementary School Students." Pp. 13-52 in *Brookings Papers on Educational Policy 2002*, edited by Diane Ravitch. Washington, DC: The Brookings Institution.

Nagaoka, Jenny and Melissa Roderick. 2004. *Ending Social Promotion: The Effects of Retention*. Chicago, Illinois: Consortium for Chicago School Research.

National Research Council. 1989. *A Common Destiny: Blacks and American Society*. edited by Gerald David Jaynes and Robin M. Williams, Jr., Committee on the Status of Black Americans, Commission on Behavioral and Social Sciences. Washington, D.C.: National Academy Press.

———. 1993. *The Future of the Survey of Income and Program Participation*. Edited by Constance Citro and Graham Kalton, Panel to Evaluate the Survey of Income and Program Participation, Committee on National Statistics, Commission on Behavioral and

Social Sciences and Education. Washington, D.C.: National Academy Press.

National Research Council. 1999. *Evaluation of the Voluntary National Tests : Year 2 : Final Report*. eds Lauress L. Wise, Richard J. Noeth, and Judith A. Koenig. Washington, D.C. : National Academy Press.

National Research Council. 2003. *Scientific Research in Education*. editors Richard J. Shavelson and Lisa Towne. Washington, DC: National Academy Press.

National Research Council. 2004. *The 2000 Census: Counting Under Adversity*. Constance F. Citro, Daniel L. Cork, and Janet L. Norwood, Eds. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, D.C.: National Academy Press.

National Research Council, Committee on Appropriate Test Use. 1999. *High Stakes: Testing for Tracking, Promotion, and Graduation*. edited by Jay Heubert and Robert M. Hauser. Washington, DC: National Academy Press.

National Research Council, Committee on Embedding Common Test Items in State and District Assessments. 1999. *Embedding Questions: The Pursuit of a Common Measure in Uncommon Tests*. edited by Daniel M. Koretz, Meryl W. Bertenthal, and Bert F. Green. Washington, DC: National Academy Press.

National Research Council, Committee on Equivalency and Linkage of Educational Tests. 1999. *Uncommon Measures: Equivalence and Linkage Among Educational Tests*. edited by Michael J. Feuer, Paul W. Holland, Bert F. Green, Meryl W. Bertenthal, and F. C. Hemphill. Washington, DC: National Academy Press.

National Research Council. Committee on the Prevention of Reading Difficulties in Young Children. 1998. *Preventing Reading Difficulties in Young Children*. Edited by Catherine E. Snow, M. S. Burns, and Peg Griffin. Washington, DC : National Academy Press.

National Research Council, Panel on Institutional Review Boards, Surveys, and Social Science Research. 2003. *Protecting Participants and Facilitating Social and Behavioral Sciences Research*. Edited by Constance F. Citro, Daniel R. Ilgen, and Cora B. Marret. Washington, DC: National Academies Press.

National Research Council, Panel on Poverty and Family Assistance. 1995. *Measuring Poverty: a New Approach*, Edited by Constance F. Citro and Robert T. Michael. Washington, D.C: National Academy Press.

Phelps, Richard P. 2000. "High Stakes Testing for Tracking, Promotion, and Graduation." *Educational and Psychological Measurement* 60(6):992-99.

Phelps, Richard P. 2003. *Kill the Messenger : the War on Standardized Testing*. New Brunswick, N.J. : Transaction Publishers.

Roderick, Melissa. 1993. *The Path to Dropping Out: Evidence for Intervention*. Westport, Connecticut: Auburn House.

Roderick, Melissa, Anthony S. Bryk, Brian A. Jacob, John Q. Easton, and Elaine Allensworth. 1999. *Ending Social Promotion: Results From the First Two Years*. Chicago, Illinois: Consortium for Chicago School Research.

Roderick, Melissa, Jenny Nagaoka, Jen Bacon, and John Q. Easton. 2000. *Update: Ending Social*

Promotion - Passing, Retention, and Achievement Trends Among Promoted and Retained Students, 1995-1999. Chicago, Illinois: Consortium for Chicago School Research.

