University of California at Berkeley

Is There

Solid Evidence of Positive Effects for High School Students?

David Stern, Jean Yonemura Wing

IS THERE SOLID EVIDENCE OF POSITIVE EFFECTS FOR HIGH SCHOOL STUDENTS?

David Stern and Jean Yonemura Wing

University of California, Berkeley

January 24, 2004

Career Academy Support Network
University of California, Berkeley
Graduate School of Education
Berkeley, CA 94720-1670
http://casn.berkeley.edu
ask_casn@berkeley.edu

Prepared for a conference on "High School Reform: Using Evidence to Improve Policy and Practice", organized by MDRC, New Orleans, January 22-23, 2004.

IS THERE SOLID EVIDENCE OF POSITIVE EFFECTS FOR HIGH SCHOOL STUDENTS?

In this paper we illustrate the use of a strict standard for evaluating evidence on programs and strategies designed to improve outcomes for high school students. We explain what we mean by solid evidence, and present examples from multi-site evaluations of three programs. After that we examine some of the evidence on high school size, and explain why clear inferences about cause and effect remain elusive. We also look at examples of studies that use data for the whole student population in large districts, as an approach to reduce possible selection bias. We conclude with a predictable recommendation for more rigorous evaluation, and a programmatic suggestion.

Although we concur that random-assignment studies provide the best support for inferences about cause and effect, we reject the idea that random assignment should be the method of all or most educational research. Before a program, strategy, or intervention can be tested by random assignment, it has to be formulated from exploratory research, and tried out in non-experimental settings. History, ethnography, case study, design study, and other kinds of research all contribute to the understanding of educational phenomena and the development of new ideas. Even at the final stage of testing the effectiveness of a particular intervention, qualitative information about the experience of participants is useful in suggesting why effects do or do not occur, and how the program or strategy might be further improved. Also, as mentioned below, some hypotheses do not lend themselves to testing by random assignment.

Cautionary tales

"Over the past 30 years, more than two dozen comparison-group studies have found hormone replacement therapy for postmenopausal women to be effective in reducing the women's risk of coronary heart disease, by about 35-50 percent. But when hormone replacement therapy was finally evaluated in two large-scale randomized controlled trials — medicine's 'gold standard' — it was actually found to do the opposite: it increased the risk of heart disease, as well as stroke and breast cancer." (U.S. Department of Education 2003, section I.C.2; emphasis in original)[1]

Education, more than modern medicine, is notoriously susceptible to fads. Remember the school-to-work movement? At one high-level meeting in the mid-1990s, the front of the conference folder had "The School-to-Work Movement" printed on stick-on labels that were not quite firmly attached. The labels covered what had first been printed on the folder by mistake: "The School-to-Work Moment." In retrospect, this was probably a more accurate title. There are a number of reasons why the school-to-work movement did not last.[2] Some of the reasons are political: the federal law that fueled the movement was allowed to sunset. But another important reason for the movement's demise was the lack of strong evidence that "school-to-work" reforms produced positive results. There was evidence of sorts, but it was not compelling. For instance, none of the positive results cited in two reviews of the evidence were produced by random-assignment evaluations (Stern et al. 1995, Urquiola et al. 1997). Of course no one can know whether stronger evidence would have persuaded policy makers to sustain the school-to-work movement. But this is one among many examples of reform movements in education that have come and gone, leaving behind too little enduring knowledge.

Efforts to improve education should use and produce solid evidence

[1] It is important to note that these random-assignment trials tested effects of a combined dose of estrogen and progestin, but in practice estrogen is often given alone.
[2] In some ways and in some places the movement continues, but it has certainly lost momentum.

Efforts to improve education should be guided by evidence of what has worked in the past. And current efforts should continue to collect evidence to inform the future. Few would argue with these assertions.

The more important and difficult question is what evidence to believe. The general criterion we would use is that claims of cause and effect should be clear and subject to a minimum of reasonable doubt.[3] For practical purposes, this kind of evidence gives the greatest assurance that a particular strategy will produce the desired effect when applied in new situations.

Among the many perils and pitfalls researchers and evaluators face in trying to make clear causal inferences, we would highlight two: reciprocal causation and selection bias. Both of these are well known, and we have nothing original to say about them. But we find that both are sometimes overlooked in discussions of what is known about effects of high school reform.

Reciprocal causation means that two variables may each be a cause of the other. For instance, suppose a study finds that motivation and grades are positively correlated among a group of students. The explanation could be that stronger motivation has caused some students to study harder and achieve better grades. Or achieving better grades may have caused those students to feel more motivated. Or both could be true. This is one reason for the adage that correlation does not imply causation.

Selection bias occurs when participants in a program or treatment differ from non-participants on one or more unmeasured variables that are related to the outcome of interest. For instance, suppose extra instruction were offered to students after the end of the regular school day, and an evaluation compared gains over time for students who did and did not participate. If the participants tend to be students who volunteer because they are more motivated, and if the study does not adequately measure motivation, then the evaluation would overestimate the effect of the program. On the other hand, if teachers specifically recruited the least motivated students and motivation were not measured, the evaluation would underestimate the program's impact. As Heckman's (1979)

[3] For a careful and practical discussion of what cause and effect may mean, see Shadish, Cook, and Campbell (2002). One useful definition of cause is "an insufficient but non-redundant part of an unnecessary but sufficient condition" (p. 4).

classic paper pointed out, this kind of selection bias is part of the more general class of problems where unmeasured variables are correlated with the outcome and also with one of the measured predictor variables.

Bias can arise not only from the initial selection of participants, but also as a result of selective attrition from a program over time. It is usually a fair assumption that students who complete a program differ in unmeasured ways from students who drop out, and that those differences are relevant to what the program was trying to accomplish.

