CHAPTER 10

Relying on Evidence

Grover "Russ" Whitehurst

For most of our history the pace of cultural learning was slow, with one generation's experiences not very different from those of their forebears. But that pace is now accelerating so rapidly that the older members of overlapping generations have grown up in circumstances that are outside the experience of younger people. My grandparents lived most of their lives without antibiotics, television, commercial aviation, supermarkets, computers, and almost everything else we consider modern.

Some accounts of the explosion of knowledge and technology in fields such as health care, transportation, and communication credit the inventions in those fields for progress while paying no attention to the processes that made those inventions possible. There is a fundamental sense in which the incandescent light bulb was invented and that shared act of creation is part of the story of cultural evolution. But the translation of that invention into something utilitarian that altered how our planet looks from space was the product of experimentation by Thomas Edison's team at Menlo Park.

Experiments with physical materials such as those carried out by Edison must have had precursors that are prehistoric—imagine a Paleolithic man discovering which rocks produce sparks when struck together. Indeed, nearly all animal species are capable of at least rudimentary forms of trial-and-error learning. But what is simple when a cause that lies within an animal's behavioral repertoire is quickly and reliably followed by an effect the animal values becomes opaque when the effect is multiply determined, probabilistic, and delayed. So whereas an architect of viaducts in ancient Greece would likely have been able to understand the trial-and-error process by which Edison determined that carbon made the best light filament, neither would have been equipped by virtue of their training and experience to understand how to find out whether physical exercise affects health. Nor would either have been well-equipped to determine whether the placement of an advertisement on a page increases sales, or the training of teachers affects their effectiveness, or any other cause-effect relationship that can only be discerned through the application of methods that are capable of extracting a cause-and-effect signal from the noise of weakly probabilistic relationships.

These methods of experimenting to discover complex and probabilistic relationships are recent cultural inventions. Fields that have embraced them have shown rapid progress. Fields that have not, most certainly including education, have stood still.

James Lind's research on treating scurvy among British sailors, carried out in 1747, is generally credited as the first recorded instance of the application of a quasi-experimental design to study an intervention's impact. Lind selected twelve sailors stricken with scurvy, divided them into six groups of two, and gave each group a different dietary treatment. Those given oranges and lemons improved quickly whereas those otherwise treated, e.g., with vinegar, did not. Many aspects of modern experimental design and analysis were missing in Lind's approach. For example, he did

not randomly assign subjects to treatment conditions and did not quantify or test for the significance of the difference in outcomes between groups. Nevertheless, his approach of systematic variation of treatment and observation of results is the foundation of systematic learning through trial and error.

It was 1925 before Ronald Fisher, in his book *Statistical Methods for Research Workers,* explicated the critical role of randomization in assigning subjects to treatments and addressed the need for statistics to deal with error and variability in results. Fisher's methods were developed for use in agricultural and genetic research, but their extension to medicine and the social sciences was straightforward, if not immediate.

The randomized controlled trial of streptomycin for bronchial tuberculosis, begun in Great Britain in 1947, was the first well-implemented and documented randomized trial involving human subjects. The amount of streptomycin was limited so that it was ethically acceptable for the control subjects to be untreated by the drug. Randomization was used instead of the traditional technique of subject assignment to condition by alternation in order to conceal the allocation schedule from those who might bias selection into condition and from those reading the X-rays collected from patients. Procedures and results were meticulously documented. It remains a touchstone for experimental designs in general.

The use of experimental approaches to determine what works has proliferated in health care, industrial production, psychology, and business since their introduction in agriculture and medicine in the first half of the twentieth century. Jim Manzi, who runs a company that carries out experimental trials for business, asserts that many leading companies are relentless experimenters.[1] For example, Google carries out over twelve thousand experiments a year, about 10 percent of which lead to changes in business practices. CapitalOne (a leading credit card company) runs over sixty thousand experiments a year, to which it attributes its growth and

competitive edge. Manzi quotes a manager at Harrah's Casino in Las Vegas as saying that there are three things that could cost a manager his job there: harassing women, stealing from the company, and not having a control group.

