NBER WORKING PAPER SERIES

IMPROVING THE PERFORMANCE OF THE EDUCATION SECTOR: THE VALUABLE, CHALLENGING, AND LIMITED ROLE OF RANDOM ASSIGNMENT EVALUATIONS

Richard J. Murnane
Richard R. Nelson

Working Paper 11846
http://www.nber.org/papers/w11846

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2005

We would like to thank Greg Baker for his research assistance, and Patricia Graham and Catherine Snow for helpful comments on an earlier draft. Richard Murnane's work on this project was supported by a grant from the Spencer Foundation. Richard Nelson's work was partially supported by the Sloan Foundation. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Improving the Performance of the Education Sector: The Valuable, Challenging, and Limited Role of Random Assignment Evaluations
Richard J. Murnane and Richard R. Nelson
NBER Working Paper No. 11846
December 2005
JEL No. I21

ABSTRACT

In an attempt to improve the quality of educational research, the U.S. Department of Education''s Institute of Education Sciences has provided funding for 65 randomized controlled trials of educational interventions. We argue that this research methodology is more effective in providing guidance to extremely troubled schools about how to make some progress than guidance to schools trying to move from making some progress to becoming high performance organizations. We also argue that the conventional view of medical research – discoveries made in specialized laboratories that are then tested using randomized control trials -- is an inaccurate description of the sources of advances in medical practice. Moreover, this conventional view of the sources of advances in medical practice leads to incorrect inferences about how to improve educational research. We illustrate this argument using evidence from the history of medical research on the treatment of cystic fibrosis.

Richard J. Murnane
Graduate School of Education
Harvard University
6 Appian Way - Gutman 409
Cambridge, MA 02138
and NBER
richard_murnane@harvard.edu

Richard R. Nelson
Columbia University
rrn2@columbia.edu

INTRODUCTION

In a number of the world's wealthiest countries, including the United States, France, and Germany, the effectiveness of public education is a matter of great concern. International achievement comparisons show that students in these countries do not score nearly as well on math and science tests as students in other countries, particularly in East Asia. Evidence that the skills of a country's labor force are an increasingly important determinant of the rate of economic growth makes this pattern troubling (Hanushek and Kimko, 2000). Particularly disturbing in the U.S. is the relatively poor academic performance of African-American and Latino children, who constitute a growing portion of the nation's population.

The challenge of raising academic performance is made especially difficulty by uncertainty about the most effective strategies for improving education. This issue is especially acute in the United States, where test scores have been remarkably stable over the last 40 years despite a more than 200 percent increase in real per pupil expenditure on public education.[1] In this paper we focus particularly on attempts to improve public education in the United States. However, we believe that the themes are relevant to other countries.

Concerns about the performance of public education in the United States are not new; in fact, complaints about performance go back almost to the nineteenth century birth of public education in this country. What is relatively new, however, are the policy responses of state governments and more recently the federal government. Historically

---

[1] Expressed in 2001-2 constant dollars, expenditures per pupil were $3,066 in the 1961-62 school year and $9,553 in the 2001-02 school year. (See Table 166 on page 204 of Snyder et al., 2004.)

public education was primarily a local activity in the United States, with state and federal governments providing modest financial support and some regulation, but leaving most governance, curriculum, and resource issues to local governments. Beginning in the late 1980s, this situation has changed dramatically. Almost all states have introduced standards-based reform systems that include specification of the skills and knowledge that students should master at each grade level, tests to assess student mastery of the standards, and sanctions for students or for educators when performance is judged low. Passage in 2001 of the federal No child Left Behind (NCLB) legislation brought the federal government more centrally into the picture, and significantly added to the pressure on local schools to improve student performance on state-mandated tests by specifying a definition of the "adequate yearly progress" schools are expected to make. The stipulation that adequate progress must be made not only for the student body as a whole, but for all racial and ethnic groups, represents an attempt to keep the focus on groups that historically have not fared well in American schools.

While the long-run consequences of standards-based education – or test based accountability, as the reforms are often called – are yet to be determined, their introduction has had two striking consequences. First, they have temporarily pushed into the background questions about what constitutes a good education. At least for educational leaders responsible for urban schools, the operational definition of a good education has become one that results in consistent improvement in test scores for all children.[2]

[2] Of course, there are many who question whether higher scores on high stakes tests mean better prepared students (e.g. Koretz, 2005), but to date their influence on policy debates has been modest.

Second, test-based accountability has increased educators' search for strategies to increase test scores. The list of candidates includes more resources, new curricula, more teacher training, better incentives for teachers and students, and changes in governance structure. Every one of these options has its advocates. Yet the evidence on the consequences of adopting any of these options is murky at best. Thus, school and school district leaders face increasing pressure to improve performance, but little reliable guidance on how to do so.

It is in this context that the federal government has developed renewed interest in improving the quality of educational research. The logic is straight-forward. Formal research and development (R&D) has played an important role in improving performance in many sectors, including agriculture and medicine. Much written-about success stories include hybrid corn and the Salk and Sabin polio vaccines. Shouldn't high quality research play an equally important role in improving the performance of the education sector? Educational research, like research in agriculture and public health, has a significant component of public funding. This brings the federal government directly into the business of deciding how research should be done.

A role for the federal government in sponsoring educational research and development dates back to the creation of the U.S. Office of Education in 1867. However, over the next 90 years, the role was extremely modest (Warren, 1974). It increased somewhat with the creation of the National Science Foundation in 1950. However, in that same year a major education bill that had passed the Senate and was supported by President Truman died in a House committee because of fear that federal funding would result in federal control of local schools. It took the 1957 launching of

Sputnik by the Soviet Union to overcome these obstacles to a significant federal role. In 1958 the Congress passed the National Defense Education Act that provided significant funds for educational innovations. Over the next 20 years NSF and the U.S. Office of Education funded a variety of curriculum development and teacher training initiatives (Dow, 1991). However, the emphasis was much more on developing and disseminating new curricula than on research evaluating their impacts on student learning.

