JULY 2017

HOWARD S. BLOOM

# THE SEARCH FOR SIMPLE SOLUTIONS TO PRACTICAL PROBLEMS

## Reflections from a Career in Evaluation Research

*Howard Bloom is retiring this summer after serving as MDRC's chief social scientist for 18 years. This followed 21 years of teaching research methods, program evaluation, and applied statistics at Harvard University and New York University. At MDRC, Howard has led the development of experimental and quasi-experimental methods for estimating program impacts and worked closely with staff members to build these methods into their research; he intends to continue to work with MDRC on special projects for the foreseeable future.*

*With Howard's retirement approaching, his MDRC colleagues asked him to reflect on his career. What follows is his response.*

Throughout my career I always have derived the greatest satisfaction from exploring simple solutions to practical methodological problems in the context of substantive empirical research. And along the way, I learned some important lessons. For example:

### THE POWER OF SIMPLE HYPOTHESES

As a graduate student in political economy and government at Harvard during the early 1970s, I became interested in the relationship between macroeconomic conditions and voting behavior. At the time, there was debate about this issue between Gerald Kramer and George Stigler. Prof. Kramer began the debate with an analysis of national time-series data that demonstrated that an increase in real national income the year before a U.S. Congressional election increased the vote for the party of the incumbent president, while a corresponding

mdrc
BUILDING KNOWLEDGE TO IMPROVE SOCIAL POLICY

decrease in income reduced this vote.[1] Prof. Stigler used similar data to show that the average change in real national income during the *two years* before a U.S. Congressional election had no influence on the subsequent vote.[2] As I thought about these findings and the existing literature on voting behavior, it seemed to me that voters were probably "nearsighted" (responding most strongly to recent changes) and "asymmetric" (responding most strongly to negative changes). I thus estimated a model from national time-series data that reflected these hypotheses by specifying separate political effects of changes in national income during the first and second years before each election and separate political effects of rising and falling income. The results of this analysis strongly supported both hypotheses.[3] In addition, they explained how Prof. Stigler's null findings, produced by combining two prior years of income change into a single explanatory variable, masked the effects of recent economic changes found by Prof. Kramer. Furthermore, the results extended Prof. Kramer's findings by identifying the difference between the major political costs of economic downturns and the modest political benefits of economic upturns.

## THE POWER OF SIMPLE GRAPHS

As a visiting scholar at the Congressional Budget Office and the National Commission for Employment Policy in the early 1980s, my task was to study the impacts of the Comprehensive Employment and Training Act (CETA) — the federal job-training program at the time. This study was designed to use a national longitudinal data set for program participants and nonexperimental comparison group members called the Continuous Longitudinal Manpower Survey, or CLMS. In planning for the analysis, I read — among other things — a well-known paper by Orley Ashenfelter, which used an "autoregressive model" (with individuals' past values of the outcome as covariates) to estimate impacts of the previous federal job-training program, the Manpower Development Training Act (MDTA), from a national longitudinal data set.[4] This analysis found that MDTA program impacts were initially positive for men and women but decayed markedly thereafter for men — a pattern that was later assumed by many cost-benefit analyses. As background, Table 1 in the paper reported mean annual earnings for treatment group members and comparison group members, for each of six baseline years and five follow-up years. Out of curiosity, I graphed these points to see what they implied. Much to my surprise, they demonstrated a large program-induced earnings gain with *no sign* of decay over time. During the following months, I discovered that the difference between results from the graphs and those from autoregressive models was due to "time-varying bias" in the autoregressive models.[5] In addition, I discovered that a linear comparative interrupted time-series analysis (comparing deviations from baseline trends) of aggregate annual data (which reflects what could be seen in the graphs) produces an impact estimate that is identical to that from a much more complex longitudinal model of micro-data with separate baseline intercepts and slopes for each sample member.[6] This lent further credence to the graphical findings. I later learned in a letter from Donald Campbell that he too had noticed the inconsistency between Prof. Ashenfelter's findings and graphs of his aggregated data.[7]

