MDRC Working Paper

Measuring Treatment Contrast in Randomized Controlled Trials

Gayle Hamilton
Susan Scrivener

June 2018

mdrc
BUILDING KNOWLEDGE TO IMPROVE SOCIAL POLICY

Funding for this publication was provided by the Laura and John Arnold Foundation, the Edna McConnell Clark Foundation, the Charles and Lynn Schusterman Foundation, the JPB Foundation, the Kresge Foundation, the Ford Foundation, and the Joyce Foundation.

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 publication 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.

Abstract

Many social policy evaluations examine the effects of a new program or initiative relative to a counterfactual, commonly referred to as a "business as usual" condition, meant to represent what would happen to people if the new program being tested did not exist. In some evaluations, few if any individuals in the counterfactual situation can receive services that are similar to the ones being studied. But in many evaluations, the counterfactual condition includes services that are in the same realm as the ones being studied, such as services provided by an alternative program or services that are widely available in the community. The difference between what the program group receives and what those in the counterfactual condition receive is called the *treatment contrast*. The treatment contrast is at the heart of most evaluations of new social policy programs or initiatives; thus, measuring and understanding the treatment contrast is critical to understanding what a program's measured effects represent.

This working paper explains the importance for social policy and program evaluations of the treatment contrast and offers guidance on how to measure that contrast. It draws on the knowledge and experience built from the hundreds of diverse evaluations that MDRC, an education and social policy research organization, has conducted in the past 40 years.

The paper makes the case that assessing treatment contrast yields benefits that include helping to identify the specific questions that an impact evaluation will and will not answer, highlighting the program components that might and might not be driving a program's effects, and suggesting why a program's effects might differ across cohorts or subgroups within a site or across sites. Procedurally, the working paper suggests planning for the treatment contrast analysis early in studies, having a treatment contrast measurement plan before program effect results are known, and remaining focused on the treatment contrast throughout a study. It also argues that the measurement of treatment contrast should receive as much attention from researchers as does assessing the process of program implementation or measuring fidelity to the original planned model. In choosing the aspects of treatment contrast to measure, the working paper suggests that the theory of change or logic model for the studied program should play the central role. While questions certainly remain about studying treatment contrast, this working paper provides ideas for researchers as they seek to understand what a program's measured effects do and do not suggest regarding the best ways to improve social programs and policies — a particularly important consideration when continuous improvement in program replication is a goal.

Contents

Acknowledgments

The authors thank Carolyn Hill and Ginger Knox at MDRC for their invaluable guidance and advice on this paper. The authors thank them and Gordon Berlin, Dan Bloom, Howard Bloom, William Corrin, Jo Ann Hsueh, Dina Israel, Michelle Manno, Shira Mattera, and Charles Michalopoulos for their helpful comments on earlier drafts of this paper. Rebecca Bender edited this paper and Ann Kottner and Carolyn Thomas prepared the document for publication.

The Authors

Measuring Treatment Contrast in Randomized Controlled Trials

Introduction

Many social policy evaluations examine the effects of a new program or initiative (or of a variation of a program or initiative) relative to something else — a counterfactual. In some such evaluations, few (if any) individuals in the counterfactual situation can receive services that are similar to the ones being studied. This was particularly the case in some early social program evaluations, such as in a 1970s study of an innovative preschool intervention (the High/Scope Perry Preschool study; Schweinhart, 2004) or in studies of early welfare-to-work programs in the 1970s and early 1980s (Gueron and Rolston, 2013). But in many evaluations, the counterfactual almost always includes services that *can* be similar to the ones being studied — that is, services that are in the same realm, available through an alternative program or widely or customarily available in the community, the school or college, or some other setting in which the evaluation is being conducted. These existing services are sometimes referred to as *business as usual.* The difference between the treatment received by people with access to the new program or initiative and the treatment received by people without access to the new program but with access to existing services is called the *treatment contrast*. The treatment contrast is at the heart of most evaluations of new social policy programs or initiatives, and measuring and understanding the treatment contrast is critical in understanding what a program's measured effects represent.

Drawing on the knowledge and experience built from the hundreds of diverse evaluations that MDRC, an education and social policy research firm, has conducted in the past 40 years, this paper sets forth specific reasons why it is important for social policy and program evaluations to focus on the treatment contrast and offers guidance on how to measure that contrast in program effectiveness studies. While the focus of the paper is on randomized controlled trials — a type of research design that MDRC has used to study social policy and program innovations in over 500 communities — many of the points made in the paper are applicable to other types of research designs, such as quasi-experimental designs, as well.

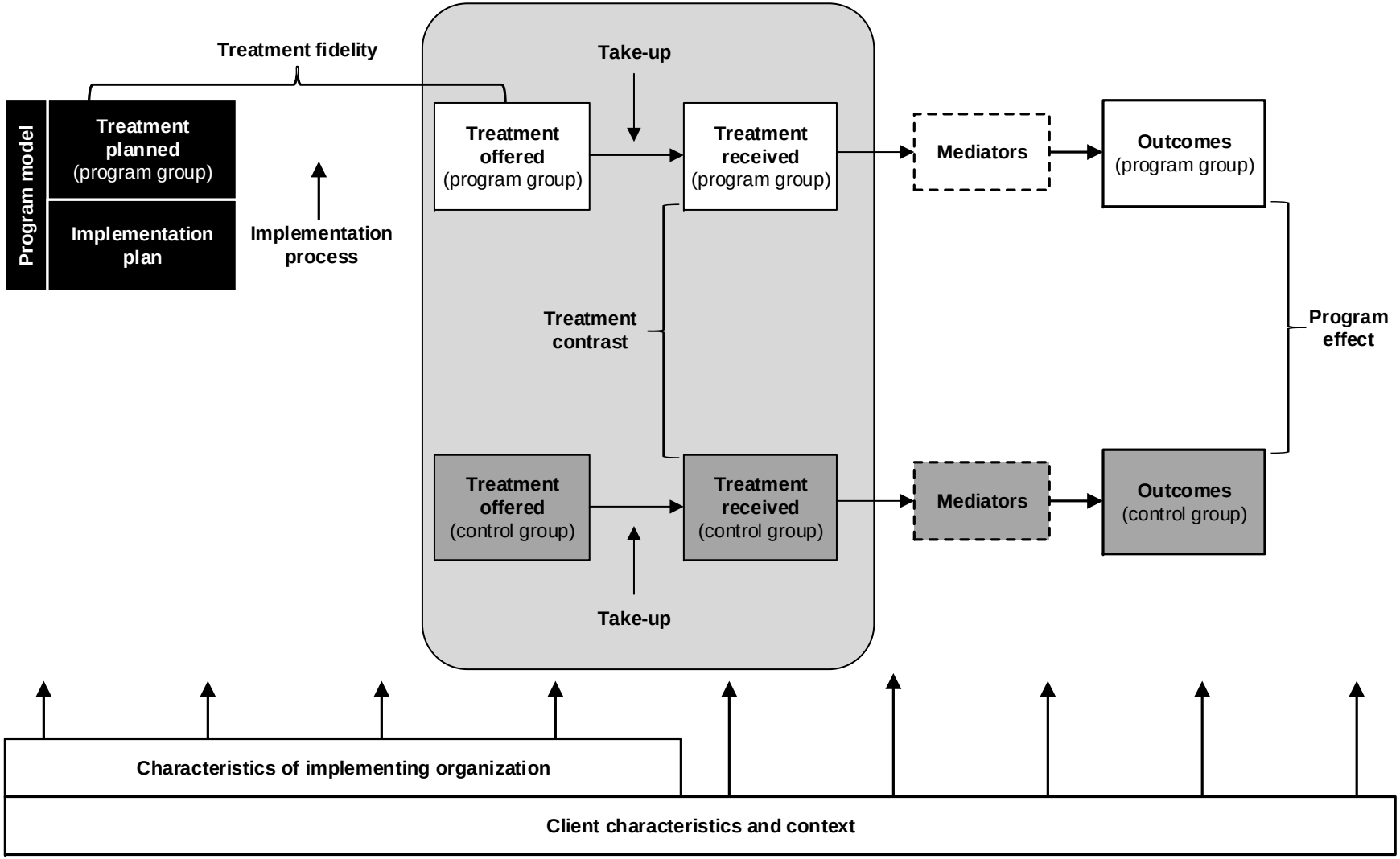
Why Is It Important to Focus on Treatment Contrast?

