LIMITATIONS OF EXPERIMENTS IN EDUCATION RESEARCH

Diane Whitmore Schanzenbach
Institute for Policy Research
Northwestern University
Evanston, IL 60208
dws@northwestern.edu

Abstract

Research based on randomized experiments (along with high-quality quasi-experiments) has gained traction in education circles in recent years. There is little doubt this has been driven in large part by the shift in research funding strategy by the Department of Education's Institute of Education Sciences under Grover Whitehurst's lead, described in more detail in his article in this issue.

1. INTRODUCTION

The field of education policy has been a relatively late convert to experimental evaluations, following far behind the lead set in areas such as job training (Lalonde 1986) and international development (Duflo 2006). One might wonder why it took so long for education research to adopt these methods. In part it might be because historically the federal government has played a relatively small role in education. In fact, two of the most prominent educational experiments—the Perry Preschool Program and Tennessee's Project STAR (Student/Teacher Achievement Ratio)—were implemented without federal funding. Another source of resistance to experiments may be the perception within the field of education that other research methods are just as good (or better).

I think the recent emphasis in education on the experimental evaluation of programs is a healthy trend. Education policy is vitally important to workforce development and the future well-being of our nation, yet there is a paucity of rigorous program evaluation on which to base those policies. To fill this gap, we should probably be conducting many more experiments and rigorous evaluations.

Nonetheless, it is important to note that experiments have significant limitations. This article provides an overview of some of these limitations. There are also good pro-experiment responses to many of the criticisms I outline below. I hope this article helps us better understand the limitations of experiments so we have a better understanding of what generalizations we can and cannot make from experimental results, and we can learn from the limitations in order to improve the design of future experiments.

2. THE EXPERIMENTAL IDEAL

In discussing the limitations of experiments, I start from the assumption that we are interested in uncovering the causal impact of programs.[1] In other words, we would like to estimate the differences in outcomes between a group of students that was exposed to a program and an equivalent group that was not exposed. For example, what happens to the distribution of test scores when high-stakes accountability is enacted? Does the adoption of computer-based instruction improve student outcomes? The challenge in estimating a policy's impact is generally establishing a valid counterfactual. In other words, what would have been this student's outcome if she had not been subjected to the policy? As Grover Whitehurst describes in his article in this issue, when the

1. Cook (2002) has an interesting overview of objections on philosophical grounds to experiments about the nature of causality.

goal is to evaluate the impact of a particular policy, randomized controlled trials are the gold standard of research.

When an experiment has not been conducted, other quasi-experimental tools of policy evaluation, such as regression discontinuity, difference in difference, or propensity score matching, must be used. Even when using these other methods, the logic of experiments underlies the approach to answering causal questions. We teach budding researchers to ask what experiment they would ideally like to design in order to answer a particular research question and to choose the quasi-experimental research method that best approximates the experimental ideal. Indeed, we judge the quality of nonexperimental methods by how closely they approximate experiments.

Today there are numerous examples of high-quality nonexperimental program evaluations in education. Unfortunately these nonexperimental studies, which are often extremely clever, are typically too complicated to be of widespread use to policy makers. (Try to explain an instrumental variables paper to a school principal sometime.) This further underlines the importance of real randomized experiments. Not only do experiments credibly establish a counterfactual so that average treatment effects can be easily calculated, but they are straightforward to explain to practitioners without advanced methodological training.

3. LIMITATIONS OF EXPERIMENTS

Even though there are many important advantages to experimental evaluations, it is also important to understand their limitations. In particular, they cannot answer every question. I next describe some of the major limitations to experiments.

Feasibility of Implementing an Experiment

Some of the most important questions in educational policy cannot feasibly be evaluated via experiments, even though one could in theory design an experiment to test a particular question. For example, many personnel issues are next to impossible to test using a randomized experiment. It is difficult to imagine states or school districts agreeing to the random assignment of collective bargaining rights or randomly laying off teachers. Similarly, while policies aimed at failing schools, such as closures or assigning them to turnaround specialists, could in theory be randomly assigned across the risk set, in practice it would be difficult to find a superintendent who is willing to go along with such a plan. In situations like these—in which an experiment could in theory be designed but is practically infeasible—quasi-experimental methods can often be adopted to tease out the parameter of interest. For example, Hoxby (1996) and Lovenheim (2009) leverage differences in timing of the

introduction of collective bargaining to isolate the impact of teachers' unionization on outcomes.