1 Kramer (1971).

2 Stigler (1973).

3 Bloom and Price (1975).

4 Ashenfelter (1978).

5 Bloom (1984b).

6 Bloom (1984b).

7 See Cook and Campbell (1979), p. 229.

## THE POWER OF SIMPLE ANALYTIC EXPRESSIONS

In 1983, my wife, Susan Bloom, and I were asked by Dennis Carey, secretary of labor for the State of Delaware (whom I had met while teaching program evaluation in a Kennedy School executive program), to conduct a randomized impact study of Retraining Delaware's Dislocated Workers, a small pilot program being planned by the state. During initial discussions about the program and its evaluation, it became clear that some persons randomized to the program would end up not participating in it. Thus I began to think about how to treat these "no-shows" in an experimental impact analysis. As I discovered, the solution to this problem was surprisingly simple. Based on one assumption — that no-shows experience no program impact — I developed an estimator of the mean impact of program participation.[8] The estimator merely *divides* an experimental estimate of the mean impact of program assignment by the proportion of treatment group members who participated in the program. This straightforward division became known as the "Bloom adjustment," which I later extended to account for control-group "cross-overs" who receive program services.[9] The approach was developed independently and formalized in the context of instrumental variables analysis by Angrist, Imbens, and Rubin and is now used widely.[10]

## THE POWER OF SIMPLE RESEARCH DESIGN PARAMETERS

In 1986 I became co-principal investigator of the National Job Training Partnership Act (JTPA) Study with Judy Gueron and Larry Orr. This study, which was the first major randomized impact evaluation of a federal job training program, was created in response to widespread dissatisfaction with the ambiguous results of previous nonexperimental evaluations of such programs. From the outset of the study, it was clear that site recruitment would be difficult because potential sites (local JTPA service delivery areas, or SDAs) had serious misgivings about participating and little or no incentive to do so (participation was neither mandated nor compensated). Although we initially planned for a balanced 50/50 treatment and control group allocation, early site reconnaissance indicated strong resistance to having so many control group members. Thus I was asked to determine how much we could reduce our control group proportion (holding the total sample size constant) without undermining our statistical power. Although the literature on statistical power analysis for experimental design was well established at the time, it was fairly technical and somewhat opaque to non-statisticians, and it did not provide a simple expression for determining power. In my search for a more convenient and transparent alternative, I came upon a simple expression for the standard error of an experimental impact estimator,[11] which I modified to focus on the influence of sample allocation. In addition, I defined a "minimum detectable effect" (a concept that meant different things to different researchers at the time)[12] to be the smallest true impact that could be detected at a given level of statistical significance with a given level of statistical power and discovered that this parameter was just a multiple

8 Bloom (1984a).

9 Bloom et al. (1997).

10 Angrist, Imbens, and Rubin (1996).

11 Pitcher (1979).

12 For example, sometimes this term was defined as the minimum value of an impact estimate that would be statistically significant and sometimes it was defined as the minimum value of a true impact that would make an intervention cost effective (Pitcher, 1979).

of the standard error of an impact estimator. The resulting expression made it easy to see that the minimum detectable effect increases very little as one departs from a 50/50 balanced sample until the allocation becomes highly imbalanced. Based on this finding, we reduced our JTPA Study control group proportion to one-third. This facilitated site recruitment (which was still a hard sell) while increasing our minimum detectable earnings effect by only 7 percent. Some years later I wrote a short paper on minimum detectable effects for applied researchers and graduate students.[13] Since then, this parameter has become the "coin of the realm" for assessing the precision of sample designs for evaluation research.

## THE POWER OF CLUSTER RANDOM ASSIGNMENT

In the mid-1990s, I was asked to help MDRC develop a research design for evaluating Jobs-Plus, a saturation-level place-based demonstration program intended to increase employment and earnings for residents of selected public housing developments in six U.S. cities. Given the nature of the intervention, it was not possible to estimate Jobs-Plus impacts with a conventional individual-level random assignment experiment. However we were able to randomly select from a matched pair or triplet of public housing *developments* in each city one treatment development and one or two control developments. At the time, we did not know how to assess the statistical precision of this design. But our instincts told us that we had far too few randomized developments (clusters) to support much precision. So we decided to use a comparative interrupted time-series analysis instead.[14] A couple of years after the Jobs-Plus study began, MDRC convened a multidisciplinary meeting of academic researchers (representing disciplines ranging from history to astrophysics) to explore alternative paradigms for studying causal relationships. At this meeting, Steve Raudenbush spoke about cluster random assignment designs,[15] which provided me with an introduction to their statistical properties. In 2003, Bob Granger, who had organized the MDRC meeting and later moved to the William T. Grant Foundation, invited Steve and me to develop a project that, among other things, helped to bring cluster randomization into evaluation research.[16] We all felt this was important, because although cluster-randomized trials had the potential to greatly extend the range of interventions that could be evaluated rigorously, these designs were poorly understood by applied researchers (some did not even believe they were true experimental designs), and when these designs had been used in the past, they often were analyzed incorrectly (as if individual sample members had been randomized). Much of the work for this project involved teaching others how to design and analyze cluster-randomized studies, the culmination of which was a three-day workshop conducted by Steve and me at the University of Michigan (with assistance from Jessaca Spybrook and Andres Martinez) for a large and diverse group of prominent researchers and research funders.