Concepts and Definitions

This paper draws from and builds upon a conceptual framework described by Weiss, Bloom, and Brock (2014). (See Figure 1, which is taken from that document.) While Weiss,

Figure 1

A Conceptual Framework for Studying Program Effects, Treatment Contrasts, and Implementation



Note: The large, shaded "Take-up" box in the center of the figure represents the focus of this paper.

Source: Weiss, Bloom, and Brock, 2014.

Bloom, and Brock's primary focus is on studying differences in program effects across different groups of people or in different circumstances, many of their key points and their framework are also relevant when focusing on treatment contrast in the evaluation of one program in one site.

Like Weiss, Bloom, and Brock's study, this paper begins with some definitions that come from the statistical literature on causal effect and are based on the concept of *potential outcomes* (for example, Rubin, 1974, 1978). In brief, potential outcomes for a person are the outcomes that the person would have under different sets of experiences or conditions. To establish the causal effect of offering a specific program to someone — sometimes referred to as the effect of the intent to treat (Angrist, Imbens, and Rubin, 1996) — it is assumed that each person has two potential outcomes: First, there is the outcome the person would experience if the person were assigned to or offered a program. This is usually referred to as the *treated outcome*. Second, there is the outcome the person would experience if the person were *not* assigned to or offered a program. This is usually referred to as the *untreated counterfactual outcome*, or as the *counterfactual* for short.

A person, of course, can only experience one condition at a time. Evaluators cannot simultaneously observe both potential outcomes for an individual, and thus they cannot observe a program (causal) effect for a particular person. Evaluators can, however, determine the average outcome for a sample of people who are offered a program (usually called the program or treatment group in an evaluation) as well as the average outcome for a sample of people who are not offered the program (usually called the control or comparison group in an evaluation). If the two samples or groups of people are virtually identical before they are or are not offered the program, then the measured difference in the two groups' average outcomes over time represents a reliable estimate of the average effect of offering the program to people (Rubin, 1974, 1978; Shadish, Cook, and Campbell, 2002; Weiss, Bloom, and Brock, 2014).

In randomized controlled trials, a program group and a control group are produced that are initially the same, on average, or at least are not systematically different.[1] The estimated effects measured in such impact evaluations compare potential outcomes under two treatment conditions: (1) access to program services (plus any other existing services) and (2) access only to other existing services. In essence, because the program effect is defined as a relative, average difference between groups' outcomes, the cause of the program effect has to be defined as a relative difference between treatment conditions as well (Cordray and Pion, 2006). It is the treatment contrast — that is, the difference in treatment between the two conditions (to which, in the case of randomized controlled trials, individuals are randomly assigned), also sometimes referred to as *achieved relative strength* (Cordray and Pion, 2006)

[1]For ease of reference, the rest of this paper uses only the term *program group* to refer to a program or treatment group and the term *control group* to encompass a control or comparison group.

— that causes program effects.[2] As Holland (1986) asserts, "the effect of a cause is *always* relative to another cause" (p. 946).[3]

As discussed in Weiss, Bloom, and Brock (2014) and shown in Figure 1, evaluations can focus on several phenomena when seeking to examine why a program had the effects that it did. Foci can include the planned treatment, the offered treatment, the treatment taken up, and the treatment contrast — in the context of a particular environment (for example, the characteristics of the implementing organizations and the individuals who were served). Weiss, Bloom, and Brock (2014), as well as this paper, defines the *implementation process* as the process leading from planned treatment to offered treatment. Both documents also define *treatment fidelity* as the similarity or difference between the planned treatment and the offered or taken-up treatment. The focus of this paper is specifically on treatment contrast, which is defined as the difference between the treatment received (or taken up) by program group and control group members.

Conceptual Advantages of Measuring Treatment Contrast

Given the centrality of treatment contrast in social policy and program evaluations, it is valuable to point out some of the specific ways in which measuring that contrast is useful. These advantages fall into several categories — informing the program impact (or effectiveness) evaluation results, informing understanding of the program's implementation, and informing program replication and scaling efforts — and are described in that order below.

- **Measuring treatment contrast identifies the specific effectiveness questions that an impact evaluation will answer.**

Most basically, measuring the treatment contrast provides information about exactly what an impact evaluation is testing — that is, what policy questions the results of an impact evaluation will and will not answer. It reveals what produced the effects that are found. Except for the rare evaluations with "no service" counterfactuals, it is not simply the new program that leads to changes in outcomes — it is the *difference* between that new program and the counterfactual.

---

[2]The treatment contrast is sometimes referred to as the *service contrast*. This paper uses the term *treatment contrast* because program interventions often do not consist exclusively of services, and evaluations often test the effects of program aspects that are not services. For example, tests can compare mandatory participation with voluntary participation, different administrative and staffing arrangements, or different institutions providing similar services.

[3]Ideally, the concept of treatment contrast also would be taken into account when designing program performance measures outside of an impact evaluation. For decades, however, policymakers have struggled over how program performance measures can take into account the program's counterfactual. This topic is beyond the scope of this paper.

Consider an example in which the implementation of a program MDRC was studying was relatively strong and remained fairly constant over time, but the counterfactual condition turned out to be different from what was originally anticipated. This was the case in an evaluation that tested whether a specialized math curriculum supported by intensive professional development could strengthen children's pre-kindergarten (pre-K) experiences and subsequent outcomes on a large scale in New York City — the Making Pre-K Count evaluation (Morris, Mattera, and Maier, 2016; Mattera, Jacob, and Morris, 2018). The study built upon extant literature and program scans in New York City showing that most preschools historically had conducted very little math instruction. These efforts suggested that the intervention might be powerful, given the comparatively weak math instructional counterfactual conditions. The study found that the implementation of the intervention was reasonably strong for most program components, and practices were in line with most fidelity indicators. However, the counterfactual conditions were not static; over the course of the evaluation, New York City began to emphasize the alignment of math and literacy curricula with Common Core standards and, during the second year of the evaluation, began to roll out universal pre-K. Given these developments, over time the business-as-usual condition became more similar to the new intervention being tested, particularly with regard to the amount of math instruction that was given. As a result, the treatment contrast in the second year of the evaluation did not represent as stark a departure from the usual math curriculum and support as had existed in prior trials of the intervention in other large cities. The specific effectiveness question addressed in the evaluation thus changed from the originally anticipated one of "What difference does Making Pre-K Count make in an environment with weak math instruction?" to the actual question of "What difference does Making Pre-K Count make in an environment with an increasing emphasis on math instruction and an increasing availability of pre-K?"