Complex Treatments

Experiments are best suited to test relatively straightforward interventions, such as implementing a particular curriculum. They are generally limited to testing the impact of pulling a single lever at a time, whether that be one program or a given bundle of programs. Educational theory, on the other hand, is usually built around larger systems in which many reforms are implemented at once and have important synergies. In his book *So Much Reform, So Little Change*, Charles Payne (2008) criticizes urban school improvement strategies and argues that successful strategies must do more than implement a string of disconnected programs. On the other hand, experiments are good at testing individual programs or bundles of programs. Without additional information about the fidelity of implementation, though, an experiment cannot determine whether the program or programs have been implemented well.

John Easton, the current director of the Institute of Education Sciences (IES), makes a related criticism in his 2010 presidential address at the American Educational Research Association meeting: "One theme I hear regularly at IES and elsewhere is that many of our disappointing evaluation results come about because programs are not implemented correctly or with fidelity" (Easton 2010).

There are several important lessons to take away from this. First, evaluators must pay closer attention to how policies are implemented in practice. This is akin to the medical concept of measuring the impact of a drug "as prescribed" versus "as taken." If the treatment ideally calls for a very complex set of related interventions but in practice these are implemented in a less than complete manner, this is important for policy makers to understand. It is possible that a program, like a medical treatment, could be effective if implemented properly but is not effective on average when it is implemented in practice. It may be the case that by collecting better measures on the fidelity of implementation—especially at the pilot program stage—programs could be streamlined so they could be implemented with more fidelity, or implementation benchmarks could be established. If, however, a policy works only under narrow circumstances when a complex implementation process is closely adhered to—and less than ideal implementation yields no results—this suggests that the policy may not be a good candidate for widespread adoption in the real world, even if its impacts would be large with ideal implementation.

Overall, the notion that education policies are most effective when they are more comprehensive and complex does not necessarily invalidate the use of experiments. If anything, it points to the need for more experimentation and

study so policy makers can understand under what conditions the programs are most effective. However, once evaluators come up with an estimated impact, it is important to keep in mind under what circumstances this impact was measured (e.g., under random assignment of students to classrooms or in schools with high levels of social capital among the faculty) and how well this is likely to be replicated in real life. It is possible that the experiment may hold so much else constant that its findings will be of much more limited use under less controlled circumstances.

Time and Timing

A related issue is that experiments are often conducted on newly implemented programs. This may be the only way to make the evaluation feasible when the intervention is not oversubscribed. For example, administrators are more apt to agree to random assignment of treatment during a one- or two-year pilot study that precedes a larger-scale rollout. It seems somehow fairer to the control group to merely delay the implementation for a year or two.

This is problematic if the program increases in effectiveness over time; the pilot program may look quite different from the mature program. This appears to be true, for example, in charter schools. There are substantial start-up costs to opening a school: policies must be set in a wide variety of areas, the school's culture must be established, in some cases teachers must become more familiar with the school's curriculum and instructional practices, and so on. As a result, the treatment effect in year 1 may be substantially different from the treatment effect in year 4.

In addition, in some cases the impacts are different for students of different ages and/or build over time. These factors suggest that the longer-run effectiveness of a school may be understated by an evaluation conducted in the first year or two. Take the case of a new charter school that admits children via randomized lottery and opens with grades K–3, with a plan to add an additional grade at the top each year it is open until it is a K–6 school. In this district students are routinely administered standardized tests starting in third grade. An experimental evaluation of this school in its first year would rely primarily on the third graders' test scores. But the third-grade cohort may be very different from the kindergarten cohort. For example, students who want to transfer to a newly opened charter school in third grade are probably more likely to be extremely dissatisfied with their current school or may be more highly mobile than those who want to enroll at a more natural entry point like kindergarten or the transition to middle school. Driven in part by the differences in underlying student characteristics, the impact of the school may differ markedly between the kindergarten and third-grade cohorts in a given year. In addition, by design the third-grade curriculum could be more effective

if the student has been exposed to the same curriculum in earlier grades—that is, the impacts could cumulate over time. As a result, the effects observed on the original kindergarten cohort when it finally reaches third grade may be substantially different from those observed for the third graders enrolled in the first year a school was open.