## THE POWER OF EMPIRICAL INFORMATION ABOUT RESEARCH DESIGN PARAMETERS

As I learned, the main weakness of cluster random assignment designs is the fact that their precision

13 Bloom (1995).

14 Bloom and Riccio (2005).

15 Raudenbush (1997).

16 For example, Bloom (2005) and Raudenbush (1997).

typically depends more on the number of clusters randomized (which is often quite limited) than it does on the number of sample members per cluster (which is often less limited). The key determinant of these two factors' relative influence on precision is a parameter called the intraclass correlation, or ICC. Two other determinants of precision are the predictive power or R-squared values of cluster-level and individual-level covariates. Thus to assess the precision of a proposed cluster randomized design requires a solid empirical basis for assuming values for these parameters. To provide this information (which was virtually nonexistent when we began) a group of colleagues and I conducted a series of empirical studies using databases from school districts and prior educational evaluations.[17] Since then, findings from these studies have been used in proposals for cluster randomized studies by many researchers. In addition, Carolyn Hill, Ximena Portilla, and I collated similar information from studies of early childhood programs to create an MDRC research design tool that we named the "Effect-O-Sizer," which has been used for numerous MDRC proposals. We subsequently took information from the Effect-O-Sizer and from our papers and integrated it into the widely used Optimal Design computer program created by Steve Raudenbush and Jessaca Spybrook for assessing the power or precision of experimental designs. Consequently, this much-needed information is now widely available.

## THE POWER OF SIMPLE MESSAGES

In the late 1990s, Jim Riccio invited me to collaborate on a study of the relationship between program implementation and program impacts, using a massive, high-quality data set from three past multisite randomized trials of work-welfare programs conducted by MDRC. The sample for this analysis contained 69,399 individually randomized persons from 59 sites located in seven states. Data for sample members included extensive baseline information plus a short-run outcome measure: total earnings during the first two years after random assignment. Most important, however, was that across the three studies MDRC had collected comprehensive and consistent data from local staff surveys about the implementation of each site's program. Early in the project we were joined by Carolyn Hill, a doctoral student at the time, who was interested in research on program implementation.[18] Over the next several years, the three of us conducted a series of analyses and published our findings.[19] These findings explored the ability of key features of program implementation, program services, program environment, and program participants to predict cross-site variation in program impacts. Although we documented the importance of numerous predictors, one stood out among all others — the strength and consistency of the *message* delivered to program participants by local program staff members about the importance of finding a job as quickly as possible, even a low-paying job. This finding (that a message can be an important medium for change) had by far the strongest influence on program impacts that we observed, it was impervious to rigorous sensitivity tests, and it accorded well with common sense. In addition, I believe it has important implications for many other types of interventions.

17 Bloom, Bos, and Lee (1999); Bloom, Richburg-Hayes, and Black (2007); Jacob, Zhu, and Bloom (2010); Zhu, Bloom, Jacob, and Xu (2012). Important work on this issue also was done at the time by Hedges and Hedberg (2007).

18 Carolyn was introduced to Jim and me by her professor, Larry Lynn, who had helped Jim to conceptualize the project and raise funding for it.