- **When an impact evaluation finds effects, measuring treatment contrast highlights what aspects of an intervention might and might not be driving the program's effects.**

Measuring treatment contrast can also provide information about what dimensions of a program may be driving its effects. One way it can do this is by identifying, in a multicomponent program, components for which there is no or very little treatment contrast; these components are unlikely to be driving effects. An evaluation of a multicomponent program for community college students that includes blocked or linked courses and almost doubled three-year graduation rates — the City University of New York's Accelerated Study in Associate Programs (CUNY ASAP) — provides an example of this (Scrivener et al., 2015). The study found that the blocked or linked courses were implemented somewhat differently across colleges, most program group students did not take a complete block of courses, and the control condition included some blocked or linked courses. As a result, the treatment contrast on that component was small. The treatment contrast on several other components was large, and the

evaluation concluded that the blocked or linked courses were unlikely to be driving the program effects that were found.

While in general in a multicomponent program, components for which there is more substantial treatment contrast are more likely to be driving the impacts, the relative size of the treatment contrast is not the only aspect that matters. More nuanced analysis is needed. For one, treatment contrast aspects should be considered and interpreted using the program's theory of change and logic model. In addition, the nature of the service for which there is contrast should be taken into account. Some services, despite substantial impacts, may be unlikely to play a large role in driving impacts. In addition, the *quality* of services in the new initiative, compared with the quality of services that are part of business as usual, may be playing a role. In general, the analysis should incorporate exploring other, detailed implementation research data, often qualitative, in addition to quantitative information, to assess what aspects of multicomponent programs might be driving impacts. For example, in one test within a large 1990s study of the effectiveness of different welfare-to-work strategies — the National Evaluation of Welfare-to-Work Strategies, or NEWWS (Hamilton and Scrivener, 2012a) — the welfare agency that was operating the program partnered with a local community college on the program design and operation, but assignments to academic college courses, as opposed to basic education or vocational training courses, were not permitted. The program increased the proportion of welfare recipients who took at least one college course for credit in the second half of the study's five-year follow-up period, relative to the control group. Some policymakers assumed that staff members did not follow the program's protocol and the increase in taking academic college courses was driving the program's substantial earnings effects. Detailed treatment contrast analysis, however, suggested that the increase in college courses was not driven by specific actions taken by program staff members and, in fact, occurred after most participants had left the program (and after the substantial earnings effects began to appear). Rather, the increase appeared to be driven primarily by welfare recipients' increased exposure to the community college system while they were participating in job search and other program activities that were hosted by the colleges, which led program participants to eventually enroll in college courses. The discovery of this added benefit of partnering with local community colleges contributed to the design of later program initiatives in this locality.

- **When an impact evaluation finds no effects, measuring treatment contrast identifies the intervention components that may not independently make a difference.**

As much as it is helpful to understand what might be driving a program's effect, it also is helpful to identify when a substantial treatment contrast actually does not make a difference in the outcomes that are being examined. In other words, it is often as valuable to understand which treatment contrasts do not make a difference as to understand which do. An example is

provided by the multi-test Employment Retention and Advancement (ERA) evaluation, which examined the effectiveness of different approaches to encouraging low-income individuals to retain employment and advance in their earnings over time. Several ERA programs offered case management on a voluntary basis to employed individuals in the form of job coaching, referrals to supportive services, assistance in developing career plans, and advice on education and training programs. While some of these programs increased program group members' participation in such case management services relative to control group levels, none of these programs increased employment or earnings (Hamilton and Scrivener, 2012b). This result led the ERA researchers to conclude that although case management may be a necessary ingredient of programs, by itself it is not sufficient to make a meaningful difference in employment outcomes unless it is combined with other, more concrete services.

- **Measuring treatment contrast discourages drawing inaccurate conclusions, based simply on participation patterns in the new initiative or program, about what is driving effects.**

Conducting a thorough assessment of treatment contrast can result in a more accurate understanding of programs and the sources of their effects. For example, several tests in the earlier-mentioned NEWWS project examined a number of programs with different emphases (Hamilton and Scrivener, 2012a). In three cities, education-and-training-first programs that required welfare recipients to participate in education or training before job search resulted in substantial numbers of individuals engaging in these activities: In the three programs combined, over a five-year follow-up period, 40 percent of program group members participated in adult basic education (ABE), 28 percent participated in vocational training, and 18 percent participated in postsecondary education.[4] On their own, however, control group members were more likely to enroll in vocational training than in ABE. Thus, a program-control differential was more common and much larger for ABE participation than for vocational training, and there was no differential for postsecondary education. As a result of this detailed treatment contrast analysis, researchers concluded that the employment and earnings impacts found for these programs stemmed (in part, given other key program components) from increased participation in ABE, to a much lesser extent from vocational training, and not at all from postsecondary education.

- **Measuring treatment contrast suggests why a program's effects might differ across cohorts or subgroups within a site or across sites.**

Measuring treatment contrast can also help identify why a program generates effects for some cohorts, subgroups, or sites but not others. Both program implementation and the counter-

[4]These three-site combined statistics were calculated from tables in Hamilton et al., 2001.

factual can vary across cohorts, subgroups, and sites, thus leading to different treatment contrasts. In other words, treatment contrast measurement can help explain impact variation within a site, as well as across sites. An evaluation of a multifaceted program designed to meet the holistic needs of secondary school students, called Diplomas Now (Sepanik et al., 2015), provides an example. The study found that, generally speaking, impacts on students' outcomes were better for the second cohort of students compared with the first cohort (and some of the differences across cohorts were statistically significant). The evaluation also found that the treatment contrast increased from the first year of implementation to the second year. Project researchers concluded that this increase reflected a combination of Diplomas Now schools maintaining or improving their implementation over time and counterfactual schools declining in their implementation of similar activities. Thus, the treatment condition remained stable or improved, and the counterfactual shifted. The researchers did not analyze associations between specific aspects of the increased contrast and the larger impacts, but the team hypothesized that the overall increased second-year treatment contrast might have contributed to the larger second-year impacts (Corrin et al., 2016).[5]

- **Measuring treatment contrast can indicate when technical assistance is needed or when an impact evaluation may not be worthwhile.**

Assessing treatment contrast early in a study can provide information about whether an impact study will answer the intended question and, in demonstrations or other projects in which technical assistance is provided, whether technical assistance is needed, and what that assistance should aim to change. All entities involved in impact evaluations — evaluators, funders, and the program's developers and operators — must be certain that a new program is different from what is already available ("business as usual") and that the impact evaluation will truly add knowledge about the benefits that a program or intervention might be able to confer. While researchers often focus early in an impact evaluation on the extent to which a program has fidelity to its intended form, early examinations of relevant dimensions of the counterfactual are also often useful. In a study of three new service approaches intended to help address fathers' personal or societal barriers to positive involvement with their children, called Building Bridges and Bonds (Harknett, Manno, and Balu, 2017), for example, the pilot phase of the impact evaluation was used to examine the planned counterfactual within each study site and to make adjustments to the offered services in sites where there seemed to be a danger that the

[5]The Diplomas Now evaluation also suggested another advantage of exploring treatment contrast in detail: In some schools that implemented the new program, implementation fidelity of some components of the program (for example, the coaching of the teachers) did not reach the thresholds the program developer had hypothesized would produce a meaningful effect (Corrin, Sepanik, Rosen, and Shane, 2016). But, in contrast to the services provided at the control group schools, the new program still had an effect. This finding suggests that treatment contrast analysis also can help refine fidelity standards for new initiatives.

treatment contrast will be too narrow. The project is monitoring the potential for the control group to access similar services in communities as local policy contexts change, and technical assistance is being used to increase program group participation in the add-on services that are being tested.