Both problems — reciprocal causation and selection bias — can be avoided by designing an intervention that is relevant to the question being asked, and assigning participants at random. For instance, if the question is how much does motivation affect grades, the intervention could be some kind of counseling or experience designed to increase motivation. Randomly assigning participants would ensure that their unmeasured characteristics do not differ much, on average, from non-participants if both groups are large.[4]

Random-assignment studies in education have limitations, which are also well known. Some questions do not readily lend themselves to experimental manipulation. For instance, if the question is how much do grades affect motivation, it would be difficult to justify random assignment of students' actual grades. Even in situations where an experimental intervention can be designed, it is generally difficult to arrange a uniform non-treatment condition for the control group. Unlike medical research, educational evaluations usually cannot administer a placebo, so the control group receives a mix of "brand X" or "regular school" experiences, and the evaluation becomes a comparison of a fairly well-defined treatment versus a less well-defined set of alternatives. In some educational evaluations, students assigned to the control group even have managed somehow to sneak into the treatment group.

More common is movement in the other direction: some students who are randomly assigned to a program never actually participate at all. Others begin but leave before completing the program. This creates a common dilemma for evaluators: excluding no-shows or early leavers from the treatment group would

[4] How big the samples have to be in order to reduce the average difference to a given level depends on the distribution of the unmeasured variable.

defeat the purpose of random assignment, because these students may well differ in unobserved ways from those who do show up and complete the program. But including them dilutes the measured impact of the program. One common procedure to correct for no-shows is to divide the measured impact by the proportion of students assigned to the program who do at least begin it (Bloom 1984). This procedure makes some plausible assumptions about unmeasured differences between treatment and control groups.[5] Various attempts have also been made to correct for attrition from the treatment group, but these require stronger assumptions about absence of unmeasured differences between early leavers and program completers. If such assumptions were plausible, there would be less need for random assignment in the first place.

Despite these problems with random-assignment studies, a well-implemented random-assignment design provides the clearest and strongest evidence about cause and effect (Mosteller, Light, and Sachs 1996; U.S. Department of Education 2003).[6] For this paper, therefore, we tried to find random-assignment evaluations of programs that had been implemented in multiple sites. We wanted multiple-site programs because it is most useful to know about strategies that have already been successfully replicated. We also limit consideration here to programs that bear on the institutional design of high schools — arrangements of activities in time and space — not including methods for teaching specific subjects in classrooms.

Some examples of solid evidence

Here we provide brief summaries of three programs that have produced positive impacts for high school students. Each program has been replicated at multiple sites, and has been evaluated using random assignment. The impacts

[5] The assumptions are that the program has no effect on students who did not show up, and the probability of being a no-show would have been the same in the control group as in the treatment group (Myers and Schirm 1999, p. B-7).

[6] In situations where random assignment cannot be done, other designs may offer the best evidence possible. Such designs include careful matching of individual participants with non-participants who are very similar, use of exogenous instrumental variables as proxies for endogenous differences in educational experience, or interrupted time series analysis of schools or districts before and after a particular intervention. See Shadish, Cook, and Campbell (2002), Slavin (2002).

we present here are the most conservative estimates reported in the source documents; they are not adjusted for no-shows or degree of participation.[7] Among other outcomes, we focus especially on high school completion, because a high school diploma becomes more and more important in the labor market as the options available to high school dropouts continue to shrink (Levin 2001). We know these three programs are not the only ones that have been evaluated by random assignment, and we do not claim they are the only programs for which there is solid evidence of positive impacts. We present these as exemplars, and would be happy to know there are others.[8]

Quantum Opportunity Program (QOP). Two separate evaluations, both using random assignment, have found that QOP (prononounced quop) significantly increased high school completion rates, among other positive outcomes. A community-based organization at each site is responsible for putting in place the QOP model which combines the following features:

- Each participant has an adult counselor who acts as case manager and advocate. In theory, and often in practice, counselors are accessible to students by telephone or pager 24 hours a day, seven days a week.
- Participants remain in the program whether they change schools, drop out, become incarcerated, or move out of state. The program's motto is, "once in QOP, always in QOP."
- Educational services include individual assessment and planning, tutoring in high school subjects, and computer-assisted instruction.
- Developmental activities promote life skills and employment readiness, in addition to cultural exposure and recreation.
- Participants perform services that benefit the community.

[7] Some of the impacts we report here have been adjusted to take into account measured differences between treatment and control groups. As described in the source documents, these adjustments used either regression or propensity scores. Such adjustments reduce the standard errors of estimated impacts, but do not make assumptions about unmeasured variables.

[8] Our search for solid evidence was greatly facilitated by the excellent compendia of programs compiled by the American Youth Policy Forum (1997, 1999; Jurich and Estes 2000; James, Jurich, and Estes 2001), and the meta-analysis of Comprehensive School Reform model results by Borman et al. (2003).

- Participants are paid about a dollar per hour spent in QOP activities other than recreation or mentoring, and an equal amount is deposited in an accrual account to be used for postsecondary education or training.

Participants typically engage in roughly 200 to 300 hours of QOP activity each year. The cost per participant per year is on the order of $5,000 in 2004 dollars. Implementation, participation patterns, and cost vary considerably over time and among sites (Hahn 1999; Maxfield, Castner, Maralani, and Vencill 2003; Schirm, Rodriguez-Planas, Maxfield, and Tuttle 2003).

Andrew Hahn and associates at Brandeis evaluated the Ford-funded QOP pilot program in five cities from 1989 to 1993. Hahn (1999) summarized the findings. At each site the evaluators randomly assigned 25 students to QOP and 25 to a control group, from a list of exiting eighth graders whose families were receiving one or more forms of public assistance. Hahn emphasizes that the evaluation deliberately did not require students to apply to QOP, in order to test program operators' ability to recruit low-income students to participate. The Brandeis researchers administered questionnaires in the fall for five years and in the spring of senior year. They also administered tests of academic and functional skills during each year of high school. After the first two years, test scores improved for the QOP group relative to controls. In the fall after scheduled graduation, the survey found these statistically significant differences, among others (p. 247):

	Assigned to QOP	Control Group
Percent graduated[9]	63	42
Percent dropped out[10]	23	50

[9] Hahn (1999) does not indicate whether this includes recipients of GEDs as well as regular diplomas.
[10] Defined as not having graduated and not currently in school.