Such a situation presents real challenges to the evaluator. For an experimental evaluation in a school's first year, there is no way to tell the difference between the temporary null results that will likely mature to sizable impacts within a few years and null results from an ineffective program that will never yield positive impacts. In this case it seems that the ultimate evaluation of the program's promise has to be based on some combination of the measured impact along with other observations on processes, strength of leadership, and so on. The problems are exacerbated when parents and especially funders want to see measurable positive impacts immediately and are unwilling or unable to take a wait-and-see approach.

This is also problematic if the short-term measurable impacts do not proxy the long-run impacts very well. In general when we intervene with children we are ultimately interested in improving long-run outcomes such as lifetime wages or other measures of well-being. Since we generally do not want to wait twenty or more years for wage data to be available, we look for impacts on short-run measures such as test scores that are thought to be good proxies for the longer-term outcomes. Sometimes, though, the long-term outcomes are not well predicted by short-run measures. For example, Krueger and Whitmore (2001) find that test score impacts from being randomly assigned to a small class fade out substantially in the years immediately after the intervention's termination. The impact on college going (as proxied by whether the student took a college entrance exam) was much larger. As a result, the long-term effectiveness of the intervention is larger when the long-run outcomes are actually observed than when they are weakly proxied by test scores. Similar disparities between actual long-run outcomes and short-run proxies are also observed in studies of the Head Start preschool program (Currie and Thomas 1995; Deming 2009).

In addition, placing high stakes on short-run outcomes may alter the relationship between the proxy measures and the long-run outcomes. A widely used rule of thumb from Neal and Johnson's (1996) work is that a 1 standard deviation improvement in childhood test scores increases adult wages by 10 percent. Today, though, under high-stakes accountability systems, schools face pressure to improve test scores without a parallel improvement in students' (unobserved) skill levels. As a result, down the road we may find that test score improvements do not translate into the improvements in long-run outcomes that we would have expected.

A related criticism is that experiments are useless to policy makers in the short run. By the time an experiment is designed, implemented, and evaluated, it is often true that the policy debate has moved ahead and the results are no longer of direct policy interest. But what type of research is useful under such circumstances? Certainly not a correlational study, which, although it can be executed more quickly than an experiment, will fail to yield an estimate of the policy's causal impact. Unfortunately policy makers often turn to such studies, then find themselves disappointed with real-life results that do not live up to whatever promises were made based on poorly designed research that yielded biased estimates of the program effects.

This suggests to me that we should be conducting not less but *more* experimental (and quasi-experimental) research and setting our sights more broadly to include questions that are not of immediate policy relevance but that may inform the debate in the future. For example, early work on school accountability systems in Florida by David Figlio (Figlio and Lucas 2004; Figlio and Rouse 2006; Figlio and Getzler 2006), in Chicago by Brian Jacob (Jacob 2003, 2005), and in North Carolina by Thomas Kane and Douglas Staiger (Kane and Staiger 2002) presaged much of the later debate surrounding No Child Left Behind. I also note that we have learned more from the STAR experiment than just the impact of class size. Researchers have used the data to tease out estimates of teacher effects (Chetty et al. 2011), peer effects (Cascio and Schanzenbach 2007; Graham 2008; Sojourner 2011), teacher-student match quality (Dee 2004), and more. This suggests that the field would be well served by having more large-scale random experiments and making the data available to other researchers. I fear the current research funding environment has shifted too much toward funding work of immediate policy interest and does not facilitate the types of basic research that may be policy relevant in the future.

External Validity

A limitation of both experiments and well-identified quasi-experiments is whether the estimated impact would be similar if the program were replicated in another location, at a different time, or targeting a different group of students. Researchers often do little or nothing to address this point and should likely do more. One straightforward approach is to report a comparison of the experiment's control group to the population of interest and reweight the sample to be representative of that population. This again suggests the need for more experimentation across a wider variety of settings. This approach was taken in the evaluation of welfare-to-work programs in which experiments were conducted across twenty separate programs (Michalopoulos, Schwartz, and Adams-Ciardullo 2001). Some researchers have used these experiments

in conjunction with structural modeling to predict impacts of other policies under consideration (e.g., Pepper 2003). Education policy could likely benefit from a similar approach.