During the time in the twentieth century in which other fields were embracing systematic experimentation as the fundamental process for learning what works, education was pre-scientific.

In 1971, the President's Commission on School Finance commissioned the Rand Corporation to review research on what was known about what works in education, reasoning, "The wise expenditure of public funds for education . . . must be based on knowledge of which investments produce results, and which do not." Rand concluded:

> The body of educational research now available leaves much to be desired, at least by comparison with the level of understanding that has been achieved in numerous other fields. . . . Research has found nothing that consistently and unambiguously makes a difference in student outcomes.

Almost thirty years later, in 1999, the National Academies of Science came to essentially the same conclusion:

> One striking fact is that the complex world of education—unlike defense, health care, or industrial production—does not rest on a strong research base. In no other field are personal experience and ideology so frequently relied on to make policy choices, and in no other field is the research base so inadequate and little used.

In comparison to the sad state of affairs that existed heretofore, the twenty-first century has seen an explosion of rigorous and relevant research, providing for the first time a foundation for evidence-based education, which is the use of the best currently available empirical evidence in making policy and practice decisions in

education. Evidence-based education requires a supply of evidence that is relevant to policy and practice, methods for vetting the quality of evidence, processes for synthesis and dissemination of research findings, and demand for evidence among practitioners and policymakers. Substantial progress has occurred in each of these areas, although demand lags behind supply because competitive pressures that create incentives to adopt more effective practices are relatively weak in education compared to many other fields.

The Institute of Education Sciences (IES) within the US Department of Education has played an important role on the supply side of evidence-based education. IES was established in 2002 with the mission of producing rigorous and relevant research to support education policy and practice. Under IES the federal government for the first time established a clear set of priorities for education research funding, articulated standards for research quality that emphasized the validity of causal claims, created a set of processes for research grant competitions that were orderly, predictable, and grounded on systematic peer review, and garnered a budget from Congress that was sufficient to fund nearly all grant applications that were deemed by peer reviewers to be of the highest quality. IES also launched rigorous evaluations of federal education programs, funded university-based doctoral training programs in the education sciences, and directed hundreds of millions of dollars toward the establishment of statewide longitudinal databases of student and teacher records that could be grist to the mill of education research. Through its What Works Clearinghouse (WWC), IES took responsibility for vetting and disseminating findings from studies on the effectiveness of individual programs and practices, as well as publishing practice guides in which consensus panels synthesize recommendations for practitioners from the existing research base.

In part because of the investments and focus of IES, a new community of researchers has arisen that is committed to conducting

rigorous and relevant research on education. The community includes people who had been doing rigorous and relevant research in education for the whole of their careers, but often in isolation from others doing such work in education. It also includes those with established careers in cognate fields such as psychology and economics who shifted their attention to education and significant and growing numbers of newly minted researchers trained in interdisciplinary doctoral programs in education science who are grounded in the normative canons of the social, behavioral, and cognitive sciences.

Progress in Knowledge of What Works

Considerable progress has been made in identifying particular programs and practices that have an impact on student achievement.

The What Works Clearinghouse (WWC) has been in operation for almost a decade with the goal of being the central and trusted source of scientific evidence for what works in education. To date the WWC has conducted systematic reviews of 9,325 research studies, of which 654 were determined to either meet all methodological standards or meet standards with reservations. These rigorous studies enabled the WWC to identify 105 separate interventions with positive effects on student outcomes within the domains of literacy, mathematics, science, student behavior, dropout prevention, early childhood education, English language learners, and students with disabilities. This is a far cry from the conclusions of Rand over forty years ago that research has found nothing that consistently makes a difference in student outcomes.

Progress in Knowledge of What Makes a Difference

The WWC examines the impact of branded interventions that are intended to affect student outcomes. Meanwhile, a different body

of research employs different methods that address the influence of the organization and process by which education is delivered. The preferred method for studies of what works is the carefully planned and executed randomized trial. Typical methods for studies of what makes a difference are epidemiological, i.e., they involve an examination of naturally occurring patterns of association among input and output variables in education. The difference in methodological approach between studies of what works vs. studies of what makes a difference is not so much a matter of choice as of necessity. Whereas discrete interventions lend themselves readily to carefully planned and implemented experiments and quasi-experiments, the broader governance arrangements in which education is delivered and the types of policies decided by district, state, and federal officials are usually difficult to vary experimentally. They often need to be evaluated post-hoc because they are instituted with no thought to their evaluation. As state- and district-level longitudinal education databases have come online in the last decade, the field of education epidemiology has grown by leaps and bounds, both in the volume of published work and in methodological sophistication.

