MDRC Working Papers on Research Methodology

A Conceptual Framework for Studying the Sources of Variation in Program Effects

Michael J. Weiss
Howard S. Bloom
Thomas Brock

June 2013

Acknowledgments

The ideas, writing, editing, and review of this paper involved a great many people who vastly improved its quality. In particular, we would like to acknowledge the contributions of MDRC staff members Gordon Berlin, Ginger Knox, Shira Mattera, James Riccio, and Marie-Andrée Somers, and consultant Kay Sherwood. They provided guidance on the ideas addressed by the paper, suggested many of the examples that are used to illustrate these ideas, and in these ways and others added depth to the paper. A special thanks to MDRC's Caitlin Platania who provided significant support developing figures, editing and formatting text, checking references, and a long list of other miscellaneous tasks; her contributions were critical to completing this paper. We also would like to thank the William T. Grant Foundation for its support. Not only did the Foundation fund this work but its president, Robert C. Granger, and program officer, Kim DuMont, provided meticulous feedback and valuable advice on early versions of the paper. In addition, we thank Joseph Durlak, Naomi Goldstein, Jim Kemple, Mark Lipsey, Steven Raudenbush, and Lauren Supplee for reviewing and commenting on earlier drafts of this work. That said, all positions taken in this paper and errors that it might contain are solely the responsibility of its authors.

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, The George Gund 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 Web site: www.mdrc.org.

Abstract

Evaluations of public programs in many fields reveal that (1) different types of programs (or different versions of the same program) vary in their effectiveness, (2) a program that is effective for one group of people might not be effective for other groups of people, and (3) a program that is effective in one set of circumstances may not be effective in other circumstances. This paper presents a conceptual framework for research on such variation in program effects and the sources of this variation. The framework is intended to help researchers — both those who focus mainly on studying program implementation and those who focus mainly on estimating program effects — see how their respective pieces fit together in a way that helps to identify factors that explain variation in program effects and thereby support more systematic data collection on these factors. The ultimate goal of the framework is to enable researchers to offer better guidance to policymakers and program operators on the conditions and practices that are associated with larger and more positive effects.

Contents

List of Exhibits

Table

Figure

Introduction

Recent advances in evaluation research have greatly increased the number of high-quality studies of the causal effects produced by public programs in many different fields. Therefore, much is now being learned about the average effects of specific programs for specific groups. As this evidence base has grown, so has the realization that (1) different types of programs (or different versions of the same program) vary in their effectiveness, (2) a program that is effective for one group of people might not be effective for other groups of people, and (3) a program that is effective in one set of circumstances might not be effective in other circumstances. Thus, average program effects may not tell the whole story. With this in mind, a small number of studies have been conducted to examine sources of variation in program effectiveness. For example:

- A meta-analysis of past studies of the effects of 69 after-school programs on social, behavioral, and academic outcomes for young people identified a subgroup of programs that had large favorable effects, while in stark contrast finding no identifiable effects for other types of programs.[1] Programs that were found to be effective shared four characteristics: a sequenced approach to program activities, an emphasis on active learning strategies, a focus on a limited number of goals, and activities that were explicitly tied to these goals (Durlak, Weissberg, and Pachan, 2010).

- Based on a lottery that was used to select students for oversubscribed charter schools that were run by a charter management organization (CMO), a national study found that the effects of 68 schools on student reading and math scores varied substantially.[2] Favorable program effects were associated with two school practices: establishment of clear behavior standards, such as zero tolerance for negative behavior, and an emphasis on intensive coaching for new teachers (Lake et al., 2012).

- Randomized trials of welfare-to-work programs in 59 local welfare offices from seven states produced estimates of average program effects on participants' earnings during their first two years after entering the programs that range from negative (-$1,412) to positive ($4,217). The study also revealed that, other things being equal, program offices that adopted certain practices — such as counseling clients to obtain jobs quickly or offering

[1]The Winsorized study level effects ranged in value from -0.16 to +0.85.