19 Bloom, Hill, and Riccio (2003).

## THE POWER OF SIMPLE BENCHMARKS

In the early 2000s, Judy Gueron asked me to think about the appropriate role of standardized mean difference effect sizes as a metric for reporting MDRC findings. Judy's question arose from the fact that MDRC had recently begun to evaluate educational interventions, whose impacts are often reported as effect sizes (in standard deviation units), whereas MDRC's many evaluations of employment and training programs had typically reported impacts in dollars (for earnings and welfare receipt) or percentages (for employment and welfare receipt rates). As I began to explore the effect size literature in response to Judy's question, it quickly became clear that except for the widely used and somewhat arbitrary rule of thumb proposed by Jacob Cohen[20] — to consider effect sizes near 0.2, 0.5, and 0.8 standard deviations to be small, medium, and large, respectively — there was little guidance for interpreting effect-size findings. Thus I began to think about a project that would produce empirical benchmarks for this purpose. To help me, I contacted Mark Lipsey, whose work on interpreting effect sizes I admired,[21] and Carolyn Hill, with whom I had worked previously. Together we conducted an empirical study that produced three types of effect-size benchmarks for educational research: (1) the effect-size equivalent of annual growth in reading and math achievement by grade based on information for the national norming samples of seven major standardized tests; (2) the effect-size equivalent of existing achievement gaps by race/ethnicity, gender, and economic disadvantage; and (3) effect-size findings from past random assignment studies of educational interventions.[22] Since then, these benchmarks have been used by many researchers to plan evaluation studies and interpret their findings.

## THE POWER OF NATURALLY OCCURRING LOTTERIES

Around 2005, Richard Kahan, who had founded an organization that had created several innovative new small public high schools in New York City, asked a group of us at MDRC whether there was some way that we could measure the impacts of those schools (which we subsequently named small schools of choice, or SSCs). In addition to describing the new schools, his colleague Saskia Levy Thompson also described the new algorithm that assigns entering ninth-graders to high schools citywide. Because this algorithm contains an element of randomness, we all wondered whether somewhere inside of it was the statistical equivalent of naturally occurring lotteries. I subsequently identified those lotteries, which made it possible to rigorously estimate the impacts of almost 100 SSCs using data for over 20,000 individually randomized students, most of whom were disadvantaged students of color. Our findings demonstrated that SSCs markedly increased rates of high school graduation and postsecondary enrollment on average, and for a broad range of student subgroups defined by race/ethnicity, gender, prior academic performance, and economic disadvantage.[23] Using the experimental framework produced by SSC lotteries, we also were able to demonstrate that the cost per graduate for SSCs is *less* than that for their counterfactual counterparts.[24] These unusually rigorous findings for such a large educational innovation led to the SSC model becoming eligible

20 Cohen (1988).

21 Lipsey (1990).

22 Bloom, Hill, Black, and Lipsey (2008); Hill, Bloom, Black, and Lipsey (2008).

23 For example, Bloom and Unterman (2014); Unterman (2014).

24 Bifulco, Unterman, and Bloom (2014).

for funding under federal School Improvement Grants. Currently, MDRC is continuing to follow sample members to assess SSC impacts on their persistence in and graduation from postsecondary education. MDRC is also exploring factors that are hypothesized to influence SSC impacts. In addition, MDRC is implementing plans to follow sample members into the labor market to assess SSC impacts on their earnings and employment. None of this would have been possible without the naturally occurring lotteries that exist within New York City's annual high school assignment process.

## THE IMPORTANCE OF STUDYING CROSS-SITE IMPACT VARIATION

As part of our work for the William T. Grant Foundation, Steve Raudenbush and I began to focus on studying cross-site variation in program impacts using data from multisite randomized trials. This focus grew out of the longstanding desire of researchers, practitioners, and policymakers to learn when, why, and how interventions work, combined with the recent accumulation of high-quality multisite randomized trials in education research and related areas. As a first step, Mike Weiss, Tom Brock, and I developed a conceptual framework for identifying factors that produce impact variation.[25] In addition, Steve and I wrote an overview paper about opportunities for using multisite trials to study impact variation.[26] Furthermore, Steve and I (with Sean Reardon, Guanglei Hong, and Lindsay Page) led a two-day workshop at the University of Chicago on studying impact variation for a large group of prominent social science researchers, methodologists, and funders. Soon after this workshop, Mike McPherson, president of the Spencer Foundation, asked MDRC to organize an ambitious project that brought together leading methodologists from academia and the three organizations that had conducted most of the large multisite trials in education research: MDRC, Mathematica Policy Research, and Abt Associates, Inc. The goal of the project is to develop methods for studying the magnitude and sources of variation in program impacts and apply these methods to data from existing multisite trials. One focus of my role in the project is on the detection and quantification of cross-site impact variation. This work has produced three papers that will be published together in a forthcoming issue of the *Journal of Research on Educational Effectiveness*. One paper presents the statistical method we developed for estimating a cross-site mean and standard deviation of program impacts.[27] The second paper describes how to assess the precision of such estimates.[28] The third paper presents estimates of cross-site impact variation — plus an exploratory analysis of factors that influence the magnitude of this variation — based on data from 16 large-scale multisite trials.[29] My other main focus in the project involves participating in a team led by Sean Reardon that is developing an instrumental variables method for studying mediators of program impacts and using this method to explore potential mediators of the impacts of SSCs. I am excited by what our project has accomplished and hope that it will have a lasting influence on evaluation research.