Percent in postsecondary education or training	42	16

Allen Schirm and colleagues at Mathematica evaluated the QOP demonstration funded by Ford in two cities and the U.S. Department of Labor in five cities from 1995 to 2001. Maxfield, Castner, Maralani, and Vencill (2003) describe implementation results. Schirm, Rodriguez-Planas, Maxfield, and Tuttle (2003) give a detailed analysis of impacts on students. Maxfield, Schirm, and Rodriguez-Planas (2003) summarize both implementation and impacts. Like the Brandeis study, the Mathematica evaluation deliberately did not ask students to apply to QOP. Instead, participants and controls were randomly selected from the population of students in the bottom two-thirds of the GPA distribution among those entering ninth grade for the first time at a high school where the dropout rate was at least 40 percent.[11] However, of the 2,550 students who met these criteria, only 1,069 returned signed consent forms to participate in the evaluation, so there was an element of volition.

In addition to the intake information used to select the sample, Schirm and associates conducted an in-person survey and achievement test in the spring of year four, and a telephone survey in year five. They also tried to collect transcripts from all high schools participants attended. Midway through the year after scheduled high school graduation, the following statistically significant differences emerged (Schirm et al. 2003, Tables V.1, V.3):

	Assigned to QOP	Control Group
Percent graduated from regular high school	46	40
Percent with regular diploma or GED certificate or still in high	79	72

[11] Students deemed by the school to be too disabled to participate in the program were excluded.

school or a GED program		
Percent attending postsecondary education or training[12]	32	26

The Mathematica evaluation found smaller impacts than the Brandeis study, but they do confirm the earlier findings. These results are important because random-assignment studies of other programs to reduce high school dropout rates often have failed to find significant impacts (e.g. Dynarski et al. 1998, Kemple 2001). These two evaluations provide solid evidence that QOP boosts educational attainment by students in populations where high school completion rates are low.

Upward Bound. Created by the federal Higher Education Act in 1965, Upward Bound is a long-established, well known, and widely distributed program to increase access to college for students whose families have low incomes or whose parents have not attended college. In 1992 the U.S. Department of education began the first large-scale, random-assignment evaluation of Upward Bound. The first phase of the study followed most students through high school and some of the older students into postsecondary education (Myers and Schirm 1999).

Most Upward Bound projects are operated by institutions of higher education, which provide academic counseling, tutoring, and enrichment to participating high school students during the school year and, usually, intensive academic programs on the college campus during the summer. The evaluation classified all Upward Bound projects by type of college sponsor — public or private, two- or four-year — and by urban or rural location, then drew a stratified random sample of projects to represent the program population. Within each project, eligible applicants were randomly assigned to Upward Bound or the control group. In a number of projects, applicants were first

[12] Includes armed forces.

classified by characteristics such as race or gender, then randomly assigned within strata. The assignment process occurred over a 14 month period from 1992 to 1994.

In addition to questionnaire data collected on applicants at the time of selection, the first phase of the evaluation conducted telephone surveys and collected school transcripts in 1994-95 and 1996-97. Most students were in grade 9 or 10 when the study began, and in 1996-97 their high school status was as follows (Myers and Schirm 1999, Table III.2). The difference in the percent still in high school is statistically significant; the differences in the other two rows are not.

	Assigned to Upward Bound	Control Group
Percent graduated from high school[13]	59	63
Percent still in high school	35	28
Dropped out	6	9

For the sample as a whole, the only other significant impacts as of 1996 were that students assigned to Upward Bound had formed higher expectations regarding their eventual education attainment, and they had completed more high school credits in math and social studies.

The evaluation found more statistically significant impacts for particular subgroups of students (Table V.1). Among students who initially indicated they did not expect to complete a bachelor's degree (21 percent of the study sample), those assigned to Upward Bound were more likely to have graduated from high school by 1996, and they were less likely to have dropped out (Table III.7). Students below the median on an index of academic performance in grade 9

[13] Report does not indicate whether this includes recipients of GEDs as well as regular diplomas.

were also less likely to drop out and more likely to graduate by 1996 if assigned to Upward Bound (Table III.15). Students from low-income families (82 percent of the sample), Hispanics (23 percent) and whites (21 percent) assigned to Upward Bound also were less likely to drop out, and the Hispanic students were more likely to be still attending high school (Tables III.11, III.13). Boys (29 percent of the sample) were less likely to have dropped out if assigned to Upward Bound (Table III.9). In addition to these impacts on high school status, the evaluation found significant impacts for these same subgroups on educational expectations and the number of credits earned in various high school courses.

A report on the second phase of the evaluation was made available to us as a draft for review (Myers et al. 2003). This incorporated results from a survey in 1998-2000. By then, 90 percent of the sample had graduated from high school, 3 percent had obtained GED certificates, and 7 percent had dropped out. There were no significant differences in these outcomes between students who had or had not been assigned to Upward Bound. The only significant difference in high school performance for the sample as a whole was that students assigned to Upward Bound completed more credits in math (Myers et al. 2003, Table II.5). Impacts of Upward Bound on high school graduation, dropout rates, and GED completion also were no longer significant among subgroups defined by low initial educational expectations or weak educational records in grade 9 (Tables II.6, II.7).

The 1998-2000 survey contained questions about postsecondary education, including names of any schools attended. Evaluators then attempted to obtain respondents' transcripts from those schools. These attempts produced information that either verified the respondent's claim, falsified the claim, or were ambiguous.[14] Using only verified enrollment to calculate enrollment rates may understate true enrollment rates, but using unverified enrollment would overstate them. Myers et al. (2003) therefore present both sets of results. For the sample as a whole, the only impact of Upward Bound on postsecondary

[14] An example of an ambigous result is an institution responding that it could not release student records without written permission, from which the evaluators could not tell whether the particular student was enrolled or not.

enrollment was in increase in enrollment at four-year colleges, which was significant in the unverified but not quite significant in the verified data (Tables III.1, III.2).

Among students who had initially indicated they did not expect to obtain bachelor's degrees, the impact on enrollment and credits earned in four-year colleges was significant using both kinds of data (Tables III.3, III.4). Dividing students by academic records in grade 9, the unverified data showed positive impacts on four-year college enrollment for both high and low achievers, but the verified data showed the impact was significant only for the students who did better in grade 9 (Tables III.5, III.6). Both verified and unverified data showed a positive impact for Hispanics on enrollment and credits at four-year colleges or other postsecondary schools (Tables III.9, III.10). For both males and females, the unverified data indicated a positive impact on four-year college attendance, but the verified data showed only a positive impact for males on attendance at any postsecondary school (Tables III.11, III.12).