Consider the question of how much teachers, schools, and districts matter to student outcomes. Thinking on this topic through the 1990s was heavily influenced by the landmark 1966 report, *Equality of Educational Opportunity*, by sociologist James Coleman. This was a huge study employing sixty thousand teachers in grade six and beyond in over three thousand schools. The principal finding was that nearly all of the variability in what students achieved was attributable to their socioeconomic background rather than to their schools and teachers. On the subject of teachers, Coleman wrote, "A list of variables concerning such matters as teachers' scores on a vocabulary test, their own level of education, their years of experience, showed little relation to achievement. . . ."

Coleman's insight that schools should be evaluated on their outcomes, not their resources, and his attempt to do so scientifically

FIGURE 1. Comparison of One Standard Deviation of Teacher/Classroom, School, and District Differences on Student Achievement*

(* 0.10 of a student standard deviation = roughly 25% of a school year of learning)

were major advances in education research. But his methods are now understood to have been flawed. All of his analyses were conducted on data that had been aggregated to the school level. For example, the average vocabulary score for all teachers in a school was related to the average test score for all children in a school. We now have available statistical methods that are able to isolate the influence of different levels of the education system on student outcomes. These multilevel approaches generate very different conclusions from those that were the received wisdom of the last century.

Fig. 1 represents the results from a recent study based on statewide data on fourth and fifth grade student achievement in reading and mathematics from North Carolina and Florida for the 2008–2009 school year.[2] It addresses the relative influence of teachers, schools, and classrooms by mapping each to a common unit of a standard deviation of difference in student achievement. One way to think about a standard deviation is that it corresponds to a difference between roughly the thirtieth and seventieth percentiles of performance. Thus, based on the data represented in the figure, students in a classroom with a teacher at the seventieth percentile would be about 0.16 standard deviation ahead of students in the classroom of a teacher at the thirtieth percentile, which is nearly 40 percent of a school year. As indicated in the figure, the impact of schools and districts is less than that of teachers, but still a difference of months of a school year.

We now know that teachers, schools, and districts matter for student achievement, i.e., that demographics and family background are not everything. However, we have not yet translated this understanding into the design and implementation of interventions that can be shown to improve student achievement. It is as if Edison demonstrated that there was variation in the efficacy of light bulb filaments but never got to the point of identifying a practical design. In other words, we know what makes a difference but not what works.

A Look to the Future

Never make predictions, especially about the future.
—Casey Stengel

The transformation of education from a field based on intuition, historical practice, and fad and fancy to a field based on evidence has been thwarted until recently by an inadequate supply of rigorous and relevant research. The supply side of evidence-based education has advanced rapidly in the last decade. But demand is still weak.

The essential question for those interested in the advance of evidence-based education is whether demand is weak because there is something wrong with the research that is being provided or because there is something about the way the field of education is organized that suppresses the market for evidence-based approaches. The problem lies in both areas.

The research and development enterprise in education needs to invest more deeply and systematically in process innovations that will serve the practical needs of school districts and schools. We are unlikely to get dramatically better at educating students until we have a cadre of researchers whose job is to engineer more efficient and effective processes for carrying out the work of schools. Education has an increasingly strong research community, but it

lacks more than a few people trained and employed to improve the workaday processes of delivering education through systematic experimentation.

If Harrah's intends to increase the number of its mid-week customers from Southern California through the mailing of a discount offer, the person responsible for designing the solicitation will lose his job if there isn't a control group. If Google wants to find the display size for search results that generates the most click-throughs on smart phones, it will systematically vary that parameter across different randomly selected groups of users. It has employees and contractors to design the intervention and analyze the results. Its business success depends on finding the best answer.