The need for a systematic educational research program was a primary reason for the creation of the National Institute of Education. Seen an analogous to the National Institutes of Health, NIE was announced with great fanfare and the promise of dramatic increases in funding. Yet before a decade was past, NIE was dead. Factors contributing to its demise included weak political support for educational research, political errors by its early leaders,[3] the creation of the Department of Education which demoted the agency from equivalence with the US Office of Education, and, finally, the election in 1980 of a president who wanted the federal government to be less involved in education. All of these factors contributed to the sense that NIE had failed to live up to expectations, a judgment typically made without consideration of what realistic expectations might have been.

Over the next 25 years, the federal government has continued to provide some support for educational research and development, albeit with a much lower profile. Most of the U.S. Department of Education's modest budget for R&D has gone to educational research centers and labs, typically connected to schools of education. The centers, which are awarded in periodic competitions, investigate particular themes seen as

[3] For example, an early director chose to attend a meeting in Paris rather than the Senate appropriations committee hearing on NIE's budget.

important to American education. These include improvements in math, science, reading, and math instruction, strategies for improving high schools, methods for increasing adult literacy and learning, research on evaluation, standards, and testing, and policy areas including the role of the states in promoting educational reform. The regional labs provide technical assistance to states and school districts.

While advocates for the labs and centers have touted their contributions, to most legislators they have seemed extremely modest, especially when compared to hybrid corn and the polio vaccines. One oft-proposed explanation is that educational research has been insufficiently "scientific." Of course, this raises the question of what scientific research in education would look like.

In the fall of 2000 the U.S. Department of Education asked the National Research Council to form a committee to address this question. The Committee's 157 page report, released in 2002, was a carefully worded document, proposing six principles for defining scientifically based research, and arguing that choice of research methods needs to be tailored to the question under investigation (Shavelson and Towne, 2002).

While the NRC Committee was deliberating, the Congress passed the No Child Left Behind Act of 2001. The definition of scientifically based research in NCLB was more narrow than that described in the yet-to-be-released NRC report. While the definition is lengthy, the part that the U.S. Department of Education seemed to embrace was the emphasis on the value of randomized experiments.

On November 5, 2002, President Bush signed into law the Education Sciences Reform Act, which authorized the creation of a new organization to sponsor educational research, the Institute of Education Sciences. In its first 28 months of operation, IES

provided funding for 65 random assignment evaluations, including studies of reading comprehension programs, violence prevention programs, after-school programs, teacher preparation, teacher induction, and teacher professional development programs, school choice programs, and educational technology initiatives.[4] The explicit goal of these studies is to provide better evidence to state and local educational policymakers about "what works."

To facilitate access to such information, the U.S. Department of Education has provided significant funding for the "What Works Clearinghouse," an organization with the following purpose:

> On an ongoing basis, the What Works Clearinghouse (WWC) collects, screens, and identifies studies of the effectiveness of educational interventions (programs, products, practices, and policies). We review the studies that have the strongest design, and report on the strengths and weaknesses of those studies against the WWC Evidence Standards so that you know what the best scientific evidence has to say. [5]

To gain a sense of the challenge involved in providing education practitioners with sound advice on what works, it is instructive to review the accomplishments of the What Works Clearinghouse to date. According to its website, as of September 2005, after three years of work, the What Works Clearinghouse has completed the review of research in one study area, middle school math curricula. It examined 77 studies conducted since 1983. Only 10 met its standards of evidence, and only four of those

[4] The figure of 65 random assignment evaluations was provided by Dr. Lynn Okagaki, Deputy Director for Science, IES, in an email message on March 29, 2005. The information on the types of studies funded by IES was taken from the following U.S. Department of Education website: http://www.ed.gov/rschstat/eval/resources/studyplans.html (accessed March 24, 2005).

[5] DoE's $18.5 million dollar grant to the What Works Clearinghouse was awarded in August 2002, a few months prior to the passage of ESRA, the legislation that authorized the creation of the Institute for Education Sciences.

were random assignment evaluations. It concluded that two of the curricula interventions achieved positive statistically significant benefits compared to alternatives. [6]

We note that there are two somewhat different roles that can be played by educational research and by randomized experiments as a research methodology. First, research can be directed toward the creation, development, and evaluation of practices that are significantly different from those broadly used. The purpose here is to advance the frontiers of educational practice. Second, research can be aimed at assessing the effectiveness of different educational practices already in use in some settings, with the aim of defining, and then disseminating, best practice. In the language of introductory economics, the first objective is to expand the production possibility frontier. The second is to help educational organizations move closer to the current frontier.

We think it fair to say that during the NIE years the most common view was that educational research was mainly aimed at the first objective, advancing the frontier of what is possible to accomplish, as research had indeed done in the fields of agriculture and medicine. The more recent orientation represents greater emphasis on the second objective, the identification and dissemination of best practices.

Of course, the distinction between these two objectives for educational research is not clear-cut. In many cases educational research is concerned with exploring new ways of applying ideas that have been around for some time and that some teachers have used in one way or another in their practice. Teaching mathematics through a curriculum that prompts children to develop an understanding of the properties of number systems is one

[6] The information summarized in this paragraph was taken from the following website: http://www.whatworks.ed.gov/ and from the report, math_topic_report.pdf, accessed at that website on Sept. 5, 2005.

example. Today's new math curricula have deep roots in the "new math" of the 1960s (Sarason, 1971).