Although pre-enrollment analysis is not the main focus of this paper, treatment contrast is sometimes assessed even before sample members are enrolled in a study. The research design phase for a study of an education-conditioned internship program for disconnected youth — Project Rise — provides an example (Manno, Yang, and Bangser, 2015). At the start of the design work for a planned randomized controlled trial, interviews with New York City officials and nonprofit administrators indicated that the nonprofit organizations running Project Rise could also access a citywide internship program. After random assignment, then, it was likely that individuals who were assigned to the control group would be referred to internships that would be fairly similar to those offered in Project Rise. As a result, researchers did not initiate a randomized controlled trial, instead conducting a detailed implementation study of the then-new Project Rise model. Most fundamentally, the implementation study suggested that the offered paid internships, which were expected to be the most significant draw for participants, were not as much of a participation inducement as were the free General Educational Development (GED) preparation classes that were offered as part of the program. (The study also included an assessment of whether an eventual randomized controlled trial was warranted in the future.)

Finally, even after study enrollment has begun, early treatment contrast analysis — done before any program effectiveness impact analyses — sometimes reveals that a certain site has no treatment contrast, and thus implementing an impact evaluation in that site will not indicate the potential benefits of the intervention that is under study. In such cases, technical assistance is often used to improve the situation. If there is not sufficient improvement, sites occasionally may be dropped from the evaluation, after extensive discussions within the evaluation teams and with the program operators.

- **Measuring treatment contrast can inform replication and scale-up efforts.**

Information about treatment contrast is also critical when replicating or scaling up programs or initiatives. Understanding the treatment contrast in the original study helps to clarify what drove the effects. And, in order to generate program effects that are similar in type, direction, and magnitude to those seen in the original evaluations, it must be determined whether the program or initiative that will be replicated or scaled up is sufficiently different from what currently exists in the new specific contexts and for the new specific target populations. As Knox, Hill, and Berlin (2018) assert, it is necessary to understand the treatment contrast when a program was originally found effective, as well as in the potential new service setting, to be able to direct new programs to contexts where they can add value.

Consider the example of a well-known education reform — Success for All (SFA) — that was scaled up (Quint et al., 2015). Experimental and quasi-experimental evaluations of SFA, which aims to improve students' reading skills in the elementary grades, showed that students in SFA schools performed better on standardized tests than students who were receiving other reading programs. Between 1987, when SFA was developed, and 2010, when it began to be scaled up as part of a U.S. Department of Education i3 grant, reading instruction in the United States changed markedly. For various reasons, the emphasis on phonics — a central focus of SFA — increased and additional interventions for struggling readers were implemented. These developments narrowed the differences between schools adopting SFA and schools using other reading programs and made it more difficult for SFA to show markedly greater positive effects. In other words, the counterfactual condition shifted over time, and the later implementation of SFA yielded a smaller treatment contrast than was the case in the original studies.

Smaller treatment contrasts that occur when programs are scaled up can also result from changes in the treatment condition. Often, interventions become weakened or altered when they are implemented at scale and thus have reduced fidelity to their original model. As a consequence, they become less different from "business as usual" in their new settings. For example, in a study of a particular approach to providing early intervention services to elementary students who were at risk of reading failure or who exhibited other academic or behavioral problems — Response to Intervention, or RtI (Balu et al., 2015) — researchers found that schools had implemented a number of changes to RtI, compared with the way it was operated in earlier, much smaller and more researcher-controlled studies. The 2015 study of 146 elementary schools found that relative to the earlier studies, a different set of students (for example, some individuals reading above grade level as well as individuals reading below grade level) were assigned to intervention services, RtI instruction sometimes displaced core instructional time as opposed to supplementing it, and schools had made adaptations in who provided intervention services to students.

While condition changes similar to the ones described above complicate researchers' analysis plans, one of the reasons the social policy field tries to establish evidence about what works and does not work, in fact, is to eventually affect business as usual. One goal of research is to identify programs and practices that have positive results — that is, to generate evidence that then can provide the rationale for the studied programs and practices eventually becoming the norm. For example, in studies of welfare-to-work programs, it is increasingly common that control group recipients of TANF benefits will be offered job search, education, or training assistance by local public agencies. This trend is due, in part, to research over the last 30 years that has indicated that providing welfare recipients with employment and training services increases their likelihood of obtaining employment and increases their earnings (Hamilton,

2012). Thus, more and more, such welfare-to-work programs have become business as usual in the United States — a positive development.

Given the conceptual advantages of measuring treatment contrast that are described above, it makes sense to include at least some treatment contrast analysis in the overall research plan of impact evaluations that compare an intervention with a counterfactual condition. In other words, it is important to make room for the measurement and analysis of treatment contrast indicators within the overall research plan for an evaluation. In particular, while important aspects of implementation research in most studies include investigating general features of program implementation as well as how closely the implemented treatment hews to the planned treatment, these activities should not crowd out a focus on — and the expenditure of resources on — treatment contrast.

It is thus clear that a more explicit, systematic focus on treatment contrast is desirable as a part of evaluations' implementation research or more general research plan. The following section discusses what to ideally measure and proposes ways to measure it.

Measuring Treatment Contrast

Treatment contrast details are particular to each study. As a result, this section offers guidance and strategies to use when deciding what to measure and how to measure it when examining treatment contrast, but not hard-and-fast instructions. Regardless of the nuances of each study, however, this paper suggests that implementation researchers should follow these guidelines:

- *Involve researchers representing all major aspects of an evaluation in discussing and understanding treatment contrast conceptualization and measurement.* Various research team members will bring different skills — including perspectives (for example, impact, implementation, or cost perspectives), substantive knowledge (for example, from the field or from past studies), and measurement expertise (for example, about treatment contrast indicators that can be measured through program records, study sample member surveys, or qualitative data) — regarding different aspects of the treatment contrast. Together, the evaluation team members must ensure that the treatment contrast analysis is sufficiently planned; has a design that is fully informed by the theory of change for the intervention under study; takes advantage of the evaluation team's thorough knowledge of the intervention's context, components, and implementation; and is soundly operationalized and carried out.

- *Start planning what to measure for the treatment contrast analysis, and how to measure it, early in a project*. Such early planning will allow the evaluation team researchers, including the implementation researchers, to embed treatment contrast measures in a wide array of data-collection efforts. As one illustration, staff members may be able to collect relevant data as part of site-selection visits or early technical assistance visits, and collect them in ways that will complement other, later, treatment contrast measures, if the treatment contrast measurement plan is developed early enough. Furthermore, such data may be useful in refining the treatment contrast measurement plan. Many evaluations require data-collection plans and instruments to be developed very early in the project to facilitate Office of Management and Budget (OMB) review and approval or Institutional Review Board (IRB) evaluation, underscoring the need to design the treatment contrast measurement and analysis plans early.

- *Stay focused on treatment contrast throughout the study*. Implementation and other researchers should be attuned to changes over time in the program intervention as well as in the counterfactual environment, and to how the study might assess these changed conditions at various points in time. Researchers should also think about how to share the information from these assessments with other evaluation team members so the team can consider the implications of what is being learned about the treatment contrast for the evaluation as a whole.

- *Identify what to measure and how to measure it before the impact results are known*. As noted above, planning for the treatment contrast analysis should take place early in studies. But at the least, the treatment contrast measurement plan should be set before the impact results are known. This avoids the appearance (or actual occurrence) of "fishing" for explanations for impact findings and makes the treatment contrast analysis more credible.

- *Prioritize among all possible contrast measurement options*. In general, measuring treatment contrast in a multifaceted program will take more resources — that is, time and money — than will measuring it in a single-focus initiative. Rarely, however, will project resources permit the measurement of all aspects of the treatment contrast in a particular study. Thus, decisions about what to measure, how to measure it, and how often to measure it will all be subject to resource constraints. For this reason, it is best to prioritize among the treatment contrast measurement options — similar to what is done regarding other design decisions made in evalua-

tions — so a study can make sure that the most essential treatment contrast measurement options are definitely done and the others are done if eventual resources permit. (In deciding how much of the project resources to devote to measuring treatment contrast, researchers will, of course, need to consider not only the rest of the implementation research but also the other research in the project. This paper does not address how to prioritize among all possible research questions on a project.)