Another limitation of experiments is that they are generally best at uncovering partial equilibrium effects. The impacts can be quite different when parents, teachers, and students have a chance to optimize their behavior in light of the program. Pop-Eleches and Urquiola (2011) describe this well in their work on attending a higher-quality high school in Romania. In particular they find that behavioral responses differ both across time and by the magnitude of the program. One important conclusion they draw is that it is important to analyze the impact of educational interventions not only on student outcomes but also on the behavior of all those involved.

Black Box

Another limitation of experiments is that although they are good at testing the impact of a policy, they provide little insight into what makes the policy work.

The Project STAR class size reduction experiment is a good example here. While the experiment provided an unbiased estimate of the impact of reducing class size (holding other aspects like teacher quality constant), there is little evidence about what it is about smaller classes that led to improved outcomes. I once saw a presentation that posited the mechanism as improved oxygen quality in the classroom. That is, having fewer students in the classroom means there are fewer bodies exhaling carbon dioxide. Too high a concentration of carbon dioxide in the air can lead to loss of concentration and headaches. If this were indeed the mechanism for the impacts (and I do not think it was), the same treatment could likely be administered in a much less expensive manner such as improving ventilation systems, introducing houseplants into the room, or opening a window.

Hawthorne Effects

Another limitation of experiments is that it is possible that the experience of being observed may change one's behavior—so-called Hawthorne effects. For example, participants may exert extra effort because they know their outcomes will be measured. As a result, it may be this extra effort and not the underlying program being studied that affects student outcomes. If these effects are at play, when the program is implemented at a larger scale and is not being closely observed by researchers, the actual impacts may bear little resemblance to those projected by the pilot study. It may be the case that quasi-experimental approaches are less susceptible to Hawthorne effects.

Cost

Experimental evaluations can be expensive to implement well. Researchers must collect a wide variety of mediating and outcome variables (and, if they are fortunate, baseline measures as well). It is sometimes expensive to follow the control group, which may become geographically dispersed over time or may be less likely to cooperate in the research process. The costs of experts' time and incentives for participants also threaten to add up quickly. Given a tight budget constraint, sometimes the best approach may be to run a relatively small experimental study. Unfortunately, having a small sample size can lead to underpowered experiments. When an experiment is underpowered, of course, researchers are more likely to fail to reject a null finding, and as a result some potentially important interventions may be overlooked. This problem is exacerbated in the context of educational interventions, when the intervention is often implemented at the class or school level, resulting in larger standard errors than if it were implemented at the individual level.

Under some circumstances, experimental evaluations can be executed without much additional cost. For example, many of the recent studies of school choice primarily leverage administrative databases and do not require additional expenses to run the lotteries because schools either choose to or are required to allocate spots via lottery anyway. Although evaluations conducted in this manner are less complete than those with a comprehensive qualitative component and those that follow students who exit the school system, they are also less expensive. There is some optimal trade-off given the budget constraint between conducting a larger number of these cheaper experiments, based primarily on administrative data sets and a smaller number of more detail-rich evaluations.

When considering the costs of experiments, it is important to ask to what the cost is being compared. Demonstration or implementation studies themselves often involve substantial costs. If such studies are being planned anyway, it may be relatively inexpensive to layer on top of these an experiment or other research design that will allow researchers to evaluate the program's impact. Furthermore, it is important to recall that it is often quite costly to not have adequate information about a program's effectiveness. In its absence, promising programs may be overlooked, or money may be wasted on administering ineffective programs.

Ethics

Sometimes people object to experiments because they are unwilling to withhold treatment from any individuals serving as controls. Budgetary constraints often overrule such objections in practice. As long as there are not enough resources to treat all those who are interested, some individuals will necessarily

be left out. At other times the objections are assuaged by a guarantee that the control group will get treatment after the time frame of the study has passed. For example, the randomized evaluation of the Big Brothers Big Sisters program divided study participants by placing the control group on the regular waiting list (which was generally an eighteen-month wait) and jumping the treatment group to the top of the queue (Tierney and Grossman 1995). As a result, the impact evaluation was measurable only during the eighteen months before the members of the control group were assigned their own mentors. This necessarily prevents the measurement of long-term impacts but importantly allows the control group to participate in the program.