Another respect in which the distinction between the two objectives of education research is not clear-cut is that many new educational tools, especially technology-based tools, were developed in other sectors and are being adapted to education. Examples include the use of computer-based technology to improve in-service training for teachers and the analysis of student assessment results. Some of these initiatives may eventually have a significant impact on educational practice and performance. However, we note that these technology-based initiatives are not stand-alone new technologies akin to the polio vaccines. Instead, they are tools that may help groups of teachers to work together more effectively in improving their practice.

We will explore these differences in greater depth later. Here, we simply want to state that well designed random assignment evaluations can play a useful role in both in testing the efficacy of new educational tools and practices and in identifying and spreading the best of current practices. However, we will argue that it is important to understand the limits of what greater use of random assignment evaluations can do in education, or in medicine or agriculture for that matter.

In agriculture, in medicine, and in education, random assignment testing can be a useful tool for evaluating well specified, well controlled practices. However, in all of these fields, and certainly in education, there often is much more to effective practice than simply a set of well specified routines. Evidence that particular instructional approaches have been effective in some settings with some students experiencing particular learning problems can be useful to skilled practitioners. However, conducting

random assignment evaluations in a manner that provides such nuanced results is challenging. So is the work of teachers who must identify students with particular problems and learn by trial and error whether a particular instructional technique will help a particular student. Teachers' ability to do this will depend not only on their own skill, but on the characteristics of the school in which they work, including its priorities, support for teamwork, and the quality of diagnostic tests routinely used to assess students' skills.

While we doubt that any analyst familiar with schools would dispute the argument that the same tightly-defined practices will not work equally well in all schools, some would argue that that the challenge is simply to identify the critical interaction effects between the characteristics of the clients and the setting, on the one hand, and the most effective practices, on the other hand. We would argue, however, that it is not possible, ex ante, to identify which practices will be most effective with particular students in particular settings. Figuring this out is a key part of the work in high performing schools. Unpacking this argument is the topic of the rest of this paper.

THE ROUTINE, NON-ROUTINE, AND ORGANIZATIONAL ASPECTS OF EDUCATIONAL PRACTICE

Practice in any area of human activity involves a mix of routines that are relatively standardized, describable, replicable, imitable, and aspects that do not have these characteristics. This is so in medical practice as well as in education, as we will document later in this paper. But in our view, the non-routine aspects of educational practice are especially important to high performance. Practice in any area of activity

involves a mix of actions that are largely at the discretion of individuals, actions that involve tightly coordinated teamwork, and actions that are more loosely coordinated under the broad organizational influences that shape how individuals interact. In education there is typically both a considerable amount of autonomy for individual teachers, and considerable potential for organizational influences to shape how individual teachers act and cooperate.

We believe that random assignment testing and the What Works Clearinghouse are likely to be much more useful for schools that are struggling to achieve a modest level of performance than for schools that have met this objective and now seek to become high performing organizations. In making this point, we find it useful to consider two kinds of schools. No school exactly fits either of these models, but we believe that a great many schools are close to one or the other.

The first kind of school is struggling. Teachers are not particularly well trained, and there is significant turnover of both teachers and students. If the high rates of turnover are taken as given, the key to progress is to identify and implement a set of relatively routinized practices that work reasonably well on average, even under these adverse conditions.

The program in such schools generally will involve both a core standard set of practices, and a set of standardized remedial programs that are invoked for students who are doing especially poorly. Candidates for the latter include enrollment in supplementary classes using scripted strategies designed to improve phonemic awareness, use of computer programs aimed at improving mastery of specific skills, or mandatory summer school with a specific curriculum. Applying these special treatments to different kinds of

problem cases can be recognized as an extension of the basic idea that the key to organizational improvement is to choose the best standardized “treatments,” and train all staff to implement these faithfully.

This approach to organizational improvement dates back to the tradition of scientific management associated with Frederick W. Taylor. When introduced to a school system that previously has no coherent strategy for improving student performance, this approach can result in significant achievement gains. Schools that work largely under this model clearly can benefit greatly from a program of randomized testing of readily usable educational practices that seeks to identify what works in similar contexts.

A problem with this mode of educational practice, however, is that what it can achieve is inherently limited. Children differ significantly and some will have difficulty learning irrespective of the choices of the core curriculum and instructional techniques, and the standard remediation programs used to help students that are experiencing academic difficulties. Educational practice that sticks closely to well established routines is inherently limited in what it can achieve.

An alternative school model embraces from the outset the reality that some children will not make steady academic progress irrespective of the district’s choices of its basic curricula and instructional techniques, and its choices of standardized remedial programs. Adherents of this model of organizational improvement see the critical challenge as identifying rapidly those children not making adequate progress under the standard program, diagnosing their learning difficulties, developing individualized improvement strategies, and monitoring progress closely. They may also see as part of

their work identifying and challenging students who seem to be unusually gifted. More generally, adherents of this model see their primary challenge as developing an organization that is effective at providing consistent, high quality instruction, monitoring continuously the learning of every child, and figuring out a way to help each child who is experiencing learning problems.

We note that not all schools can operate according to this second mode. To do so requires a well educated and relatively stable core of teachers, and resources with which they can work. Many schools and school districts do not meet these conditions. Schools that are able to operate this way clearly can do better than schools that are forced to operate largely through routines. These fortunate schools also can learn lessons from resources like the What Works Clearinghouse since their districts must choose curricula. Evidence that students in similar districts using particular curricula have done especially well on particular tests is relevant to their curriculum adoption decisions. Also, information on instructional techniques that have worked with children experiencing particular learning difficulties can help skilled teachers as they search for ways to help particular children.

However, adherents to this second approach to improving schools are particularly interested in policies that can most effectively improve the skills of teachers and increase their coordination of instruction, and recognize that this is a more complicated endeavor than training teachers to follow a scripted curriculum. The results of random assignment evaluations are likely to be of very limited value in guiding this work.