[2]Nonexperimental analyses were used to estimate the effects of those CMO schools that were not oversubscribed.

clients personalized attention — had earnings effects that were much larger than average. In addition, the study found that the context of the programs also mattered. Specifically, other things being equal, programs in communities with little unemployment had larger earnings effects than did programs in communities with substantial unemployment (Bloom, Hill, and Riccio, 2003).[3]

These studies have obvious value to policymakers and practitioners who want to know why some programs are more effective than others and what it might take to design and operate more successful programs. Unfortunately, researchers often struggle to answer such questions. Although scholars have argued for decades that to fully understand programs one must know how their effects vary (for example, Bryk and Raudenbush, 1988; Heckman, Smith, and Clements, 1997; Friedlander and Robins, 1997; Raudenbush and Liu, 2000; Heckman, 2001; Abadie, Angrist, and Imbens, 2002; Bitler, Gelbach, and Hoynes, 2006; and Djebbari and Smith, 2008), to date evaluation research and public policy analysis have mainly focused on the average effects of programs and paid far less systematic attention to explaining variation in effects.

To help rectify this situation, this paper presents a conceptual framework for designing and interpreting research on variation in program effects and the sources of this variation. The framework is intended to help researchers integrate the study of variation in program effectiveness, treatment contrasts (defined later), treatment fidelity, and program implementation. The first section of the paper defines core concepts about program effects and introduces the proposed framework. The second section uses the framework to describe the proximal sources of variation in program effects. The third section uses the framework to examine the roles of more distal sources of variation in program effects such as program plans, local service delivery organizations, and other factors that determine program implementation. The fourth section presents concluding thoughts.

Throughout this paper we include concrete empirical examples to illustrate points. Despite efforts to include a wide array of examples from different fields of research conducted by different researchers, we rely disproportionately on projects conducted by MDRC since these are the programs and evaluations that we know best. Nonetheless, we expect that the issues we raise apply to many other programs and policies and hope that readers will think of how this framework relates to their own examples and experiences.

[3]Dorsett and Robins (2011) conducted a related study of factors that predict effects of welfare-to-work programs in the United Kingdom.

Section 1

Defining Concepts and Introducing the Framework

Definitions of Program Effects and Program Effect Variation

Before proceeding it is important to carefully define what is meant by a program effect for a person, an average program effect for a group, and variation in program effects across individuals and groups. The definition of a program effect for a person that we and most other program evaluation researchers use comes from the statistical literature on causal effects and is based on the concept of "potential outcomes."[4] Potential outcomes for a person are the outcomes that the person would have under a different set of experiences or conditions.

Consider the causal effect of *assigning* someone to a specific program. We refer to this as the causal effect of a program *offer*.[5] In defining this effect, it is typically assumed that each person has two potential outcomes: (1) that which he or she would experience if assigned to (offered) the program and (2) that which he or she would experience if not assigned to (not offered) the program.[6] The first of these is referred to as the "treated outcome" and the second as the "untreated counterfactual outcome" or "counterfactual" for short.[7]

For example, consider the causal effect on future earnings of assigning an eligible youth to an employment and training program that provides job-search assistance, basic education, classroom occupational skills training, and on-the-job training. What is the effect of this program offer? By definition it is the *difference* between two potential outcomes for the youth: (1) the youth's future earnings if he or she were assigned to the program and (2) the youth's future earnings if he or she were not assigned to the program. This difference represents the "value added" by the program offer over the counterfactual state of the world.[8]

---

[4]Versions of the potential outcomes framework have been attributed to Neyman (1923), Fisher (1935), Roy (1951), Quandt (1972), Rubin (1974, 1978), Holland (1986), and Heckman (2001, 2005).

[5]In the statistics literature this is typically referred to as the effect of intent to treat, or ITT (Angrist, Imbens, and Rubin, 1996).

[6]The existence of only two potential outcomes assumes that each person's outcomes are independent of others' program assignment and outcomes. The statistics literature often refers to this condition as the stable unit treatment value assumption, or SUTVA (Rubin, 1986).

[7]Strictly speaking, both potential outcomes are counterfactuals. However, it is common practice in evaluation research to call the untreated outcome the counterfactual because this is what someone who is assigned to a program would have experienced if — counter to the fact of actual assignment — he or she had not been assigned to the program.

[8]The potential outcomes framework applies equally well to outcomes that are features of settings such as schools, classrooms, after-school programs, or neighborhoods. For example, one might be interested in estimating the effect of an intervention on the quality of after school programs. For a given after school program, this effect is by definition the difference between (1) the quality of the program with the intervention and (2) the quality of the program without the intervention. For simplicity, our discussion focuses on program effects on individuals rather than program effects on settings.