Keeping the above general guidelines in mind, the remainder of this section gives more specific suggestions about how to plan treatment contrast analyses and set priorities within them.

What to Measure

As mentioned earlier, the focus of this paper is on treatment contrast, and not specifically on programs' planned treatment, offered treatment, or the treatment actually taken up — phenomena that are routinely examined as part of implementation research. As a result, this paper does not address how to measure or assess program implementation (the process leading from planned treatment to offered treatment) or fidelity (the similarity or difference between the planned treatment and the offered or taken-up treatment). Measuring and assessing these phenomena, however, often takes up much of researchers' focus, time, and resources. This paper argues that researchers should give as much attention to treatment contrast as they do to examining the process of program implementation or measuring treatment fidelity. The paper also asserts that researchers should seek to collect data on the implementation process and fidelity for both the program group (that is, data pertaining to the treatment) and the control group (that is, data pertaining to whatever services are available to the control group), and that both sets of data should inform the treatment contrast as well as feed into the general implementation research analysis. Moreover, in many cases it could be wise to start evaluation planning by focusing first on an examination of counterfactual conditions to firmly establish the definition of business as usual, to clearly define the nature and extent of the problem that a new intervention intends to solve, and to ensure that the planned new intervention is different enough from business as usual that it has a chance to make a significant dent in the problem of focus.

Continuing to follow the framework in Weiss, Bloom, and Brock (2014), there are at least four important dimensions of treatment received to *consider* measuring — for both the intervention of interest and the counterfactual condition — when assessing treatment contrast. These dimensions are described below.

Content

What services are provided? What other features of the intervention or treatment are unusual and possibly different from the counterfactual condition? Weiss, Bloom, and Brock (2014) define content as "the features, components, or ingredients of a service package that are a program's basic building blocks" (p. 10). Examples of services include subsidized employment positions; enhanced academic advising; or providing individuals with help to develop their résumés, locate job openings, and prepare for job interviews. Examples of other types of features are mandatory participation requirements (for example, a requirement to meet with a college adviser), access to a different type of counselor or a different method of counseling (for example, using motivational interviewing or cognitive behavioral therapy), or exposure to a specific emphasis within offered services — aspects of content that usually may not be seen as services. One test within the ERA study (Hendra et al., 2010), for example, compared two different versions of a "job club" — a group job-search activity designed to help unemployed individuals receiving TANF benefits find work. One job-club model urged individuals to seek jobs in their field of interest, based on the theory that this might enable them to stay in jobs longer and ultimately move into better jobs along a career path. The other model was a traditional one, which emphasized getting jobs quickly, regardless of the field. In this test, special attention needed to be paid to the content of the job clubs to ensure that there was a treatment contrast and that it was in line with the two theories being tested. Analyses of observational data, staff member interviews, and study sample member surveys led the researchers to conclude that the two job clubs did indeed deliver different messages about the types of jobs to seek.

Quantity

What is the "dosage" of treatment provided? What percentage of program group and control group members participate in services or take advantage of other treatment features, how much time do they typically spend in services or connected to other features, and over what period of time do they participate? In a study examining the effectiveness of "boot camps" meant to quickly train individuals in certain job skills needed in particular industry sectors, for example, it would be important to compare the amount of training that program group members receive with the amount of training that control group members receive. In this case, the amount could be measured by calculating, for individuals in the program group as well as the control group, the percentage who ever participated in training (frequency) multiplied by the number of days spent in training (duration) multiplied by the number of hours typically spent per day in training (intensity).

Quality

How well is the treatment provided? What is the quality of the provided treatment? For some types of interventions, it is possible to compare the quality of interactions between service providers (for example, teachers) and study sample members (for example, students), for both program and control group members, using observation instruments like the Classroom Assessment Scoring System (CLASS) or the Home Visit Rating Scales (HOVRS) or self-report measures like the Working Alliance Inventory. For other types of interventions, different instruments can assess program settings and resources in both the intervention and counterfactual environments, such as the child-care-quality rating scales used to assess preschools and Head Start centers. For still other intervention types, like job-search or job-training programs, other quality-assessment instruments or scales may exist, or may need to be created. Overall, quality is very hard to assess because even program developers do not necessarily have a good working definition of it for their particular intervention. Consequently, when preexisting quality instruments are used, evaluators should have a fair amount of confidence that those instruments themselves are evidence-based — in other words, that they have been clearly shown to be able to measure quality. (This is sometimes defined as whether the measure is predictive of future outcomes of interest. For example, a measure of classroom quality should ideally be predictive of children's outcomes.) When these types of instruments are lacking, in some cases it is possible to use proxy indicators for quality. For example, when studying a training program, it is often useful to collect information about the cost per course taken by program group members and compare it with the cost per course taken by control group members. While this is an imperfect measure of quality, it can provide a sense of the value that the market places on the various training courses. In particular fields, other types of proxies that do not require direct observation may be available. For example, the continuity of care, the intensity of care, or the extent to which staff members report that their practices adhere to particular core principles of service provision are sometimes used as proxies for quality.[6]

Conveyance

By what delivery mode, when, and by whom is the treatment provided? Usually, this is taken to mean whether services are provided in person or online, individually or in groups, and by counselors, teachers, or other types of staff members. The treatment contrast, however, can also be influenced by the characteristics of the organizations that are delivering services. In fact, some studies specifically test whether conveyance by one type of organization produces better results than conveyance by another type of organization. For example, another test within the ERA study compared three approaches to permitting employed TANF recipients to meet their state's weekly TANF participation requirement (Hendra et al., 2010).

[6]In general, when proxies for quality are used, they should have known correlations with actual quality.

Business as usual for TANF recipients working at least 20 hours per week gave individuals no flexibility to reduce or eliminate their required TANF work-hour obligation if they were participating in education or training. The two variations compared with this approach gave TANF recipients more flexibility to reduce or eliminate their TANF work-hour obligation if they were also in school or training, but they also differed from each other in terms of the institution that operated the program: One was run by the county welfare agency and one was run by the local workforce agency. Thus, in this test, it was essential to examine, for each agency, their procedures, their staff members' practices, and the level of access each agency had to different types of education or training options.

The above four dimensions usually should be assessed for program group members *as well as* for control group members when measuring treatment contrast. Without information about a dimension for the control group, it is very hard to measure the treatment contrast on that dimension. Nevertheless, in some situations it is fine to measure a certain construct in a dimension for the program group and not for the control group. This type of situation usually occurs when it is fairly impossible for the control group to receive a given service or treatment. In some early welfare-to-work studies, for example, Aid to Families with Dependent Children (AFDC) and TANF recipients in the program group were required to participate in community work experience positions; AFDC/TANF benefit recipients in the control group did not have this requirement (see, for example, Hamilton and Friedlander, 1989). Given that only individuals in the program group were offered these positions, and that at the time no other organization was offering these positions outside of AFDC/TANF, it was not necessary to expend any resources investigating control group members' take-up or participation in community work experience positions — it could be assumed to be zero. A more recent example of this situation is a study of a pilot program to simulate an expanded Earned Income Tax Credit (EITC) for low-income single workers without dependent children (Miller et al., 2017). This type of EITC is available only for the program group members in the study. Thus, take-up of this credit among control group members can be assumed to be zero.