In sum, the first phase of the evaluation indicated that Upward Bound improved high school performance especially for low-income Hispanic and white males who start high school with low educational expectations and weak academic records. However, the follow-up survey three years later, when the entire sample was past high school, found many of the earlier apparent high school impacts had attenuated or disappeared. Postsecondary impacts were absent or ambiguous for the sample as a whole and for several subgroups. But Upward Bound did increase the rate of four-year college attendance by about 20 percentage points among students who had not expected to earn bachelor's degrees at the time the evaluation began. And among Hispanics, Upward Bound boosted the four-year college-going rate by 12 to 14 percentage points.

Career academies. The term "career academy" was coined by Stern, Raby, and Dayton (1992) to describe a kind of high school program that had originated in Philadelphia in 1969, then spread to California, New York City, and eventually nationwide, encouraged in part by positive results from several quasi-experimental evaluations (e.g. Reller 1987; Stern 2003 summarizes the evidence). There is no authoritative, uniform definition of a career academy, and as the term

has become popular the variation among programs that call themselves career academies has increased.[15] Common themes for career academies are health, business and finance, arts and communications, computers, engineering, law and government.

In 1993 MDRC began the first random-assignment evaluation of career academies (Kemple and Rock 1996). MDRC abstracted three main features to define a career academy:

- School-within-a-school organization in which academy students at each grade level take a set of classes together, and stay with the same small group of teachers from one year to the next.
- Curriculum that includes academic courses meeting college entrance requirements, and technical classes, all related to the academy theme.
- Employer partnerships to provide internships and other experiences outside the classroom, related to the academy theme.

The evaluation began with ten sites, but one academy ceased operating. All nine remaining academies are in high schools with large proportions of low-income and minority students. Each was the only career academy in the school.

At the start of the evaluation, the academies recruited more applicants than they could accommodate. Applicants knew they might not be admitted. MDRC randomly assigned about two-thirds of the applicants to the academy; the others became the control group. In the ten years since the evaluation began, MDRC collected student records, surveyed students during each of their high school years, and conducted follow-up surveys one year and four years after high school.

During the high school years, career academies produced several positive impacts on students' experience and achievement. Compared to the control group, academy students reported receiving more support from teachers and from other students (Kemple 1997). They were more likely to combine academic

[15] The state of California provides grants to school districts for "partnership academies" which are defined by statute, but this definition does not apply to the hundreds of academies in California that do not receive state funding. A few other states also have funded such academies. The federal School-to-Work Opportunities Act in 1994 included career academies on a list of seven "promising practices," but did not not define them. Building on the MDRC definition, the Career Academy Support Network (http://casn.berkeley.edu) has negotiated a common definition among several networks currently promoting career academies.

and technical courses, engage in career development activities, and work in jobs connected to school (Kemple, Poglinco, and Snipes 1999). As of spring of senior year, academies retained a larger fraction of the students whose initial characteristics made them more likely to drop out (Kemple and Snipes 2000). Among students at less risk of dropping out, academies increased participation in technical courses and career development activities without reducing academic course credits (Kemple and Snipes 2000).

The first follow-up survey, one year after scheduled graduation, found no significant impacts on students' high school completion, GED acquisition, or participation in postsecondary schooling. It also showed no significant impact on employment or earnings, though students who had been assigned to career academies were working and earning somewhat more than the control group (Kemple 2001).

The most recent follow-up, about four years after scheduled graduation from high school, found large and significant impacts on employment and earnings, and no difference in educational attainment (Kemple 2003). In the full sample, students assigned to career academies earned higher hourly wages, worked more hours per week, had more months of employment, and earned about 10 percent more per month than the control group. All these differences occurred for both males and females, but they were not statistically significant for females. The MDRC evaluation distinguished between students at high, medium, or low risk of dropping out of high school, as predicted by variables measured before random assignment. Academies had significant positive impacts on average hours worked per week within the 25 percent at high risk, on average hourly wages for the 50 percent at medium risk, and on average monthly earnings for both these groups. Impacts on high school completion or postsecondary education were not significant for the sample as a whole or for any subgroup, but Kemple (2003) notes that both the academy and control groups had high rates of high school completion and postsecondary enrollment compared to national (NELS) data on urban high school students.

In sum, the MDRC evaluation found that career academies gave students more personal support, career guidance, technical classes, and school-supervised work experience during high school. Academies also succeeded in retaining

more high-risk students through spring of senior year. Eventual impacts on high school graduation or postsecondary education were not significantly positive or negative for the sample as a whole or for any subgroups. But academies had substantial positive impacts on employment and earnings after high school, especially for young men and for students whose initial characteristics indicated high or medium risk of not finishing high school.

A shared feature: accommodating student mobility. These studies provide solid evidence that some interventions have produced positive impacts for young people who start high school with poor academic records, low educational expectations, or other challenging circumstances. Although we have focused more on evaluation methods than on program design, we note that the three programs described here to some extent share a common feature: they can accommodate students who move. QOP explicitly emphasizes trying to stay connected with participants even when they move around, institutionally or geographically: "once in QOP, always in QOP." Upward Bound also can accommodate some mobility of participants among high schools, because an Upward Bound project typically serves students in several high schools near the college where the project is located. Career academies are less able to keep students who move, because an academy is rooted in its home high school. But some academies do enroll students from other high schools or districts. Accommodating student mobility is important because so many students move in and out of high school or from one school to another, sometimes in the middle of the school year, and students who move more often are less likely to finish high school.

Elusive inference: effect of small size in high schools

We turn our attention now to studies that attempt to draw strong causal inferences from evidence not produced by random assignment. To illustrate the difficulty of drawing such inferences, we focus on studies about effects of small size in high schools, a variable which has been given paramount emphasis in current reform strategy. We have not reviewed all the empirical studies on this

topic, but we have selected some of the best and most often cited. These studies are informative, and some are ingenious. But they leave considerable room for doubt about the extent to which smaller school size causes better results for students.