This is evident from evaluations of a number of initiatives aimed at building the skills of teachers and changing the ways people in schools interact. The introduction of a

new math curriculum in Pittsburgh provides one example. Among the principles underlying Everyday Math, a K-grade 5 math curriculum developed with NSF support, is that students should construct algorithms for conducting numerical calculations rather than memorizing rules. Instead of didactically explaining to students rules to follow, teachers should model solution strategies and then pose problems and encourage students to devise and practice problem-solving approaches. Students should work in groups to develop problem-solving strategies so that they become accustomed to explaining their reasoning. For most Pittsburgh elementary school teachers, Everyday Math was a quite significant departure from the way they had learned math and from the methods they were taught to teach math.

The district invested heavily in teacher training aimed at improving elementary school teachers understanding of mathematics and their skills in teaching the new curriculum. The evaluation report indicated that the average performance of students on mathematics exams aligned with the curriculum did improve between 1996 and 1998. However, there was significant variation in performance across schools. Students in schools in which Everyday math was implemented well did much better on the fourth-grade math exam than did students in schools in which the quality of implementation was poor. A careful reading of the report reveals another important statistic – out of 53 Pittsburgh elementary schools, 8 implemented the new curriculum well, 3 implemented it poorly, and the quality of implementation in the other 45 elementary schools lay between these extremes. This was after several years of quite intensive and quite expensive professional development (Briars and Resnick, 2000). So the lesson for educational leaders seems to be the awkward combination: Everyday Math is an extremely effective

curriculum when it is implemented well, but even with significant investment, it is difficult to implement well.

The random assignment evaluations of James Comer's School Development Program by Thomas Cook and his colleagues provides another example (Cook et al., 1999; and Cook et al., 2000). The essence of the Comer approach to improving schools is that adults, teachers, and parents must form new relationships with each other, and through this process engage students in new ways. The principles underlying the Comer model make sense. However, they are difficult to implement. Many schools that try to change the way adults interact with each other fail to do so. As a result, the evaluations have shown that implementation of the Comer model tends to be weak. Those schools that do implement the model well experience performance improvements. But given the difficulty in implementing the principles, the lessons for educational leaders again are not clear.

RANDOM ASSIGNMENT TESTING OF INCENTIVE SYSTEMS AND BROAD GOVERNING STRUCTURES

Many economists believe that the key to improving the performance of the education system is to design, implement, and evaluate reward systems that create strong incentives for teachers and school administrators to improve students' skills (Hanushek, 1994 and 2004). We share the view that it is important to experiment with alternative incentive systems and to rigorously evaluate the results with randomized controlled trials. We see as especially interesting the incentives a number of districts introduced to attract skilled teachers to schools serving concentrations of poor children.

At the same time, our interpretation of the quite limited available evidence is that creating strong financial incentives to improve student achievement in the face of limited knowledge about how to do this is problematic. We base this judgment on evidence about the consequences of pay-for-performance for educators and greater competition for public schools.

Pay for performance for individual teachers, -- or merit pay as it is commonly known – has a long history in American education. Literally thousands of school districts have tried merit pay at one time or another over the last century. However, almost all dropped it within five years. Since there has never been a well designed evaluation of any merit pay program, it is not clear what the consequences for students and teachers of any specific plan were. However, a common view is that the demise of the plans stemmed, at least in part, from the recognition that teachers need to work together to create an effective school and that merit pay for individual teachers generates an environment that inhibits cooperation (Murnane and Cohen, 1986).

In recent years a number of school districts and states have introduced school-based pay-for-performance plans that provide financial rewards to the faculties of schools whose students perform well on mandatory tests. The plans vary and to date little is known about their consequences. However, an evaluation of the North Carolina plan showed that the criteria for a financial reward were much more difficult to meet in schools serving large percentages of disadvantaged children (Ladd and Walsh, 2002). As a result teachers tried to avoid working in these schools. One lesson is the difficulty of getting the incentives right. A second is the pressing need for knowledge about how to improve schools serving concentrations of disadvantaged students. The reason is that

strong incentives in the face of a lack of knowledge about how to bring about desired outcomes are likely to elicit dysfunctional responses.

Another approach to improving incentives is to introduce voucher plans that provide financial support to parents that choose to send their children to non-public schools. The logic is that greater competition will induce public schools to improve their performance. In the last decade there have been several random assignment evaluations of small scale voucher programs aimed at providing greater school options for students from low income families. While advocates and opponents of vouchers emphasize different aspects of the evaluation results, most would agree that the results were mixed.[7] The main lesson we take away, again, is that strong incentives alone will not result in markedly improved performance in the face of quite limited knowledge about how to create effective education for disadvantaged students.

So while we agree that it is important to learn more about the consequences of alternative incentive regimes, this is not a substitute for research aimed at increasing knowledge of how to improve production processes in schools. This is especially true since the clients of schools are children. Even if improved incentives would lead over time to a winnowing out of ineffective schools, the cost of failures in terms of children's futures is enormous.

SIMILARITIES AND DIFFERENCES BETWEEN MEDICINE AND EDUCATION: TREATING CYSTIC FIBROSIS

[7] For example, Rouse (1998) reported that students participating in the Milwaukee voucher experiment did a little better than the control group in math, but not in reading. Howell et al. report that black children attending private schools as a result of their participation in the New York City Scholarship program scored

In this section we consider the treatment of cystic fibrosis as a way of illuminating both the similarities and differences between the roles of research in medicine and education. We draw principally on two sources. One is an article by John Littlewood (2002) that describes the series of advances in understanding and treatment of cystic fibrosis over the past sixty years that have greatly increased the expected lifespan and the quality of life of those that have the disease. The other is an article by Atul Gawande (2004), which describes the very significant differences in the effectiveness of treatment of cystic fibrosis at the present time among centers that specialize in the treatment of the disease, with reflections on the factors that lie behind those differences.