How should implementation researchers determine what aspects of the treatment contrast to measure, and which should get highest priority if there are resource constraints? This process will differ for each evaluation, given that the nature of the treatment contrast — a specific treatment compared with a specific counterfactual — is very context dependent. But this paper suggests that implementation researchers take the following steps for each study:

- *Develop a logic model*. Sometimes evaluation teams will need to start from scratch to develop a logic model — which illustrates the specifics of the theory of change — and sometimes the studied programs will already have one that can be used or refined. In addition, in some studies researchers work with program developers who can help define or refine the program's theory

of change or logic model. The logic model will help the researchers identify the key aspects or constructs of treatment contrast that are expected in the study. As a team develops or refines the logic model, the specific focus on treatment contrast also may point to inputs, outputs, or contextual dimensions that should be added to the model.

- *Create a list of all treatment contrast constructs*. These constructs should be operationally defined, a process that often results in expressing the concept as hypotheses. For example, if a central aspect of the intervention under study is individualized attention and this aspect is supposed to be a key departure from the counterfactual, this concept might be defined by identifying subconstructs of the construct: Individuals in the intervention under study, compared with individuals in the business-as-usual control condition, will be more likely to interact with X types of staff members; will have more frequent contact with staff members — at least Y times a month more often; and will characterize their relationships with staff members differently, as A, B, and C.

- *Categorize each treatment contrast construct (or subconstruct) as high or low priority for measurement*. While ideally all constructs or subconstructs would be measured in a study, very few studies have the resources to do so. Thus, some type of prioritization is needed. Categorization based on high or low priority best takes place in a two-stage process. In the first stage, priorities should be set based on the intervention's logic model. This will take into account the salience of each aspect of the intervention under study. Some aspects of the intervention might be more fundamental to the intervention's theory of change than others. At this point, it is useful to ask: What is likely to be most valuable when later interpreting a program's impacts or lack of impacts? In the second stage, the ease or difficulty of measuring each subconstruct should be assessed. This assessment will require the study team to think about how to measure each subconstruct. At the end of this second stage, the subconstructs should be re-prioritized, and the constructs with the highest priority should be considered first for investigation.

- *Consider measuring other aspects of treatment contrast — for example, mediators and unintended treatment differences*. Some interventions have mediators that help to translate the treatment that is received into the outcomes that are of most interest. In job-training programs that are expected to increase individuals' earnings, for example, the treatment may primarily consist of placing individuals, at no cost to them, into training programs that are aligned with industry needs and that train for jobs that have frequent open-

ings and pay well. A key mediator between this treatment and increased earnings, however, is participants' persistence in and completion of the training program. This type of mediator should be specified in the logic model and defined jointly with a study's impact analysts. In any case, mediators need to be measured for both treatment and control group members, similar to other treatment contrast constructs. Unintended but possibly important treatment contrasts also need to be considered. Using the above training program example, the treatment contrast analysis should consider the "opportunity cost" of participating in the training program. If, in the absence of the offered training, control group members end up being more likely to take academic college courses and make more progress toward an academic degree, this substitution effect should be measured in the treatment contrast analysis. The study's logic model can help to identify the possibility of this type of treatment contrast, also.

How to Measure It

As noted above, an intervention's logic model or theory of change, in combination with other considerations, will suggest *what* treatment contrast aspects should be considered for measurement. But what should be taken into account in deciding *how* to measure these treatment contrast aspects?

Two overall concepts are helpful to keep in mind. First, the focus of this paper is on measuring treatment contrast within the context of randomized controlled trials as opposed to other types of research designs, and randomized controlled trials make the *interpretation* of measured treatment contrasts much more straightforward. In a randomized controlled trial, at the time of study entry individuals assigned to different research groups are virtually identical — on average — on such things as their characteristics and family structures, interests, and motivations. This type of research design thus gives implementation researchers a huge advantage in interpreting measured treatment contrasts: Any post-study entry differences between the groups — in their propensity to interact with certain types of program staff members, take up services at all, participate in different types of services, stay longer in services, substitute some services for other services, or forfeit benefits when they don't participate in services — can confidently be interpreted as a result of the intervention under study. These differences are indicators of the treatment contrast. Other types of research designs require assumptions to be made about how similar the counterfactual group was to the group that was to be offered the new program at the start of the study. These assumptions, which often cannot be verified (for example, about similarities in the groups' motivational levels), must also be taken into account in interpreting any differences between the program and the counterfactual groups' take-up of services — that is, in the treatment contrast measurement. Overall, then, interpreting the

meaning of treatment contrast measures within randomized controlled trials, compared with other types of designs, requires fewer assumptions; thus, the interpretation of the measurement in randomized controlled trials is likely more straightforward and, generally, more accurate.

Second, however, the accuracy of the treatment contrast measurement depends, in part, on the extent to which actions and statuses are measured in the same way over the same time period for program group members as for control group members. Comparing reports of services received that were obtained through a telephone survey for program group members, for example, with reports of services received that were obtained through a state management information system for control group members is very likely to result in poor measurement of the treatment contrast: The different data sources are likely to affect the average reported treatment received with and without access to the intervention under study and, as a result, give the researchers an incorrect understanding of the nature of the treatment contrast. Thus, to the extent that it is possible and practical, implementation researchers should seek to design measures of treatment contrast that are identical for program and control groups.

When determining *how* to measure treatment contrast, there are four measurement dimensions to consider: the timing of the measurement, the sample members to include in the measurement, the perspectives from which to conduct the assessment, and the type of data and methods to be used in the measurement. Each of these dimensions is discussed below.

Timing

The treatment contrast can be measured over a set period of time (for example, comparing the percentage of program group and control group members who ever participated in a certain activity within a 12-month post-study-entry follow-up period or within the academic year in which they were randomly assigned), or at a specific point in time (for example, comparing the percentage of program group and control group members who were active in a specific service as of a follow-up survey interview or as of the end of a particular academic semester). Of course, all treatment contrast measurements within a study do not have to cover the same time period: Different components or dimensions of the treatment contrast can be measured over different periods or at different points in time.

Measurement of the treatment contrast also can cover the same entire follow-up period used in the impact or effectiveness analysis or specific portions of that follow-up period. A test of a program in Texas that offered several services, including earnings stipends to individuals who left the TANF program because they found work, studied as part of the ERA project (Martinson and Hendra, 2006), took "readings" of the treatment contrast periodically through sample member surveys and the collection of stipend disbursement data. The treatment contrast analysis indicated that the program and control group members received similar pre-employment services and that their treatment also was very similar during the four months after

they found employment, during which time their earnings were disregarded in the calculation of their TANF grants. Only after this disregard period was there a large treatment contrast. This pattern of treatment contrast helped to explain the pattern of earnings impacts, in that impacts did not emerge until well into the second year of study participant follow-up.

Generally, the intervention's theory of change (which, when applied to a specific program, can be translated into a logic model that shows specific links among expected inputs, outputs, and outcomes) should be the biggest driver in making decisions about the timing of measuring treatment contrast. If the theory, for example, suggests that almost all the intervention's distinctiveness from business as usual will occur shortly after individuals' entry into the study, efforts to measure treatment contrast should be concentrated in a short period. Alternatively, if the treatment's distinctiveness is likely to unfold over time, or if the distinctiveness is theorized to affect other behaviors that may not manifest themselves for a while, a much longer period should be considered. The theory of change may also suggest measuring different aspects of the treatment contrast at different points in time.

Another factor that should be considered, however, is the researchers' early confidence that a treatment contrast will, in fact, exist eventually. In situations where there is some doubt — for example, when information about business as usual at the outset of the study is limited, or when it is unclear whether staff members in the new intervention will actually be able to adjust their usual practices to be in line with those of the new intervention — it is worth considering conducting a very early exploration of the treatment contrast. In these situations, the very early investigation may indicate that more technical assistance is needed to strengthen the new intervention and bring it in line with its expected practices; that the counterfactual situation in a site or two is so similar to the new intervention that these particular sites should not be included in the evaluation; or, most drastically, that the treatment contrast is very low across all sites and is not repairable with technical assistance, so the impact evaluation should be abandoned.