Rogosa, David. 1999. *How Accurate Are the STAR National Percentile Rank Scores for Individual Students? An Interpretative Guide*. Stanford, CA: Stanford University.

Sewell, William H., Robert M. Hauser, Kristen W. Springer, and Taissa S. Hauser. 2003. "As We Age: The Wisconsin Longitudinal Study, 1957-2001." Pp. 3-111 in *Research in Social Stratification and Mobility*, vol. 20, edited by Kevin Leicht. London: Elsevier.

Shepard, Lorrie A. 2004. "Understanding Research on the Consequences of Retention." Pp. 183-202 in *Can Unlike Students Learn Together? Grade Retention, Tracking, and Grouping*, Edited by Herbert J. Walberg, Arthur J. Reynolds, and Margaret C. Wang. Greenwich, Connecticut: Information Age Publishing.

Steinhauer, Jennifer. 8 Apr 2004. "Mayor Says Prevention Is Key in Plan to Hold Back Students." *The New York Times*.

Temple, Judy A., Arthur J. Reynolds, and Wendy T. Miedel. 2000. "Can Early Intervention Prevent High School Dropout? Evidence From the Chicago Child-Parent Centers." *Urban Education* 35(1):31-56.

Temple, Judy A., Arthur J. Reynolds, and Suh-Ruu Ou. 2004. "Grade Retention and School Dropout: Another Look at the Evidence." Pp. 183-202 in *Can Unlike Students Learn Together? Grade Retention, Tracking, and Grouping*, Edited by Herbert J. Walberg, Arthur J. Reynolds, and Margaret C. Wang. Greenwich, Connecticut: Information Age Publishing.

U.S. Department of Education,Office for Civil Rights. 2000. *The Use of Tests When Making High-Stakes Decisions for Students: A Resource Guide for Educators and Policymakers*. Washington, DC: U.S. Department of Education.

United States and Congress. 2000. *Congressional Record: Proceedings and Debates of the 106th Congress, Second Session*. Daily ed. ed. Washington, DC: U.S. G.P.O.

Weinberg, Daniel H. 2002. "The Survey of Income and Program Participation: Recent History and Future Developments." *SIPP Working Papers*. No. 232. Washington, DC: U.S. Bureau of the Census.

Wise, Lauress L., Robert M. Hauser, Karen J. Mitchell, and Michael J. Feuer. 1999. *Evaluation of the Voluntary National Tests : Phase 1 Report*. Washington, D.C. : National Academy Press.