Medical scientists and physicians now know that cystic fibrosis is a genetic disease. Since 1989 they even know the location on the genome where the problem can arise. They know that the source of the disease is a mutation that reduces the ability of cells to manage chloride. The result is that various bodily secretions are thickened, which makes the body less able to absorb food, and which gradually clogs the lungs. These infirmities decrease the ability of the body to resist various infections.

Sixty years ago these facts were not known. In fact, only in 1938 was cystic fibrosis described as a definite clinical entity. While some medical scientists strongly believed that the disease was inherited in some way, others were not so sure. A prominent belief was that the disease was related to vitamin A deficiency. It was widely understood that those with the disease could not absorb food normally, and also were vulnerable to a variety of infections. The disease was understood as a children's disease because, until recently, most people that inherited it died before the age of ten.

---

better on math and reading exams than did children in the black children in the control group, but this was not the case for Hispanic children.

Since the end of World War II there have been significant advances in treatment, which increased somewhat the life expectancy of children with the disease. Most of the advances have been of two sorts. One involved discovery of nutrients that those with the disease seemed to be able to absorb. The other was the discovery of antibiotics that kept many of the infections from being fatal. There apparently was growing understanding that the root problem was inability of the body to deal with chlorides. However, Littlewood's discussion of the various advances in treatment does not indicate that this understanding had much to do with the discovery of better treatments. That discovery process seems to have been largely experimental.

As new treatments were discovered or invented and tried, a process that up until the 1950s seems largely to have been the result of the work of individual medical scientists and doctors, news of how they seemed to be working was spread through the professional community. Testing seems to have been done largely by individual physicians, physician groups, or clinics. Until the 1970s, there does not seem to have been much care taken to design careful statistical analysis, much less conduct random assignment evaluations. But certain of the new treatments had positive effects that were quickly visable to the physicians using them, and the word of these spread. Others turned out not to be effective. In some cases this became apparent relatively quickly. In other cases, it took some time for initial enthusiasm to wane. Littlewood gives the example of "mist tent therapy."

The 1950s and 1960s saw the emergence of important new institutions dedicated to collaborative research on the causes and treatment of cystic fibrosis. These included the Cystic Fibrosis Foundation and a collection of clinics that specialized in treatment of

the disease. There is little doubt that the new institutions improved the flow of information. The foundation also promoted evaluations of alternative treatments, including, during the 1970s, the first double blind nutritional intervention study. However, the role of systematic evaluation research seems to have been modest relative to the role played by the judgment of physicians based on their experiences with their own patients. And the new treatments described by Littlewood seem generally to have been of the form that skilled physicians could adopt them reasonably effectively, in some cases after training.

Littlewood judges that by the 1980s understanding of the disease had become significantly stronger. These advances in understanding led to increasingly reliable diagnostic techniques. But it is not clear that it had much of an effect on treatments, where the old experimental process seems to have continued to be the rule, with some significant improvements being found in nutrition, means of warding off infection, and physiotherapy.

Littlewood proposes that in the 1990s the new understanding of the genetic basis of cystic fibrosis had begun to influence strongly the search for better treatment. While he predicts that gene therapies will become available within ten to fifteen years, as of the time he wrote his article (2002) nothing much seems to have come from this route (although a phase 1 clinical trial involving gene therapy using compacted DNA has recently been completed).[8]

We would like to highlight several important features of this story which bear on the current debate regarding how to improve educational practice. A flow of new

[8] Information taken from the following website on September 5, 2005: http://www.cff.org/research/clinical_trials/ongoing_trials/gene_therapies/

treatments was coming out of research done primarily by physicians working with their own patients. Particularly as diagnostic tests of lung function improved, the evidence that these new treatments were effective was relatively obvious to the physicians treating the patients. Communication among the professional community was a vital aspect of the process of improvement. The system of professional publications, conventions, and personal communications, many supported by the Cystic Fibrosis Foundation, provided an equivalent to the What Works Clearinghouse.

Thus far the story we have told of progress in treating cystic fibrosis sounds a lot like the story in the minds of those who are advocating more scientific research, more systematic evaluation of alternative treatments, and a better way of communicating what works and what doesn't in education. One difference is that the rate of introduction of promising new ways of treating clients has been more rapid in treating cystic fibrosis than in educating students. But the broad belief that careful evaluation and reliable and timely dissemination of information bearing on efficacy certainly is supported by the cystic fibrosis case.

There is another aspect of the cystic fibrosis case, however, that is strongly reminiscent of our proposition that there is much about good practice in a field that is very difficult to describe in terms of a set of routines to follow or in terms of a well defined program that can be evaluated by random assignment testing.

The 2004 article by Atul Gawande on the differences in the efficacy of treatment of cystic fibrosis in different specialized centers focuses on just this. His basic theme is that there are very great differences across the hundred or so treatment centers in the United States in the average length of life of the patients they treat. Moreover, the same

centers are the best performing year after year. Little of this variation can be explained by differences in the client mixes or in the package of treatments different centers use – they all use the standard treatments that the most current research has found to be the most effective. Nor can the center to center variation be explained by differences in the credentials of staff.

This phenomenon of large site to site performance differences is not unique to treatment for cystic fibrosis. It is a common pattern in medicine and in many other areas of social services, including counseling of welfare recipients and foster children, and education. Just as with cystic fibrosis treatment centers, little of the site to site performance variation in these centers can be explained by differences in the packages of standardized treatment techniques or credentials of staff. So what is the explanation?

According to Gawande, the best centers for treating cystic fibrosis are especially effective at monitoring key indicators of patients' health, especially lung function, identifying rapidly patients whose lung function declines, training all staff to diagnose the source of the problem – often stemming from changes in patients' personal behaviors --, working with the patient to develop a tailored improvement strategy, and monitoring progress closely.