Study sample coverage

Depending on the data sources used to provide information about the treatment contrast, it may be possible to collect information for all program group and control group members involved in the study or for only a subset of them, either for a random subset or for individuals in particular study cohorts or sites.

The ideal is to collect information for all study sample members or for a random subset of all study sample members. This approach allows the treatment contrast analysis to fully inform the impact analysis. If, for example, positive impacts on the ultimate outcomes of interest (for example, earnings, school graduation rates, or criminal justice system recidivism rates) are found for the late cohorts of study sample members but not for the early cohorts, it would be unfortunate to have treatment contrast information only for the early cohorts. In this

situation, it would be hard to ascertain what drove impacts for the later cohorts but not the early cohorts. In addition, new interventions often mature and improve over time, or are more fully implemented in some sites than in others. Collecting treatment contrast information for just some cohorts or sites will prevent the evaluation from taking full advantage of this maturation process or variation in implementation practices across sites — that is, from being able to truly tie differences in treatment contrast to differences in program impacts. Tying these differences together allows researchers to learn as much as possible about the new intervention and the ways it can be operated most effectively and produce greater impacts when replicated or scaled up.

In some cases, however, it may make sense to restrict the collection of treatment contrast information to certain cohorts or sites. If a new intervention takes a long time to be fully implemented, it may not be worthwhile to examine the treatment contrast for the earliest cohort of enrollees (although the data could be used to ascertain whether treatment contrast increases as the new intervention matures). Another situation could arise if, after covering the primary aspects of the treatment contrast for the whole sample, the study team is able to delve more deeply into a narrow aspect of the treatment contrast. Sometimes that type of substudy can only be conducted for a late, nonrandom cohort but will still provide valuable information about the mechanisms by which the intervention's impacts were likely produced. Similarly, if a certain site simply fails to implement the intervention, it is unlikely to be useful to collect treatment contrast information for that site.

Perspectives from which to collect data

Some new interventions aim to deviate from business as usual by changing the behavior or practices of different types of actors — for example, when classroom teachers present material in a different way, when staff members in welfare-to-work programs refer clients to different types of activities, or when guidance counselors meet with students more frequently or use new methods to advise them. Other new interventions aim to change institutional practices or rules — for example, when community colleges offer different types of classes or structure them in different ways, when employers provide different types of supports to encourage employment retention and advancement among low-wage workers, when public housing authorities allow individuals to save increased earnings that would otherwise be put toward rent increases, or when criminal justice institutions immediately place formerly incarcerated individuals into subsidized jobs when they are released. Additionally, some new interventions seek to alter practices at several levels — for example, a new community college intervention may modify course requirements and course structures and also change the way college advisers conduct their sessions with students. Finally, most interventions ultimately aspire to change the behavior or practices of the clients they serve — for example, by increasing enrollment in and

completion of training by welfare recipients or by increasing course completion among community college students.

Again, the theory of change for the studied intervention should help to suggest the perspective from which treatment contrast information should be collected. The theory will suggest whether most, if not all, change from business as usual is likely to occur at the staff action level as opposed to the institutional level.

But at a minimum, the same perspective (or unit of analysis) that is used in the study's impact analysis should be used in the study's treatment contrast analysis. If the impact analysis is measuring the effects of the intervention on students, then the treatment contrast should — at the least — measure the treatment contrast for students. If the impact analysis is measuring the effects of the intervention on teachers, then the treatment contrast should — at the least — measure the treatment contrast for teachers.

To fully understand what underlies program impacts (or lack of impacts), however, researchers usually need to go beyond measuring the treatment contrast from only the same perspective (or unit of analysis) used in the impact analysis. If a theory of change stipulates that most, if not all, change from business as usual is likely to occur at the staff action level, resources should be expended on documenting and measuring the practices undertaken by staff members working with program group members and the practices undertaken by staff members working with control group members. This information is needed in order for the implementation researchers to understand why there was (or was not) a treatment contrast apparent at the individual or client level in the study. Similarly, if a new initiative involves changes in institutions' practices or rules, then the treatment contrast analysis should collect data on the extent to which the institutions' practices or rules truly differed from business as usual, whether the practices or actions of staff members differed according to program and control group status, and, if the impact unit of analysis is students or clients, whether the practices or actions of the students or clients differed according to program and control group status.

Type of data and methods

Both qualitative and quantitative data, from different sources and collected in a variety of ways, can be used to measure treatment contrast. Observations and semi-structured interviews (with clients, staff members, and administrators), pertaining to both the new intervention and business as usual, can contribute to the treatment contrast analysis. Management information system data, survey data, and administrative records (such as student records) — usually collected on an individual level but also possibly collected on an aggregate level — also can be used in the treatment contrast analysis, provided that they are available for both the new intervention and the counterfactual condition. Data can be collected in a variety of ways, with the proviso that they need to be collected in the same way for the new program as for the

business-as-usual situation. Notably, the data and tools used to measure treatment contrast almost always will need to go beyond those used to measure implementation fidelity.

Summary and Open Questions

This paper echoes other recent exhortations to value the examination of treatment contrast — to increase researchers' understanding of how and why programs make a difference and ultimately to inform program improvement and support "value-added" replication and scaling-up efforts. (See, for example, Knox, Hill, and Berlin, 2018; Manno and Gaubert, 2018; Corrin and Martinez, 2017.) Specifically, the paper sets forth several reasons why it is important to assess treatment contrast in randomized controlled trials: Treatment contrast analysis helps to identify the specific questions that impact evaluations will and will not answer; highlights what program components might and might not be driving a program's effects; suggests what program features might be unable, on their own, to produce effects; suggests why a program's effects might differ across cohorts or subgroups within a site or across sites; and can indicate when technical assistance is needed or when an impact evaluation may not be worthwhile.

Procedurally, the paper suggests proactively starting the planning for the treatment contrast analysis early in studies, having a set treatment contrast measurement plan before program impact results are known, and staying focused on the treatment contrast throughout the study. It also argues that the measurement of treatment contrast should receive as much attention from researchers as does assessing the process of program implementation or measuring treatment fidelity. In addition, it highlights the advantages to focusing first on examining the counterfactual conditions and, subsequently, on the intervention or program of interest. In choosing the aspects of treatment contrast to measure, the paper suggests that the theory of change or logic model for the studied intervention should play the central role. When deciding how to measure treatment contrast, the paper stresses the need to measure treatment contrast dimensions in the same way for program group and control group members.

While this paper has covered a lot of ground, there are a number of treatment contrast issues that the evaluation field has either not widely examined or not resolved. These include the following: Is it necessary for treatment contrast measures to be statistically significantly different for program and control group members? Should treatment contrasts that are found for some types of measures be viewed as more consequential or important than contrasts that are found for other measures? Does what researchers measure, or how they measure it, change according to the most immediate or important use for the treatment contrast findings — for example, identifying technical assistance needs versus interpreting impacts? Finally, what are the advantages and disadvantages of adopting a broad treatment contrast measurement strategy at the start of a study (that is, "covering all possible bases") versus focusing more on measuring specific treatment contrast aspects?

In spite of the above open issues, however, the importance of measuring and understanding the treatment contrast in evaluations of new social policy programs or initiatives is clear. The suggestions and cautions outlined in this paper, which emanate from MDRC's experience in focusing on treatment contrast, are designed to provide ideas for researchers as they continually seek to understand what programs' measured effects do and do not suggest regarding the best ways to improve social programs and policies.