## THE POWER OF GREAT COLLEAGUES

I conclude these reflections by highlighting the most gratifying aspect of my research journey — the wonderful multidisciplinary colleagues with

25 Weiss, Bloom, and Brock (2014).

26 Raudenbush and Bloom (2015).

27 Bloom, Raudenbush, Weiss, and Porter (2016).

28 Bloom and Spybrook (2017).

29 Weiss et al. (2017).

whom I have had the privilege to work. I think of this experience as a play in four acts. Act One began in 1976, when I became an assistant professor at Harvard, where I collaborated with Sunny Ladd and Johnny Yinger using natural experiments to study a range of issues in state and local public finance. Act Two began in 1986, with the National JTPA Study, which gave me the opportunity to work closely for many years with Larry Orr and his colleagues at Abt Associates and Judy Gueron and her colleagues at MDRC. Act Three began in 1998, when I came to MDRC, and since then I have had the opportunity to work closely with two of its presidents, Judy Gueron and Gordon Berlin, and many of its wonderful staff members. Act Four began in 2003, when I started to collaborate with Steve Raudenbush, which pushed me far beyond what I thought my analytic capacity might be and opened the door for me to work with a wonderful group of colleagues from academia. When all is said and done, it is these colleagues that I hold most dear.

## REFERENCES


Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91, 434: 445-455.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics* 63 (February): 47-57.

Bifulco, Robert, Rebecca Unterman, and Howard S. Bloom. 2014. *The Relative Costs of New York City's New Small Public High Schools of Choice*. New York: MDRC.

Bloom, Howard S. 1984a. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review* 8, 2: 225-246.

Bloom, Howard S. 1984b. "Estimating the Effect of Job-Training Programs Using Longitudinal Data: Ashenfelter's Findings Reconsidered." *Journal of Human Resources* 19, 4: 544-556.

Bloom, Howard S. 1995. "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs." *Evaluation Review* 19, 5: 547-556.

Bloom, Howard S. 2005. "Randomizing Groups to Evaluate Place-Based Programs." Pages 115-172 in Howard S. Bloom (ed.), *Learning More from Social Experiments: Evolving Analytic Approaches*. New York: Russell Sage Foundation.

Bloom, Howard S., Johannes M. Bos, and Suk-Won Lee. 1999. "Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs." *Evaluation Review* 23, 4: 445-469.

Bloom, Howard S., Carolyn J. Hill, Alison Rebeck Black, and Mark W. Lipsey. 2008. "Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions." *Journal of Research on Educational Effectiveness* 1, 4: 289-328.

Bloom, Howard S., Carolyn J. Hill, and James A. Riccio. 2003. "Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments." *Journal of Policy Analysis and Management* 22, 4: 551-575.

Bloom, Howard S., Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos. 1997. "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Ttraining Partnership Act Study." *Journal of Human Resources* 32, 3: 549-576.

Bloom, Howard S., and H. Douglas Price. 1975. "Voter Response to Short-Run Economic Conditions: the Asymmetric Effect of Prosperity and Recession." *American Political Science Review* 69, 4: 1240-1254.

Bloom, Howard S., Stephen W. Raudenbush, Michael J. Weiss, and Kristin Porter. 2016. "Using Multisite Experiments to Study Cross-Site Variation in Treatment Effects: A Hybrid Approach with Fixed Intercepts and a Random Treatment Coefficient." *Journal of Research on Educational Effectiveness*. Published online December 29. doi:10.1080/19345747.2016.1264518.

Bloom, Howard S., and James A. Riccio. 2005. "Using Place-Based Random Assignment and Comparative

Interrupted Time-Series Analysis to Evaluate the Jobs-Plus Employment Program for Public Housing Residents." *Annals of the American Association for Political and Social Science* 599 (May): 19-51.

Bloom, Howard S., Lashawn Richburg-Hayes, and Alison Rebeck Black. 2007. "Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions." *Educational Evaluation and Policy Analysis* 29, 1: 30-59.