The main problem here is the influence of unobserved variables. For example, several frequently cited studies found that smaller high schools have lower dropout rates (Fetler 1989, Franklin and Crone 1992, Howley and Bickel 1999, Pittman and Haughwout 1987). Each study compared high schools in a state or national sample at one point in time. Some of the smaller high schools would be located in smaller, close-knit suburban or rural communities — the kind of place where teachers and administrators send their own children to the school where they work. Students who cut classes are more liable to be caught if they live in a community where more people know one another, so cutting classes would be less likely, and would less often lead to dropping out of school entirely. In big cities, more of the small high schools would be magnets or other schools of choice. In these situations as well, stronger social cohesion and shared values among parents and teachers could account for the lower dropout rates. The density of personal connections and strength of shared expectations among parents and school staff are unmeasured variables in these studies. Socioeconomic variables used as statistical controls do not capture these differences. The association between smaller school size and lower dropout rates, therefore, could be at least partly due to smaller high schools occurring in particular kinds of circumstances that account for the better results.

Unmeasured variables also may influence the selection of certain kinds of students into particular small schools, or into smaller subschools within large high schools. Various studies have found that students in smaller schools are relatively less alienated, more engaged, and more likely to pass courses and earn credits toward graduation (see reviews by Cotton 1996, Gladden 1998, Raywid 1995). Studies also have found better student performance in smaller learning communities (SLCs) within large urban high schools (McMullan et al. 1994, Oxley 1990, Wasley et al. 2000).[16]

[16] Stern (2003) reviewed these studies in more detail.

However, these results may be largely attributable to small schools or SLCs enrolling students whose unmeasured, pre-existing characteristics would have made them more likely to perform better in any situation. In metropolitan areas, small schools are often magnets, alternative schools, or other schools of choice. Similarly, SLCs within larger high schools usually enroll students who choose to be there. Students who are more motivated or better organized, or whose parents are more concerned about their schooling, may be more likely to exercise choice in the first place. Schools and SLCs naturally seek to enroll and retain students with these kinds of qualities. These characteristics of students and families, not measured by researchers, could account in part for the students' better performance. The ongoing process of mutual selection may result in small schools or SLCs enrolling more students whose unmeasured, pre-existing characteristics would make them more likely to succeed anywhere. One indication of this dynamic is the finding by Wasley et al. (2000) that a lower dropout rate among SLC students occurred in high schools where only some students were in SLCs, but not in high schools where all students were in SLCs. In instances where converting an entire school to SLCs has led to better outcomes (e.g. McPartland et al. 1998), it may not be clear whether some low-performing or misbehaving students who would have attended the school before the transformation do not enroll there after the change.

One way to avoid selection bias in testing whether small school size causes better student performance would be to use random assignment. Students could be randomly assigned to large schools, small schools or SLCs. We have not yet found such studies. High school programs in the random-assignment evaluation of the School Dropout Demonstration Assistance Program all had small enrollments (Dynarski et al. 1998). None of these programs increased the proportion of students earning regular high school diplomas, but the focus of this evaluation was not small size per se.

Another approach would be to randomly assign entire high school attendance zones or school districts to enroll in large or small schools. We have not yet found such a study. Gottfredson (1985) did observe what happened in five high schools where major enrollment changes suddenly occurred as a result of district reorganization. In two high schools that became bigger, there was no

change in reported drug use or delinquency, teachers' expressed feeling of safety decreased in one school, and students' reports of victimization by other students increased in one school. In three high schools that became smaller, reports of drug use and delinquency increased in two schools, teachers' feeling of safety improved in one school, and students' reports of victimization increased in one school. These results probably reflect some changes in student population as well as change in school size, but they do not indicate that the size change was decisive.

Current strategies to improve high schools seldom rely on smaller size alone. Lee and Smith (2001) argue that small size itself is not a direct cause of better student performance, but "smaller school size is a facilitating factor for creating organizational features of schools that we have shown to be important determinants of learning." (p. 157) Those organizational factors include teachers' sense of collective responsibility for learning, students taking more math and science courses, and use of more authentic instructional practices (Table 6.3). These findings are derived from an elegant statistical analysis of NELS data, using hierarchical models to distinguish between the connection of school characteristics to average achievement (excellence), and their connection to the within-school correlation of achievement with socioeconomic status (equity).

Lee and Smith's analysis of high schools is theoretically strong and empirically sophisticated. Nevertheless, it leaves open several questions about the effects of school size. Lee and Smith (2001) do not present evidence that smaller school size is associated with teachers' sense of collective responsibility for learning, students taking more math and science courses, or use of more authentic instructional practices. Even if these characteristics are more apt to be present in smaller schools, the observed association between these school features and student learning could be attributed to reciprocal causation. For instance, Lee and Smith measure teachers' sense of collective responsibility by their responses to 12 survey items including, "I can get through to the most difficult student," "Teachers make a difference in students' lives," and (with reverse scoring) "Students are incapable of learning the material" (p. 190). Teachers may be more inclined to give positive answers to these and the other

items as a result of being in a school where students are more successful. Likewise, students may take more math and science courses, and may be exposed to more challenging instructional methods, because they are successful learners. So it is not clear to what extent these school characteristics are the cause or effect of student learning.

Except in one chapter, Lee and Smith's (2001) statistical models include high school enrollment as as single number among other school characteristics in a linear combination of predictors. But in a separate chapter focusing on size itself, Lee and Smith divide schools into eight categories by enrollment, and find that students in schools with enrollments of 600 to 900 had the biggest average gains in achievement, compared to larger or smaller schools.[17] This result raises additional questions. Are the school characteristics they found to be associated with student learning also most prevalent in this same size category, compared to schools that are larger or smaller? Are there other, unmeasured characteristics of schools that may be concentrated in this size range? For instance, community characteristics may be different in very small rural schools or very large urban schools, compared to medium-sized schools in suburbs or small towns. Or a larger proportion of schools in the 600 to 900 range may be magnets or other schools of choice.

A study that viewed size as one factor among others, and also gave careful attention to student selection, is the account by Darling-Hammond, Ancess, and Ort (2002) of changes at Julia Richman High School in New York City. They point out that small size is not a sufficient condition for improvement: "Not all small schools are successful." (p. 642) They describe the transformation of Julia Richman from a large high school into a set of small, autonomous schools sharing the same site. The new, small schools built strong relationships between and among students and faculty by reducing the pupil load for each teacher and creating new advising structures; developed more coherent curriculum; engaged students in active learning; used portfolios and exhibitions to assess students' work; and provided time for teachers to collaborate.