REFERENCES

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

Balu, Rekha, Pei Zhu, Fred Doolittle, Ellen Schiller, Joseph Jenkins, and Russell Gersten. 2015. *Evaluation of Response to Intervention (RtI) Practices for Elementary School Reading*. New York: MDRC.

Cordray, David S., and Georgine M. Pion. 2006. "Treatment Strength and Integrity: Models and Methods." Pages 103-124 in Richard R. Bootzin and Patrick E. McKnight (eds.), *Strengthening Research Methodology: Psychological Measurement and Evaluation.* Washington, DC: American Psychological Association.

Corrin, William, and John Martinez. 2017. "Social Progress's True Metric: Compared to What?" Stanford Social Innovation Review. Website: https://ssir.org/articles/entry/social_progresss_true_metric_compared_to_what.

Corrin, William, Susan Sepanik, Rachel Rosen, and Andrea Shane. 2016. *Addressing Early Warning Indicators: Interim Impact Findings from the Investing in Innovation (i3) Evaluation of Diplomas Now*. New York: MDRC.

Gueron, Judith M., and Howard Rolston. 2013. *Fighting for Reliable Evidence*. New York: Russell Sage Foundation.

Hamilton, Gayle. 2012. *Improving Employment and Earnings for TANF Recipients*. Washington, DC: The Urban Institute.

Hamilton, Gayle, Stephen Freedman, Lisa Gennetian, Charles Michalopoulos, Johanna Walter, Diana Adams-Ciardullo, Anna Gassman-Pines, Sharon McGroder, Martha Zaslow, Jennifer Brooks, Surjeet Ahluwalia. 2001. *National Evaluation of Welfare-to-Work Strategies: How Effective Are Different Welfare-to-Work Approaches? Five-Year Adult and Child Impacts for Eleven Programs.* Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families and Office of the Assistant Secretary for Planning and Evaluation; and U.S. Department of Education, Office of the Deputy Secretary, Planning and Evaluation Service, and Office of Vocational and Adult Education.

Hamilton, Gayle, and Daniel Friedlander. 1989. *Final Report on the Saturation Work Initiative Model in San Diego*. New York: MDRC.

Hamilton, Gayle, and Susan Scrivener. 2012a. *Facilitating Postsecondary Education and Training for TANF Recipients*. Washington, DC: The Urban Institute.

Hamilton, Gayle, and Susan Scrivener. 2012b. *Increasing Employment Stability and Earnings for Low-Wage Workers: Lessons from the Employment Retention and Advancement (ERA) Project*. New York: MDRC.

Harknett, Kristen, Michelle S. Manno, and Rekha Balu. 2017. *Building Bridges and Bonds: Study Design Report.* OPRE Report 2017-17. Washington, DC: Office of Planning, Research, Administration for Children and Families, U.S. Department of Health and Human Services.

Hendra, Richard, Keri-Nichole Dillman, Gayle Hamilton, Erika Lundquist, Karin Martinson, and Melissa Wavelet. 2010. *How Effective Are Different Approaches Aiming to Increase Employment Retention and Advancement? Final Impacts for Twelve Models*. New York: MDRC.

Holland, Paul. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81, 396: 945-960.

Knox, Virginia, Carolyn Hill, and Gordon Berlin. 2018. *Can Evidence-Based Policy Ameliorate the Nation's Social Problems?* New York: MDRC.

Manno, Michelle S., and Jennifer Miller Gaubert. 2018. "How Early Implementation Research Can Inform Program Scale-Up Efforts." *Implementation Research Incubator* (blog), January. New York: MDRC. Website: https://www.mdrc.org/publication/how-early-implementation-research-can-inform-early-scale-efforts.

Manno, Michelle S., Edith Yang, and Michael Bangser. 2015. *Engaging Disconnected Young People in Education and Work: Findings from the Project Rise Implementation Evaluation*. New York: MDRC.

Martinson, Karin, and Richard Hendra. 2006. *The Employment Retention and Advancement Project: Results from the Texas Site*. New York: MDRC.

Mattera, Shira Kolnik, Robin Tepper Jacob, and Pamela Morris. 2018. *Strengthening Children's Math Skills with Enhanced Instruction: The Impacts of Making Pre-K Count and High 5s on Kindergarten Outcomes*. New York: MDRC.

Miller, Cynthia, Lawrence F. Katz, Gilda Azurdia, Adam Isen, and Caroline Schultz. 2017. *Expanding the Earned Income Tax Credit for Workers Without Dependent Children: Interim Findings from the Paycheck Plus Demonstration*. New York: MDRC.

Morris, Pamela, Shira Kolnik Mattera, and Michelle Maier. 2016. *Making Pre-K Count: Improving Math Instruction in New York City.* New York: MDRC.

Quint, Janet, Pei Zhu, Rekha Balu, Shelley Rappaport, and Micah DeLaurentis. 2015. *Scaling Up the Success for All Model of School Reform: Final Report from the Investing in Innovation (i3) Evaluation*. New York: MDRC.

Rubin, Don. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Education Psychology* 66: 688-701.

Rubin, Don. 1978. "Bayesian Inference of Causal Effects: The Role of Randomization." *Annals of Statistics* 6, 1: 34-58.

Schweinhart, Lawrence J. 2004. *The High/Scope Perry Preschool Study Through Age 40: Summary, Conclusions, and Frequently Asked Questions.* Ypsilanti, MI: High/Scope Press.

Scrivener, Susan, Michael J. Weiss, Alyssa Ratledge, Timothy Rudd, Colleen Sommo, and Hannah Fresques. 2015. *Doubling Graduation Rates: Three Year Effects of CUNY's Accelerated Study in Associate Programs (ASAP) for Developmental Education Students*. New York: MDRC.

Sepanik, Susan, William Corrin, David Roy, Aracelis Gray, Felix Fernandez, Ashley Briggs, and Kathleen K. Wang. 2015. *Moving Down the Track: Changing School Practices During the Second Year of Diplomas Now*. New York: MDRC.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton Mifflin Company.

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.

About MDRC

MDRC is a nonprofit, nonpartisan social and education policy research organization dedicated to learning what works to improve the well-being of low-income people. Through its research and the active communication of its findings, MDRC seeks to enhance the effectiveness of social and education policies and programs.

Founded in 1974 and located in New York; Oakland, California; Washington, DC; and Los Angeles, MDRC is best known for mounting rigorous, large-scale, real-world tests of new and existing policies and programs. Its projects are a mix of demonstrations (field tests of promising new program approaches) and evaluations of ongoing government and community initiatives. MDRC's staff members bring an unusual combination of research and organizational experience to their work, providing expertise on the latest in qualitative and quantitative methods and on program design, development, implementation, and management. MDRC seeks to learn not just whether a program is effective but also how and why the program's effects occur. In addition, it tries to place each project's findings in the broader context of related research — in order to build knowledge about what works across the social and education policy fields. MDRC's findings, lessons, and best practices are shared with a broad audience in the policy and practitioner community as well as with the general public and the media.

Over the years, MDRC has brought its unique approach to an ever-growing range of policy areas and target populations. Once known primarily for evaluations of state welfare-to-work programs, today MDRC is also studying public school reforms, employment programs for exprisoners, and programs to help low-income students succeed in college. MDRC's projects are organized into five areas:

- Promoting Family Well-Being and Children's Development
- Improving Public Education
- Raising Academic Achievement and Persistence in College
- Supporting Low-Wage Workers and Communities
- Overcoming Barriers to Employment

Working in almost every state, all of the nation's largest cities, and Canada and the United Kingdom, MDRC conducts its projects in partnership with national, state, and local governments, public school systems, community organizations, and numerous private philanthropies.