Gawande's explanation is consistent with several points Littlewood makes. First, more than 1,000 mutations of the gene causing cystic fibrosis have been identified and patients with different mutations experience different symptoms and need different treatments. Second, the environment in which patients live affects the treatments they need. Third, patients' health depends critically on their behaviors including what they eat and the diligence with which they carry out daily treatment regimens. Since patients

often notice the effects on their health only after an extended period of neglecting their daily treatments, there is a tendency to skip these treatments. However, doing so results in loss of lung function that is very difficult to reverse. Consequently, part of the work in the best treatment centers is to convince patients to carry out their treatments faithfully, to monitor indicators of health closely, and to attempt to intervene quickly when even a small decline in lung function is detected.

An implication of Gawande's explanation is that research aimed at explaining performance variation across centers needs to focus on why some organizations are more effective learning organizations than others are. This includes paying attention to improving patient monitoring systems, building the capacity of staff to identify and diagnose problem cases, creating incentives for staff to do this critical work, and developing mechanisms to learn from failures.

MAKING THE RIGHT INFERENCES FROM MEDICAL RESEARCH

As we have noted, much of the policy discussion about the role research could and should play in the advance of educational practice takes modern medicine as a model. However, the view of how progress occurs in these fields and the roles research plays tends to be somewhat oversimplified. We hope that a more accurate understanding of how advances in medical practice actually occur can help orient the discussion of how educational research can contribute to expanding the frontier of best practice in education. In the next section, we attempt to provide perspective by arguing that while educational research will not approach the power of research in advancing practice in medicine, we believe it may have at least as much scope and power as research on

business practice. We think the similarities and differences there are important to reflect upon.

Much of the discussion comparing educational research with research in the field of medicine assumes several things about medical research, explicitly or implicitly, that are not quite accurate. The first is that research exploring ways to improve practice has been sharply oriented by strong scientific understanding. The second is that significant improvements in practice almost always are the result of prior advances in scientific understanding. The third is that virtually all advances in the efficacy of medical practice take place through the same series of steps: advances in scientific understanding lead to more effective research to find better practice, which when achieved is handed down to the medical practice community.

The history of advances in the treatment of cystic fibrosis, which we presented earlier, should cast doubt on all of these assumptions. Much of the research that in fact achieved better practice was guided only very loosely by sophisticated scientific understanding. Rather, it was exploratory and experimental, guided by broad understanding of the disease and the needs of patients, with deep scientific understanding largely in the background. While in many cases the significant advances that were achieved involved the use of new materials (for example, antibiotics) and equipment (imaging devices) that had been developed upstream from the fibrosis research endeavor, and usually for different purposes, it does not appear that advances in deep understanding of fibrosis have played a particularly major role in enabling advances in practice, although this may be changing. And there has been and continues to be a considerable two-way interaction between biomedical scientists doing research concerned with cystic

fibrosis and physicians treating patients and conceiving as well as testing new procedures.

The way advances have occurred in the treatment of cystic fibrosis is not atypical. The same kind of a story can be told in many disease areas. Among the various medical technologies, the development of new pharmaceuticals stands out in terms of the extent to which the bulk of the work leading to advances is done in laboratories. The advance of surgical practice, on the other hand, proceeds as much in practice as it does in a laboratory setting (Gelijns, 1992). And even regarding pharmaceuticals, a lot is learned in practice about side effects, effective doses, the kinds of diseases and patients for which they are effective. Learning about the efficacy and safety of pharmaceuticals does not stop with the FDA mandated clinical trials.

We think it also useful to understand that randomized control trials, while an important part of the processes of evaluation of practices in modern medicine, is only one part. The cystic fibrosis case shows the important role that physician evaluation plays in the process, as well as communication among physicians and physician groups. On the other hand, the new pharmaceuticals that the physicians worked with in their efforts to deal better with this particular disease had demonstrated safety and efficacy in carefully controlled clinical trials, if usually on diseases other than cystic fibrosis, before they were admitted to the set of tools available to physicians. And we have noted the occasional undertaking of double blind studies to assess the efficacy of treatments that were in experimental use. The use of randomized trials to test the efficacy of evolving practice in treatment of cystic fibrosis probably has been less extensive than it has been in the cases of other human ailments where the population at risk is much greater, for example heart

disease, and various forms of cancer. However, even in these areas, it important to understand the use of random assignment testing as an important part of the system of evaluation of practices, but far from the whole story. We think the same understanding ought to prevail regarding the use of RA in education.

EVALUATIONS OF INNOVATIONS IN BUSINESS PRACTICES?

Business practice is an interesting contrast with medical practice. It is extremely difficult to do effective off-line experimentation that provides much guidance regarding how a new business practice will work on-line. It is, of course, possible to test the technical performance of new hardware systems, like a new computer, or an intra-company telephone system. But it is a rule, not an exception, for a company to have considerable difficulty in integrating the new hardware system into its organization so that it enhances productivity rather than diminishes it. The newspapers are full these days with stories of new software systems carefully designed off-line that just don't work in practice.

One of the authors of this paper has studied two well known innovations in business practice: the multi-divisional form of organizing decision making and authority in multi-product line business (often called the M form) and Quality Circles . In neither case was much off-line research done prior to the first attempts of business to implement these new ways of doing things. In the case of the M form, the basic idea certainly has proved useful, but it has taken a lot of learning and adjustment on the part of companies who adopted the broad idea to get the system running satisfactorily. And Quality Circles

turned out to be largely a fad, whose time now has largely passed (Chandler, 1990; and Cole and Scott eds., 2000).