[17] The distribution of achievement was least associated with socioeconomic status in all size categories below 600, compared to bigger schools.

Darling-Hammond, Ancess, and Ort (2002) paid careful attention to possible selection bias. In the new schools' first year, the student body was comprised mostly of students from the Julia Richman attendance zone "who had not applied elsewhere or had been rejected by their chosen school." (p. 645) Seventy percent were eligible for subsidized lunch, compared to 32 percent of students at Julia Richman in the previous year. Some selective attrition occurred in the first couple of years, as many students who "had not proactively chosen the schools" moved out (p. 648). But analysis of students who entered the new schools as ninth graders in 1994, excluding transfers in or out, found a four-year graduation rate of 73.3 percent, "significantly higher than the comparable New York City rate of 49.7 percent for the same cohort." (p. 649) Six-year graduation rates were also higher. Even though some transfers out of the comparison schools would graduate from a school other than the one they entered in ninth grade, these results suggest that the new schools at the Julia Richman site had stronger than average holding power. In addition, eleventh graders (presumably including those who transferred in) at the new small schools outperformed students in similar schools on New York State Regents examinations for reading and writing, though not for math. Among graduates from the new schools, college-going rates were 86 percent in 1997 and 91 percent in 1998.

The attention given to selection and attrition by Darling-Hammond, Ancess, and Ort (2002) makes this a more persuasive study. How much of the observed effect is attributable to the new schools' small size remains unclear.

Selection and choice. The likelihood of selection bias pervades much of the existing research on effects of small high schools and small learning communities (SLCs) within large high schools. If the apparent positive results of small size are largely due to selection of students with positive unmeasured traits such as motivation, then transforming all large high schools into smaller ones would not accomplish much. In addition, teachers in new small schools or SLCs also may be self-selected. If these teachers possess more motivation, commitment, energy, creativity, or other positive traits, then positive results from these small settings may not generalize to the system as a whole. There is a danger that current attempts to downsize all high schools may be based on a

fallacy of composition, a mistaken hope that what is observed in specific cases can be generalized to the whole high school population.

Although self-selection of students and teachers makes it more difficult to draw clear causal inferences, self-selection could be a good thing in a programmatic sense. It is possible that particular high schools or SLCs are good for students who choose them, but not for other students. If that were the case, the best arrangement might be to let students choose from an array of large and small schools or SLCs. Some large districts are already doing that.

Whether expanding school choice improves outcomes for students is itself a vigorously contested empirical question. Studies to date have focused mainly on elementary schools, but Cullen, Jacob, and Levitt (2003) have studied effects of high school choice in Chicago. To control for the possibility that students applying to a particular school might share certain unobserved characteristics, they focused on 19 high schools that used random lotteries to select students. They found that students who won a lottery at the time they entered ninth grade did not perform better academically in grade 9 or 10, compared to students who did not win in the same lottery. As economists, they viewed these findings as "surprising" (p. 4). In another paper (forthcoming), the same authors used proximity to different kinds of high schools as exogenous instruments to estimate the effects of choosing to enroll in one of 12 high-achieving schools, 10 career academies, or 39 other schools. They focused on whether students successfully completed high school, and found positive effects only for the career academies.[18]

Studies using data for whole school districts

One way to reduce possible selection bias is to study the whole student population in a big school district. For example, if a school district increased the number of small high schools or SLCs, evidence on districtwide trends in student performance could reveal the extent of gains for students choosing these options,

[18] Career academies in Chicago are different from the model described above and evaluated by MDRC. Chicago career academies are full-sized high schools that emphasize career and technical education.

as well as any possible negative trends among the students left behind.[19] McMullan, Sipe, and Wolf (1994) did this kind of analysis in Philadelphia, where the district, encouraged by the Pew Charitable Trust, greatly expanded the number of high school SLCs (called "charters") from 1988-89 to 1993-94. The proportion of high school students enrolled in SLCs rose steadily over this period, but most districtwide indicators of academic performance, after some initial gains, leveled off or went back down. The authors suggested that the gains due to SLCs might have been offset in the later years by changes in district policy that moved more over-age middle-school students into high schools and also cut summer school.

Another study of districtwide effects was the evaluation by Bohrnstedt et al. (1999) of Equity 2000, a program by the College Board to increase math course-taking, college preparation, and college enrollment among low-income Hispanic and African American students. Results are reported for six urban districts that enacted policies for all students to take first-year algebra by grade 9 and geometry by grade 10, and provided various kinds of support for teachers to make this happen. Course taking and other outcomes were measured by surveys given to all graduating seniors in three successive cohorts.[20] Results show larger proportions of students in the later cohorts took algebra by grade 9 and geometry by grade 10. Increases in geometry course enrollment by grade 10 were greater for Hispanics and African Americans than for Asians or whites. However, there were no apparent gains in the proportions of students taking advanced math courses or college entrance examinations.

Snipes, Doolittle, and Herlihy (2002) used districtwide data in a study of successful urban districts (Snipes, Doolittle, and Herlihy 2002). Several urban districts were chosen from different parts of the country based on evidence that student achievement had improved for at least three years, and that differences in average achievement between white and minority students had narrowed. As in the "effective schools" studies of the 1970s and 1980s, the purpose here was not to test the impact of an intervention that was defined ex ante, but instead to

[19] Selection bias may still occur if there is substantial change in the numbers or kinds of students who enroll in private schools or other districts.
[20] Response rates by cohort were 52, 61, and 64 percent (p. 14).

try to identify practices that might account for these districts' apparent success. To strengthen inferences about which practices mattered, the MDRC study also visited two comparison districts that were similar in some ways but had not improved student achievement. We mention this study as an example of the districtwide method, but, unfortunately for our purposes, the districts studied had not experienced gains in student achievement at the high school level (pp. 138-141, 106-109).