Other objections are based on the opinion that some potential participants need the program more than others and that an observer can accurately assess this need. An experimental approach can still be used in this case by assigning potential participants to priority groups and then randomly allocating treatment within those groups. This approach is often used in urban charter schools, where some students—typically those who either live in a targeted neighborhood or already attend a failing school—are given top enrollment priority. Next priority may go to, for example, students with siblings who attend the school, and so on. Say there are one hundred open spots in the school and seventy-five students in each of the top, second, and third highest priority groups. In this case all the students in the top group would be awarded admission, and the remaining twenty-five slots would be randomly allocated among the second priority group so that each student in the second group had a one in three chance of getting a slot. In this example, no students in the lowest priority group would be offered admission. One limitation of this approach is that in this case an evaluation will measure only the impact on the second priority group. An experimental evaluation requires there to be both lottery winners and lottery losers within a group. Since there are no losers in group 1, there is no control group with which to compare outcomes. Similarly, since no one in group 3 received treatment, there is no way to measure a program impact on that group. This unfortunately limits the usefulness of the exercise, as we cannot infer from the well-measured impact on group 2 what the impact would have been on groups 1 and 3.

It is also worth noting that many ethical objections rest on the assumption that the treatment actually improves outcomes. In fact, the reason an experimental evaluation is being considered in the first place is to determine whether and to what extent this is the case. There are famous counterexamples—hormone replacement therapy and "scared straight" juvenile delinquency prevention programs—that appeared to actually harm participants. These

examples should remind us to approach program evaluation with a healthy dose of skepticism.

Violations of Experimental Assumptions

The validity of an experiment hangs or falls on whether random assignment was implemented correctly, whether treatment was actually higher for the treatment group, and whether the survey response rates were both high enough and the same across treatment and control groups. If any of these conditions fails, the program's impact can no longer be measured by a simple comparison of treatment and control group means. Skeptics sometimes argue that this possibility of failure is a reason to not implement an experiment in the first place.

The mere possibility of failure is not reason enough to give up on the experimental approach (especially if it would be replaced with a research design that is not capable of estimating causal effects). In addition, there are steps researchers can take to minimize this risk. For example, sometimes researchers have baseline characteristics from application data or administrative records. It is straightforward to check for covariate balance on these baseline characteristics prior to implementing the experiment. We know that on occasion just by random chance one will draw a sample that is not balanced. If researchers can determine this upfront, they can determine whether the imbalance is so stark that they want to modify their approach or even draw another sample. Similarly, researchers can monitor follow-up response rates, and interviewing efforts can be intensified if the rates are either too low or imbalanced.

The Temptation to Data Mine

Another limitation of experiments is that it is perhaps too easy to mine the data. If one slices and dices the data in enough ways, there is a good chance that some spurious results will emerge. This is a great temptation to researchers, especially if they are facing pressure from funders who have a stake in the results. Here, too, there are ways to minimize the problem. Whenever possible, researchers should specify their hypotheses prior to analyzing data and make certain that these are guided by theory and/or prior related research. Researchers should also show sensitivity tests in order to give readers more detailed information about how robust the findings are.

Recent Critiques from Economists

Several prominent and highly regarded economists have recently argued against what might be characterized as an overemphasis on experiments (and program evaluation) in the field of economics. Deaton (2009), referring in particular to the trend toward experiments in the field of development economics, argues that experiments should hold no special priority in research

and that sometimes the parameters that are uncovered are of little or no use to policy makers. For example, in the urban charter school lottery described above, using the experiment one can identify the impact only on the group of students who had some chance of winning and some chance of losing. It takes further assumptions or structure to then predict the potential impact on other groups of students. Deaton further argues that experiments as they are currently being conducted are unlikely to yield insight into larger scientific questions regarding economic development. He is more enthusiastic about experiments that are designed to measure mechanisms that can be used to inform economic models other than those that are primarily concerned with program evaluation.

Neal (2010) focuses his critique directly on the literature on the economics of education and argues that today there is an overemphasis on program evaluation. He contends that economists are better suited to stick closer to their comparative advantage of modeling theory and designing mechanisms and should leave more of the experimental design and implementation to other social scientists. I think these critiques will likely push the field forward in important ways in the coming years. As far as I can tell, though, none of these critics would want us to revert to the style of research that was common in education prior to the recent emphasis on experiments.

4. CONCLUSIONS

Given the limitations outlined above, what can we conclude about the role of experiments in educational research? I argue that they are an important tool but of course are not the only worthwhile research approach. Experiments are best at addressing simple causal questions. They are less useful when the program to be evaluated is more difficult to implement—for example, if it is labor intensive or high levels of skill are needed for proper implementation. When drawing conclusions from experiments, one must take care to understand the limitations, especially those regarding external validity.