While we do not see research on business organizations as a model educational research should emulate, we do see some useful themes in the work W. Edward Deming pioneered. The first is the importance of developing and maintaining a variety of fine-grained measures of performance, recognizing that no one measure tells the complete story. This is especially important in education today, given the pressure many schools face to improve students' performances on state-mandated tests. Some teachers, as well as some researchers, question whether improved scores on the high-stakes tests mean that students have acquired skills that will be important in their lives. Developing a variety of indicators of success is critical to judging whether the incentives provided by high stakes testing are constructive.

A second theme is the importance of developing a systematic strategy for monitoring a variety of dimensions of performance and for analyzing the sources of lagging performance. This is an important part of the work of the most successful cystic fibrosis centers. A third theme is that most research that will be valuable to schools striving to be high performance organizations must be done in close collaboration with the schools themselves, rather than in off-site R&D centers.

SERP: A NEW MODEL FOR EDUCATION RESEARCH

Partly in response to the failure of the National Institute of Education and its successor organizations to develop a research program that was both productive and seen as such by the Congress, the National Research Council established a committee in 1996

to consider new designs for educational R&D. In its 1999 report, the committee called for a large-scale program of research development and evaluation, with most of the work embedded in school settings (Donovan et al., 2003).

In 2001 the NRC established a second committee with the mandate to outline just what the new research program should look like. The 2003 report of this committee, *The Strategic Education Research Partnership,* called for a new kind of partnership among research, practitioners, and policy makers. The characteristics of the proposed SERP program included focusing the research agenda on problems of practice, setting the research in schools, bringing to bear a variety of sources of knowledge and expertise, rigorous attention to replication and the systematic building of knowledge, the development of mechanisms to disseminate knowledge effectively, and the importance of coordinating work across sites and projects (Donovan et al., 2003). The first SERP field site is in the Boston Public Schools and focuses on improving literacy instruction in middle schools.

The SERP design is attractive for several reasons. It specifically embraces the idea that skilled practitioners are sources of new ideas for improving teaching and learning. It acknowledges that many innovations appear effective in particular sites at particular times, but it is difficult to transfer the successes to new sites. By paying close attention to interactions among innovations and particular characteristics of the original setting, SERP research projects strive to distill what is needed for successful transfer. By incorporating partnerships among researchers, practitioners, and policy makers in all projects, the designers of SERP hope to keep work focused on critical problems of

schools, expose practitioners to new insights from scholarly disciplines, and develop support for dissemination of effective practices.

Given the ambitious nature of the SERP agenda, it will take considerable time to learn whether it is successful in improving education and in developing and sustaining the political support required for continued funding. However, a necessary condition for success is educating policy makers and potential funding organizations to recognize that that success is much more likely to come in the form of many small ideas for improving practice than from a blockbuster innovation akin to the Salk vaccine. It will be important to develop a fine-grained, multi-dimensional set of indicators of instructional quality and student performance that can be used to document important, but perhaps subtle improvements in teaching and learning.

THE ROLE OF RANDOMIZED EXPERIMENTS: IMPORTANT, CHALLENGING, LIMITED

Randomized experiments are likely to be an especially powerful strategy for increasing knowledge of the consequences of particular educational interventions when the following conditions hold:[9]

1. *The treatment is well defined.* If this is not the case, it is difficult to make inferences from the evaluation results.
2. *The treatment is relatively easy to implement.* Poor implementation is a common explanation for findings of "no effects."

[9] This paragraph draws heavily from Thomas Cook (2002) and from Cook's paper in this issue.

3. *The effects of the treatment are evident in a relatively brief period of time.* Selective attrition that undermines the validity of the random assignment design becomes more severe over time.
4. *The effects of the treatment do not vary among a great many subgroups of the intended population.* The greater the number of interaction effects, the larger must be the experiment to identify these effects.
5. *Feedback effects are modest.* The presence of significant feedback effects means that the consequences of "going to scale" with the intervention might be very different from the consequences of the random assignment evaluation.

There is no shortage of educational interventions that satisfy the criteria above. Indeed, considerable sums of money currently pay for a wide variety of programs that could be evaluated with randomized experiments. They include nutrition programs, after-school programs, and targeted intervention programs designed to address specific problems for specific well defined groups of students. For this reason we applaud the resolution of IES to support randomized evaluations of such programs. We believe that the results, presumably publicized through the What Works Clearinghouse website, can play an *important* role in helping schools that are floundering badly to make some progress.

However, it is particularly *challenging* to develop random assignment evaluations that can provide educators with information useful in moving schools from "making some progress" to becoming high performance organizations that are effective in helping all students to master demanding state-specified learning standards. The reason is that no set of prescribed routines will enable all children to learn. There are many reasons

individual students have difficulty acquiring specific skills. So part of the challenge is identifying the difficulties individual lagging students are experiencing and developing a tailored remediation strategy. As explained above, the results of random assignment evaluations showing that particular instructional techniques are effective in helping some students with particular learning difficulties in particular types of settings would be useful to skilled teachers searching for solutions to particular students' learning problems. However, it would take very large scale, carefully designed experiments to provide such fine-grained results. For that reason we expect that random assignment evaluations will be of only *limited* value in helping schools to become high performing organizations.

Also, convincing students of the need to do their part, even when they see no immediate reason to do so, is as much a part of the work in high performing schools as it is for physicians in higher performing treatment centers for cystic fibrosis. Creating an organization that is skilled at doing this work effectively is much more complex than applying a well defined treatment. It seems unlikely that random assignment evaluations can be very useful in identifying the conditions necessary for doing this work well. To go further, some analysts (e.g., Richard Elmore, 2004) make the compelling argument that adopting a significant number of modular instructional initiatives eventually poses an obstacle to continued school improvement. The reason is that such "programitis" hinders development of a coherent and consistent program of instruction and a school-wide coordinated strategy for monitoring the progress of individual students.