Studies of comprehensive high school reform models

A considerable amount of recent and ongoing effort has been focused on evaluating federally identified "comprehensive school reform" (CSR) models, but strong evidence is not yet available on CSR models at the high school level. An extensive meta-analysis by Borman et al. (2003) summarized the effects on student achievement of 29 widely implemented CSR models. Only two models were designed specifically for high schools, grades 9-12: High Schools That Work (HSTW) and Talent Development High Schools (TDHS). In contrast, there were four CSR models for elementary grades K-5, seven models for grades K-8, and 16 CSR models for grades K-12. Some of the K-12 models have been studied in high schools, but the meta-analysis combined into one category all studies that included any students in grades 6 through 12.

Of the 232 studies that met the inclusion criteria for the CSR meta-analysis, 45 reported measures of student achievement from HSTW and one from TDHS (McPartland et al. 1998). All these studies were sponsored by the models' developers, except one study of HSTW. None of these studies used random assignment. The HSTW studies also rely mainly on senior-year data from successive cohorts of students who completed defined sequences of academic and vocational courses. Changes in HSTW results over time may reflect changes in the composition of the students selected.

Evaluations of CSR models are continuing, however. It is possible that one or more models may yet produce solid evidence of effects for high school students.

Recommendations

Our purpose here was to illustrate the application of strict scrutiny to claims of cause and effect in studies of programs or strategies for high school students. We described three examples of multi-site evaluations that produced solid evidence of positive impacts. We hope there are other examples already published or forthcoming. Since our search was not exhaustive, we do not claim

to provide a comprehensive review of everything known to be effective for high school students. Nevertheless, we will offer two recommendations.

First, increase investment in long-term evaluations using random assignment. The three evaluations we describe each took about a decade to produce clear findings. Given the severity of problems in American high schools, attempts to make improvement must go forward. But more of these attempts should be accompanied by random-assignment evaluations. In some situations, such as initiatives to expand choice among schools or small learning communities within schools, use of lotteries to select students provides a natural opportunity for this kind of evaluation. Even when it is not built into the program, random assignment of students, classrooms, schools, or entire districts should be done more often, and more resources should be spent on data collection and analysis. Assigning people to control groups does bar them from interventions that may be beneficial, and spending more on evaluation may take money from program operation. But these harms may be less than the possible damage caused by promoting massive changes without good evidence that they are producing desired results.

Second, on a more positive note, the random-assignment evaluations of QOP, Upward Bound, and career academies have produced solid evidence on which to build. In addition to justifying more replication of these programs in their current forms, the results may point the way to further development, evolution, or hybridization of these initiatives.[21] We noted, in particular, that the three programs, especially QOP, all accommodate students who move. This is important because students who change schools more often are less likely to finish high school successfully. Many current high school reforms are attempting to build small learning communities, intended to nurture sustained interpersonal relationships, from which students can benefit only if they stay there for some period of time. QOP has shown that it is possible to form a relationship that continues for several years between a high school student and a caring adult, even when the student does not remain in one place.

[21] Any new versions or hybrids should be rigorously evaluated, of course!

References

American Youth Policy Forum (1997). *Some things DO make a difference for youth: A compendium of evaluations of youth programs and practices.* Washington, DC: AYPF.

American Youth Policy Forum (1999). *More things that DO make a difference for youth: A compendium of evaluations of youth programs and practices, Volume II.* Washington, DC: AYPF.

Bloom, H. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review* 8:225-246.

Borman, G. D., Hewes, G. M., Overman, L. T., and Brown, S. (2003, Summer). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research 73*(2), pp. 125-230.

Bohrnstedt, G., Jakwerth, P., Rodriguez, C., and Quiñones, S. (1999, September). *The senior survey analysis of Cohorts 1, 2, and 3.* Report No. 87. Palo Alto, CA: American Institutes for Research, John. C. Flanagan Research Center.

Cotton, K. (1996). *Affective and Social Benefits of Small-Scale Schooling* (ERIC Digest No. RC 96-5). Charleston, WV: Clearinghouse on Rural Education and Small Schools. (ERIC Document Reproduction Service No. ED 401 088)

Cullen, J. B., Jacob, B. A., and Levitt, S. D. (2003, November). The effect of school choice on student outcomes: Evidence from randomized lotteries. Working Paper 10113. Cambridge, MA: National Bureau of Economic Research.

Cullen, J. B., Jacob, B. A., and Levitt, S. D. (forthcoming). The impact of school choice on student outcomes: An analysis of the Chicago Public Schools. To be published in the *Journal of Public Economics.*

Darling-Hammond, L., Ancess, J., and Ort, S.W. (2002). Reinventing high school: Outcomes of the Coalition Campus Schools Project. *American Educational Research Journal* 39(3):639-673.

Dynarski, M., Gleason, P., Rangarajan, A., and Wood, R. (1998, June). *Impacts of dropout prevention programs: Final Report.* MPR Rerference No.: 8014. A research report from the School Dropout Demonstration Assistance Program Evaluation, submitted to the U.S. Department of Education

Planning and Evaluation Service. Princeton, NJ: Mathematica Policy Research, Inc.

Elliott, M. N., Hanser, L. M., and Gilroy, C. L. (2000). *Evidence of positive student outcomes in JROTC career academies.* Santa Monica, CA: RAND Corporation.

Fetler, M. (1989). School dropout rates, academic performance, size, and poverty: Correlates of educational reform. *Educational Evaluation and Policy Analysis, 11,* 109–116.

Franklin, B. J., and Crone, L. J. (1992, November). *School accountability: Predictors and indicators of Louisiana school effectiveness.* Paper presented at the meeting of the Mid-South Educational Research Association, Knoxville, TN.

Gladden, R. (1998). The small school movement: A review of the literature. In M. Fine and J. I. Somerville (Eds.), *Small schools, big imaginations: A creative look at urban public schools* (pp. 113–137). Chicago: Cross City Campaign for Urban School Reform,

Gottfredson, D. (1985). *School size and school disorder.* Baltimore, MD: Johns Hopkins University, Center for Social Organization of Schools.

Hahn, A. (1999). Extending the time of learning. In D. J. Besharov (Ed.), *America's disconnected youth: toward a preventative strategy.* Washington, DC: Child Welfare League of America Press.

Heckman, J. J. (1979, January). Sample selection bias as a specification error. *Econometrica*, 47(1): 153-161.

Howley, C. B., and Bickel, R. *The Matthew Project: National report.* Albany, OH: Ohio University.