In contrast, if a large school district wants to redesign its processes for recruiting new teachers by changing when applications are due and offers of employment are made, it would be exceedingly rare if it either had anyone on staff or could find anyone in a local university who would be interested and able to carry out an experiment on the issue. The education research community, which is predominantly comprised of academics, is not interested in such atheoretical, small-bore questions. But these are the types of issues that education administrators address, whereas broad questions of education policy seldom are within their bailiwick. And because the managers of schools and school districts have rarely if ever been supplied with research that directly addresses the decisions they have to make, they have not had the opportunity to develop an appetite for evidence-based education.

Those who have responsibility for the supply of education research, including universities and funding agencies, need to create a pipeline that is primed with practical research of immediate relevance to everyday education decisions. This will require not only a redirection of the goal of much education research but also much better access by the research community to the administrative data at the state, district, and school level on which the research would draw. It will also require a new channel of federal funding

for short-term projects that are of immediate practical significance and that can be reviewed and funded within a few months. This is in contrast to the current modus operandi in which applications for research grants can take a year or more to make their way through the review system and are typically for multiyear projects.

Whether a supply of immediately practical research findings will increase demand for evidence-based education is an empirical question. I expect it would be useful but that its impact would be muted by the same factors that suppress that uptake of evidence that already exists on the impact of broader policies and programs.

The reason that businesses such as Google, Harrah's, and CapitalOne have an appetite for evidence of what works is that avoidable errors in their business decisions go directly to their bottom line, for which managers at many levels and the CEOs are accountable. Google needs to figure out how to maintain its search dominance on mobile devices or others will take its market share. There are lots of casinos in Las Vegas. Harrah's success depends on competing successfully for customers. And so on.

Education, in contrast, is by and large a public monopoly. A recruitment process for new teachers that is much less effective than it might be does not result in the school district losing students or revenue, at least not within a time span or through a series of events that would make the connection discernible. A truly dysfunctional management process may call attention to itself and the administrator responsible for it, but there is no incentive built into the system to experiment with improvements in processes that seem to be OK.

The most powerful way to incentivize evidence-based decision-making in education would be a system of delivery in which schools compete for students and their funding and in which the jobs and compensation of school employees and managers are conditional on their success in attracting and retaining students. Until errors in decision-making have palpable consequences for those responsible

for those decisions, the demand for evidence that will enable better decisions will be weak.

In short, the education research community needs to prime the pump of evidence-based education with a supply of research findings that are of immediate relevance to workaday decision-making, e.g., recruiting tools that enhance the effectiveness of the workforce; ways to increase the productivity of the central office; and differences in the impact of available curriculum materials for particular types of teachers and students. If this is to have more than marginal impact, it will need to be accompanied by a redesign of the delivery of education services such that schools and those who work in them are subject to market forces.

The nation can no longer tolerate vast differences in the quality of its schools and classrooms. Residential geography and quirks of school choice and classroom assignment cannot continue to define the education destiny of individual students. We need for all of our schools to be good enough to do the job that is expected of them. This will require nothing less than a relentless effort to engineer processes that assure the best possible outcomes and that result in continuous improvement. Much of this work will be down in the weeds and the results of any single effort will be incremental. Examined within a short time frame, those results may not look like they are going very far. But it is the accumulation and progression of those incremental improvements that will ultimately be transformational for student achievement and the nation's future.

Evidence-based education has shown progress over the last decade that seemed unimaginable twenty years ago. A foundation is in place for the kind of explosive growth in knowledge and application of what works that has been seen in other fields. Improvements in the relevance of the supply of research and the incentives for educators to make the best possible decisions are the necessary ingredients for the next stage of the reform of our education system.

Notes

1. Jim Manzi, *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society* (New York: Basic Books, 2012).
2. Grover J. "Russ" Whitehurst, Matthew M. Chingos, and Michael R. Gallaher, "Do School Districts Matter?" (Washington, DC: Brookings Institution Press, 2013), http://www.brookings.edu/~/media/research/files/papers/2013/3/27-school-district-impacts-whitehurst/districts_report_03252013_web.pdf.