Figure 1. Cohort Trends in Age-Grade Retardation

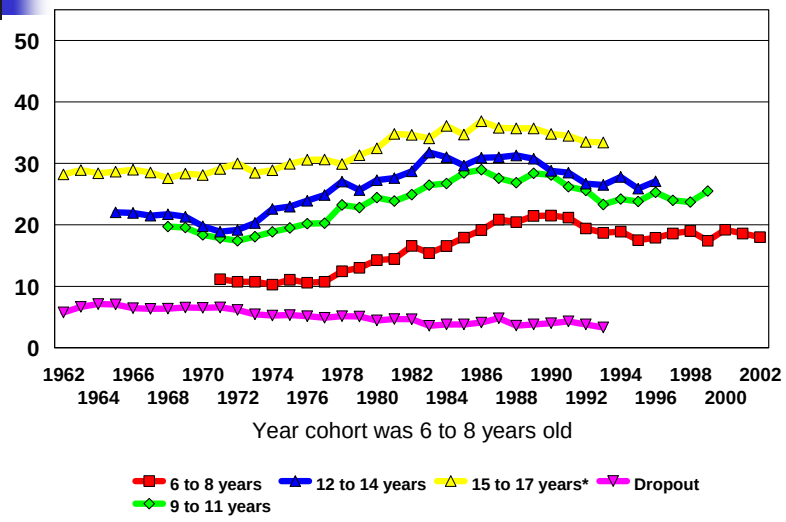

Figure 2. Cohort Trends in Smoothed Conditional Age-Grade Retardation

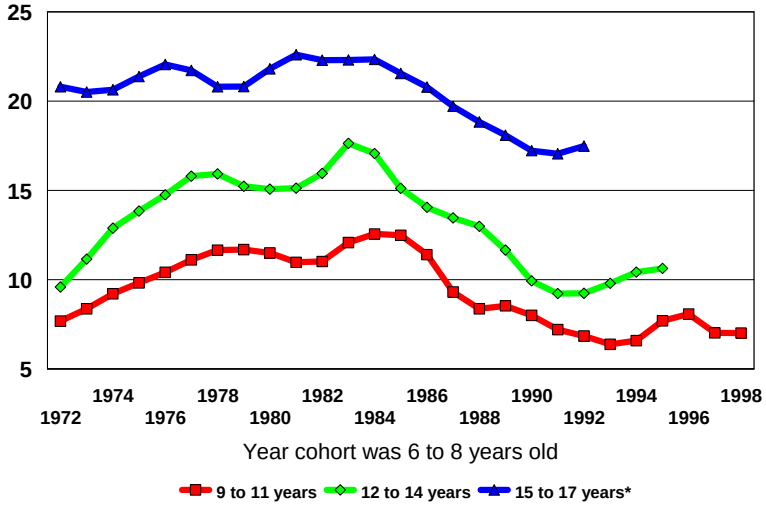

Figure 3. Smoothed Conditional Age-Grade Retardation, Whites

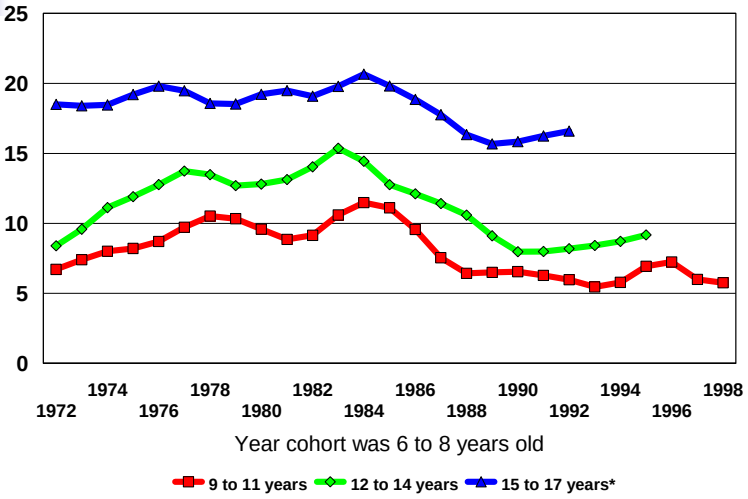

Figure 4. Smoothed Conditional Age-Grade Retardation, Blacks

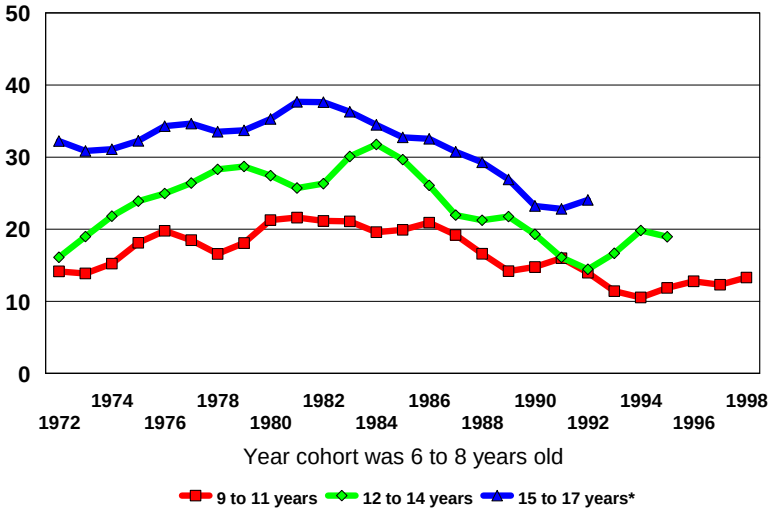

Figure 5. Smoothed Conditional Age-Grade Retardation, Hispanics