I am grateful to Lisa Barrow, Tom Cook, David Figlio, and Kirabo Jackson for helpful discussions. All errors are my own.

REFERENCES

Cascio, Elizabeth, and Diane Whitmore Schanzenbach. 2007. First in the class? Age and the education production function. NBER Working Paper No. 13663.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *Quarterly Journal of Economics* 126(4): 1593–1660.

Cook, Thomas 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* 24(3): 175–99.

Currie, Janet, and Duncan Thomas. 1995. Does Head Start make a difference? *American Economic Review* 85(3): 341–63.

Deaton, Angus. 2009. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. NBER Working Paper No. 14690.

Dee, Thomas S. 2004. Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics* 86(1): 195–210.

Deming, David. 2009. Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics* 1(3): 111–34.

Duflo, Esther. 2006. Field experiments in development economics. In *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, Vol. 2, edited by Richard Blundell, Whitney K. Newey, and Torsten Persson, pp. 322–48. Cambridge, UK: Cambridge University Press.

Easton, John Q. 2010. Out of the tower, into the schools: How new IES goals will reshape researcher roles. Presidential address to the American Educational Research Association Conference, Denver, CO, May.

Figlio, David N., and Lawrence Getzler. 2006. Accountability, ability and disability: Gaming the system? In *Improving school accountability (advances in applied microeconomics)*, Vol. 14, edited by Timothy J. Gronberg and Dennis W. Jansen, pp. 35–49. Bingley, UK: Emerald Group Publishing.

Figlio, David N., and Maurice E. Lucas. 2004. What's in a grade? School report cards and the housing market. *American Economic Review* 94(3): 591–604.

Figlio, David N., and Cecilia Elena Rouse. 2006. Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics* 90(1–2): 239–55.

Graham, Bryan S. 2008. Identifying social interactions through conditional variance restrictions. *Econometrica* 76(3): 643–60.

Hoxby, Caroline Minter. 1996. How teachers' unions affect education production. *Quarterly Journal of Economics* 111(3): 671–718.

Jacob, Brian A. 2003. A closer look at achievement gains under high-stakes testing in Chicago. In *No Child Left Behind? The politics and practice of school accountability*, edited by Paul Peterson and Martin West, pp. 269–91. Washington, DC: Brookings Institution.

Jacob, Brian A. 2005. Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics* 89(5–6): 761–96.

Kane, Thomas J., and Douglas O. Staiger. 2002. The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16(4): 91–114.

Krueger, Alan B., and Diane M. Whitmore. 2001. The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *Economic Journal* 111(468): 1–28.

Lalonde, Robert J. 1986. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76(4): 604–20.

Lovenheim, Michael F. 2009. The effect of teachers' unions on educational production: Evidence from union election certifications in three midwestern states. *Journal of Labor Economics* 27(4): 525–87.

Michalopoulos, Charles, Christine Schwartz, and Diana Adams-Ciardullo. 2001. *What works best for whom: Impacts of 20 welfare-to-work programs by subgroup*. Report prepared for the U.S. Department of Health and Human Services, Administration for Children and Families. New York: MDRC.

Neal, Derek. 2010. Education policy as a mechanism design problem. President's address to the Midwest Economics Association Annual Meeting, Chicago, March.

Neal, Derek, and William R. Johnson. 1996. The role of pre-market factors in black-white wage differences. *Journal of Political Economy* 104(5): 869–95.

Payne, Charles M. 2008. *So much reform, so little change: The persistence of failure in urban schools*. Boston: Harvard Education Press.

Pepper, John. 2003. Using experiments to evaluate performance standards: What do welfare-to-work demonstrations reveal to welfare reformers? *Journal of Human Resources* 38(4): 860–80.

Pop-Eleches, Cristian, and Miguel Urquiola. 2011. Going to a better school: Effects and behavioral responses. NBER Working Paper No. 16886.

Sojourner, Aaron. 2011. Identification of peer effects with missing peer data: Evidence from Project STAR. IZA Discussion Paper No. 5432.

Tierney, Joseph P., and Jean Baldwin Grossman. 1995. *Making a difference: An impact study of Big Brothers/Big Sisters*. Philadelphia: Public/Private Ventures.