The emphasis on a coherent instructional program and coordinated monitoring and intervention strategies does not mean that there is no room for new initiatives. For example, many schools are engaged in promising efforts to examine student assessment

results regularly and systematically. The idea is that this work can help school faculty to pinpoint aspects of their instruction that is not working well for significant numbers of students.[10] While promising, this initiative does not lend itself to random assignment evaluation for two reasons. First, it is not a modular activity that can make a difference by itself. It can help schools to improve only if it is part of a coherent school improvement strategy. Second, the work is not a well defined activity involving faithfully carrying out a set of prescribed routines. Instead, school faculties need to figure out a way of doing the work that builds on the skills and resources they have and that keeps the activity focused on improvement rather than on the limitations of particular teachers' instruction. While researchers have offered general guidelines for approaching this work, they are far from well developed routines.

In conclusion, we want to call attention to a provision of the No Child Left Behind legislation that will cause problems down the road unless educational research becomes more effective. This provision is that all schools continue to improve their performance over time. We believe that the second model of school improvement described above (developing a coherent instructional program and coordinated monitoring and intervention strategies) has greater potential for continual improvement than the first model. However, even schools with the stable faculties needed to embrace the second model will experience diminishing returns to improvement efforts unless there are advances in the state of the art. While we do not think that the rate of advance in best practice in education will ever be close to the rate of advance achieved in some areas of medicine, including the treatment of cystic fibrosis, we do think that suitably oriented

[10] See, for example, Boudett, City, and Murnane (2005).

educational research can help to advance the state of the art. Random assignment evaluations can and should play an important role in educational research, including the SERP program. However, the role, while important, should be limited. Random assignment evaluations will be no more effective in closing the gap between the best performing schools and schools making “some progress” than they have been in closing the gap between the best performing cystic fibrosis centers and the competent, but less successful centers.

References

Boudett, K.P., E.A. City, and R.J. Murnane, eds, (2005). *Data Wise: a step-by-step guide to using assessment results to improve teaching and learning* (Cambridge: Harvard Education Press).

Briars, D. and L. Resnick (2000). Standards, Assessments—What Else? The Essential Elements of Standards-Based School Improvement. CSE Technical Report 528 (UCLA, National Center for Research on Evaluation, Standards, and Student Testing/University of Pittsburgh, Learning Research and Development Center).

Chandler, A. (1990). *Scale and Scope: The Dynamics of Industrial Capitalism.* Cambridge: Harvard Univ. Press).

Cole, R. and W.R. Scott, eds. (2000). The Quality Movement and Organization theory. Thousand Oaks, Ca.: Sage).

Cook, T. D. (2002). Randomized Experiments In Educational Policy Research: A Critical Examination Of The Reasons The Educational Evaluation Community Has Offered For Not Doing Them. *Educational Evaluation and Policy Analysis*, *2*4(3), 175- 199.

Cook, T.D., F. Habib, M. Phillips, R.A. Settersten, S.C. Shagle and S.M. Degirmencioglu (1999). Comer's School Development Program in Prince George's County: A theory-based evaluation. *American Educational Research Journal* 36(3): 543-597.

Cook, T.D., H.D. Hunt, and R.F. Murphy (2000). Comer's School Development Program in Chicago: A theory-based evaluation. *American Educational Research Journal* 37(2): 535-597.

Donovan, M.S., A.K. Wigdor, and C.E. Snow, eds. (2003). *Strategic Education Research Partnership.* Committee on a Strategic Education Research Partnership. (Washington, DC: The national Academies Press).

Dow, P.B. (1991). *Schoolhouse Politics: Lessons from the Sputnik Era* (Cambridge, MA.: Harvard University Press).

Elmore, R.F. (2004). *School Reform from the Inside Out: Policy, Practice, and Performance* (Cambridge: Harvard Education Press).

Gawande. A. (2004). The Bell Curve. *The New Yorke*r (Dec. 6,).

Gelijns, A. C. (1992). Technology and Health Care in an Era of Limits (Washington, D.C.: National Academies Press).

Hanushek, E.A. (1994). *Making Schools Work: Improving Performance and Controlling Costs* (Washington, D.C.: Brookings).

Hanushek, E.A. (2004). What If There are No Best Practices," *Scottish Journal of Political Economy* 51(May, 2004)2: 156-172.

Hanushek, E.A. and D.D. Kimko (2000). Schooling, Labor Force Quality, and the Growth of Nations. *American Economic Review* 90(5): 1184-1208.

Koretz, D. (2005). Alignment, High Stakes, and the Inflation of Test Scores," in (eds., J. Herman and E. Haertel) *Uses and Misuses of Data in Accountability Testing* (National Society for the Study of Education Yearbook).

Ladd, H. F. and R.P. Walsh (2002). Implementing value-added measures of school effectiveness: getting the incentives right, *Economics of Education Review,* 21(1): 1-17.

Littlewood, J. (2002. The history of the development of cystic fibrosis care. Available at: http://www.cysticfibrosismedicine.com (accessed November 23, 2005).

Murnane, R.J. and D.K. Cohen (1986). Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive. *Harvard Educational Review* 56(February 1986)1: pp. 1-17

Rouse, C.E. (1998). Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program. *Quarterly Journal of Economics* 113 (2):553-602.

Sarason, S. (1971). The Culture of the School and the Problem of Change (Boston: Allyn and Bacon).

Shavelson, R.J. and L. Towne, eds. (2002) *Scientific Research in Education* (Washington, National Academy Press)

Snyder,T.D.,Tan,A.G.,and Hoffman,C.M.(2004). *Digest of Education Statistics 2003,*(NCES 2005 –025).U.S. Department of Education, National Center for Education Statistics.Washington, DC: Government Printing Office.

Warren, D.R. (2004). *To Enforce Education: A History of the Founding Years of the United States Office of Education* (Detroit: Wayne State University Press).