Bloom, Howard S., and Jessaca Spybrook. 2017. "Assessing the Precision of Multisite Trials for Estimating the Parameters of a Cross-Site Population Distribution of Program Effects." *Journal of Research on Educational Effectiveness*. Published online March 10. doi:10.1080/19345747.2016.1271069.

Bloom, Howard S., and Rebecca Unterman. 2014. "Can Small High Schools of Choice Improve Educational Prospects for Disadvantaged Students?" *Journal of Policy Analysis and Management* 33, 2: 290-319.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimental Design and Analysis in Field Settings*. Chicago: Rand McNally College Publishing Company.

Hedges, Larry V., and E. C. Hedberg. 2007. "Intraclass Correlation Values for Planning Group Randomized Trials in Education." *Educational Evaluation and Policy Analysis* 29, 1: 60-87.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey. 2008. "Empirical Benchmarks for Interpreting Effect Sizes in Research." *Child Development Perspectives* 2, 3: 172-177.

Jacob, Robin, Pei Zhu, and Howard S. Bloom. 2010. "New Empirical Evidence for the Design of Group Randomized Trials in Education." *Journal of Research on Educational Effectiveness* 3, 2: 157-198.

Kramer, Gerald H. 1971. "Short-Term Fluctuations in U.S. Voting Behavior, 1896-1964." *American Political Science Review* 65 (March): 131-143.

Lipsey, Mark W. 1990. "Effect Size: the Problematic Parameter." Chap. 3 in *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, CA: Sage Publications.

Pitcher, Hugh M. 1979. "A Sensitivity Analysis to Determine Sample Sizes for Performing Impact Evaluation of the CETA Programs." Pages 37-78 in Farrell E. Bloch (ed.), *Evaluating Manpower Training Programs*. Research in Labor Economics, Supplement 1. Greenwich, CT: JAI Press.

Raudenbush, Stephen W. 1997. "Statistical Analysis and Optimal Design for Group Randomized Trials." *Psychological Methods* 2, 2: 173-185.

Raudenbush, Stephen W., and Howard S. Bloom. 2015. "Learning About and From a Distribution of Program Impacts Using Multisite Trials." *American Journal of Evaluation* 36, 4: 475-499. doi:10.1177/1098214015600515.

Stigler, George J. 1973. "General Economic Conditions and National Elections." *American Economic Review* 63 (May): 160-167.

Unterman, Rebecca. 2014. *Headed to College: The Effects of New York City's Small High Schools of Choice on Postsecondary Enrollment*. New York: MDRC.

Weiss, Michael J., Howard S. Bloom, and Thomas Brock. 2014. "A Conceptual Framework for Studying the Sources of Variation in Program Effects." *Journal of Policy Analysis and Management* 33, 3: 778-808.

Weiss, Michael J., Howard S. Bloom, Natalya Verbitsky-Savitz, Himani Gupta, Alma E. Vigil, and Daniel N. Cullinan. 2017. "How Much Do the Effects of Education and Training Programs Vary Across Sites?" *Journal of Research on Educational Effectiveness*. Published online April 19. doi:10.1080/19345747.2017.1300719.

Zhu, Pei, Howard S. Bloom, Robin Jacob, and Zeyu Xu. 2012. "Designing and Analyzing Studies That Randomize Schools to Estimate Intervention Effects on Student Academic Outcomes Without Classroom-Level Information." *Educational Evaluation and Policy Analysis* 34, 1: 45-68.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, Charles and Lynn Schusterman Family Foundation, The Edna McConnell Clark Foundation, Ford Foundation, The George Gund Foundation, Daniel and Corinne Goldman, The Harry and Jeanette Weinberg Foundation, Inc., The JPB Foundation, The Joyce Foundation, The Kresge Foundation, Laura and John Arnold Foundation, Sandler Foundation, and The Starr Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

The findings and conclusions in this report do not necessarily represent the official positions or policies of the funders.

For information about MDRC and copies of our publications, see our website: www.mdrc.org.



NEW YORK
16 East 34th Street, New York, NY 10016
Tel: 212 532 3200

OAKLAND
475 14th Street, Suite 750, Oakland, CA 94612
Tel: 510 663 6372

WASHINGTON, DC
1990 M Street, NW, Suite 340
Washington, DC 20036

LOS ANGELES
11965 Venice Boulevard, Suite 402
Los Angeles, CA 90066