James, D. W., Jurich, S., and Estes, S. (2001). Raising minority academic achievement: A compendium of education programs and practices. Washington, DC: American Youth Policy Forum.

Jurich, S., and Estes, S. (2000). Raising academic achievement: A study of 20 successful programs. Washington, DC: American Youth Policy Forum.

Kemple, J. J. (1997). *Career academies: Communities of support for students and teachers. Emerging findings from a 10-site evaluation.* New York: Manpower Demonstration Research Corporation.

Kemple, J. J. (2001). *Career academies: Impacts on Students' Initial Transitions to Post-Secondary Education and Employment.* New York: Manpower Demonstration Research Corporation.

Kemple, J. J. (December 2003 DRAFT). *Career academies: Impacts on Labor Market Outcomes and Educational Attainment.* New York: Manpower Demonstration Research Corporation.

Kemple, J. J., Poglinco, S.M., and Snipes, J. C. (1999). *Career academies: Building career awareness and work-based learning activities through employer partnerships.* New York: Manpower Demonstration Research Corporation.

Kemple, J. J., and Rock, J. L. (1996). *Career academies: Early implementation lessons from a 10-Site evaluation.* New York: Manpower Demonstration Research Corporation.

Kemple, J. J., and Snipes, J. C. (2000). *Career academies: Impacts on students' engagement and performance in high school.* New York: Manpower Demonstration Research Corporation.

LaPoint, V., Jordan, W., McPartland, J. M., and Towns, D. P. (1996). *The Talent Development High School: Essential components.* Baltimore: Johns Hopkins University and Howard University, Center for Research on the Education of Students Placed at Risk.

Lee, V. E., and Smith, J. B. (2001). *Restructuring High Schools for Equity and Excellence: What Works.* New York: Teachers College Press.

Levin, H. M. (2001). High-stakes testing and economic productivity. In G. Orfield and C. Edley (Eds.), *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: Century Foundation, pp. 39–50.

Maxfield, M., Maralani, V., and Vencill, M. (2003, August). *The Quantum Opportunity Program demonstration: Implementation findings*. MPR Reference No. 8279-080. Washington, DC: Mathematica Policy Research, Inc.

Maxfield, M., Schirm, A., and Rodriguez-Planas, N. (2003, August). *The Quantum Opportunity Program Demonstration: Implementation and short-term impacts*. MPR Reference No. 8279-093. Washington, DC: Mathematica Policy Research, Inc.

McMullan, B. J., Sipe, C. L., and Wolf, W.C. (1994). *Charters and student achievement: Early evidence from school restructuring in Philadelphia.* Philadelphia: Center for Assessment and Policy Development

McPartland, J. M., Balfanz, R., Jordan, W., and Legters, N. (1998). Improving climate and achievement in a troubled urban high school through the Talent Development Model. *Journal of Education for Students Placed at Risk, 3*, 337–361.

Mosteller, F., Light, R. J., and Sachs, J. A. (1996). Sustained inquiry in education: Lessons in skill grouping and class size. *Harvard Education Review* 66:707–842.

Myers, D., Olsen, R., Seftor, N., Young, J., and Tuttle, C. (2003 DRAFT). *The Impacts of Regular Upward Bound: Results from the Third Follow-Up Data Collection*. Washington, D.C.: Mathematica Policy Research, Inc.

Myers, D., and Schirm, A. (1999, April). *The impacts of Upward Bound: Final report for Phase I of the national evaluation*. MPR Reference No.: 8046-515. Submitted to the U.S. Department of Education Planning and Evaluation Service. Washington, DC: Mathematica Policy Research, Inc

Oxley, D. (1990). *An analysis of house systems in New York City neighborhood high schools.* Philadelphia: Temple University Center for Research in Human Development and Education.

Pittman, R .B., and Haughwout, P. (1987). Influence of school size on dropout rate. *Educational Evaluation and Policy Analysis, 9,* 337–343.

Raywid, M. A. (1995). *The subschools/small schools movement—Taking stock.* Madison, WI: Center on Organization and Restructuring of Schools.

Reller, D. J. (1987). *A longitudinal study of the graduates of the Peninsula Academies, final report.* Palo Alto, CA: American Institutes for Research in the Behavioral Sciences.

Rodriguez, C., Khattri, N., and Han, Mei (1997, June). *The Equity 2000 evaluation, a summative report: Impact and implementation*. Report No. 86,. Washington, DC: Pelavin Research Center.

Schirm, A., Rodriguez-Planas, N., Maxfield, M., and Tuttle, C. (2003, August). *The Quantum Opportunity Program Demonstration: Short-term impacts*. MPR

Reference No.: 8279-093. Washington, DC: Mathematica Policy Research, Inc.

Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Slavin, R.E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher* 31(7):15-21.

Snipes, J., Doolittle, F., and Herlihy, C. (2002, September). *Foundations for success: Case studies of how urban school systems improve student achievement.* New York: Manpower Demonstration Research Corporation.

Southern Regional Education Board. (1995). *Charting the progress of education: Annual report, June 1995.* Atlanta, GA: Author.

Southern Regional Education Board. (1997). *1997 outstanding practices.* Atlanta, GA: Author.

Stern, D. (2003). Career academies and high school reform before, during, and after the school-to-work movement. In Stull, W. J., and Sanders, N. M. (Eds.): *The School-to-Work Movement, Origins and Destinations*. Westport, CT: Praeger, pp. 239-262.

Stern, D., Finkelstein, N., Stone, J.R. III, Latting, J., and Dornsife, C. (1995). *School to Work: Research on Programs in the United States.* Washington and London: Taylor and Francis, Falmer Press.

Stern, D., Raby, M., and Dayton, C. (1992). *Career academies: Partnerships for reconstructing American high schools.* San Francisco: Jossey-Bass

Urquiola, M., Stern, D., Horn, I., Dornsife, C., Chi, B., Williams, L., Merritt, D., Hughes, K., and Bailey, T. (1997). *School to work, college and career: A review of policy, practice, and results 1993–1997.* Berkeley, CA: National Center for Research in Vocational Education.

U.S. Department of Education (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, D.C.: Institute of Education Sciences, U.S. Department of Education.