Unfortunately, it is not possible to observe both potential outcomes for a person simultaneously because we can only experience one outcome at a time. Consequently, it is not possible to observe a program effect for a person. What is possible, however, is to observe the average outcome for a sample of people who are offered a program (its program group or treatment group) and the average outcome for a sample of people who are not offered the program (its control group or comparison group). If these two groups are the same on average in all ways before program assignment, then the observed difference in their average future outcomes is an unbiased estimate of the *average* effect of their program offer.

A randomized trial is the best way to produce a program group and control group that are initially the same on average in all ways (or at least not *systematically* different). The larger the sample for such a trial is, the more similar these groups will tend to be. Thus, when feasible, randomizing a large sample of eligible persons to receive a program offer or to not receive a program offer and comparing the average future outcomes of each group is the best way to obtain an unbiased estimate of the average effect of a program offer. However, it is also possible for strong quasi-experiments to approach the rigor of a randomized trial (Cook, Shadish, and Wong, 2008). Regardless of how these effects are estimated, they compare potential outcomes under two "treatment conditions": (1) access to services from the program being offered plus any other existing services and (2) access only to other existing services.[9,10]

To this point we have defined an *individual* program effect for a person (which exists in principle but cannot be observed in practice) and an *average* program effect for a group (which exists in principle and can be estimated in practice). The next step is to define what is meant by *variation* in program effects across groups of persons who differ in their background characteristics, geographic location, and/or when they enter a study. These differences are typically represented in terms of client subgroups and program sites. Client subgroups can be defined in terms of individual background characteristics such as demographics, past outcomes, and temporal cohorts. Program sites can be identified by factors such as size, geographic locations, and administrative units. It is then possible to apply the preceding definition of an

[9]If a randomized trial compares two different programs, then the average *differential* effect of one program versus the other equals the difference in their average potential outcomes.

[10]It is often the case that some persons who are offered a program do not receive it and some persons who are not offered a program do receive it. In these situations the causal effect of a program offer is not the same as the causal effect of program *receipt*. Nonetheless, the logical basis for defining these causal effects is the same. They both compare two potential outcomes, only one of which can be observed for a given person. For the causal effect of program receipt the two potential outcomes are: (1) that which would be experienced if the program were received and (2) that which would be experienced if the program were not received. In practice, estimating the causal effect of program receipt is more difficult than estimating the causal effect of a program *offer*. However, both types of estimates can often be obtained from a well-executed randomized trial or strong quasi-experiment. For simplicity this paper does not consider the more complex analysis involved in obtaining estimates of the average effects of program receipt.

average effect of a program offer to a client subgroup or a program site. Variation in these average effects across subgroups and/or sites is what we refer to as variation in program effects, or "effect variation" for short.[11] The next section of the paper discusses the sources of effect variation in detail, but first we provide an overview of our proposed framework.

Proposed Framework

Figure 1 illustrates our proposed framework, which we discuss starting with program effects and working backward (from right to left) to program implementation. This order reflects our emphasis on explaining variation in program effects. Throughout the paper we refer to factors on the right-hand side of the figure as being "downstream" in the causal pathway that runs from program implementation to program effects. We refer to factors on the left-hand side as being "upstream."

This framework represents a rigorous evaluation of a program's effects where the effects are estimated by comparing *outcomes* for a program group with outcomes for a control group (in an experiment) or for a comparison group (in a quasi-experiment). We sometimes refer to the outcome of interest for an evaluation as its "target outcome" to distinguish it from possible intermediate outcomes. For example, the target outcome for an employment and training program might be average future earnings, whereas an intermediate outcome might be the rate of client participation in job-search assistance. [12] The difference between average target outcomes for the two study groups is an estimate of the average effect of the program offer for them (labeled as *program effect*).

Two boxes to the left of the outcomes for sample members (skipping mediators, for the moment) is the *treatment they receive* with access to the program (for program group members) and without access to the program (for control or comparison group members). We refer to the difference between the average treatment received with and without access to the program as the *treatment contrast*. An intermediate effect of the program offer, the treatment contrast is also the cause of program effects — a point that we believe cannot be overemphasized and that we return to frequently.[13]

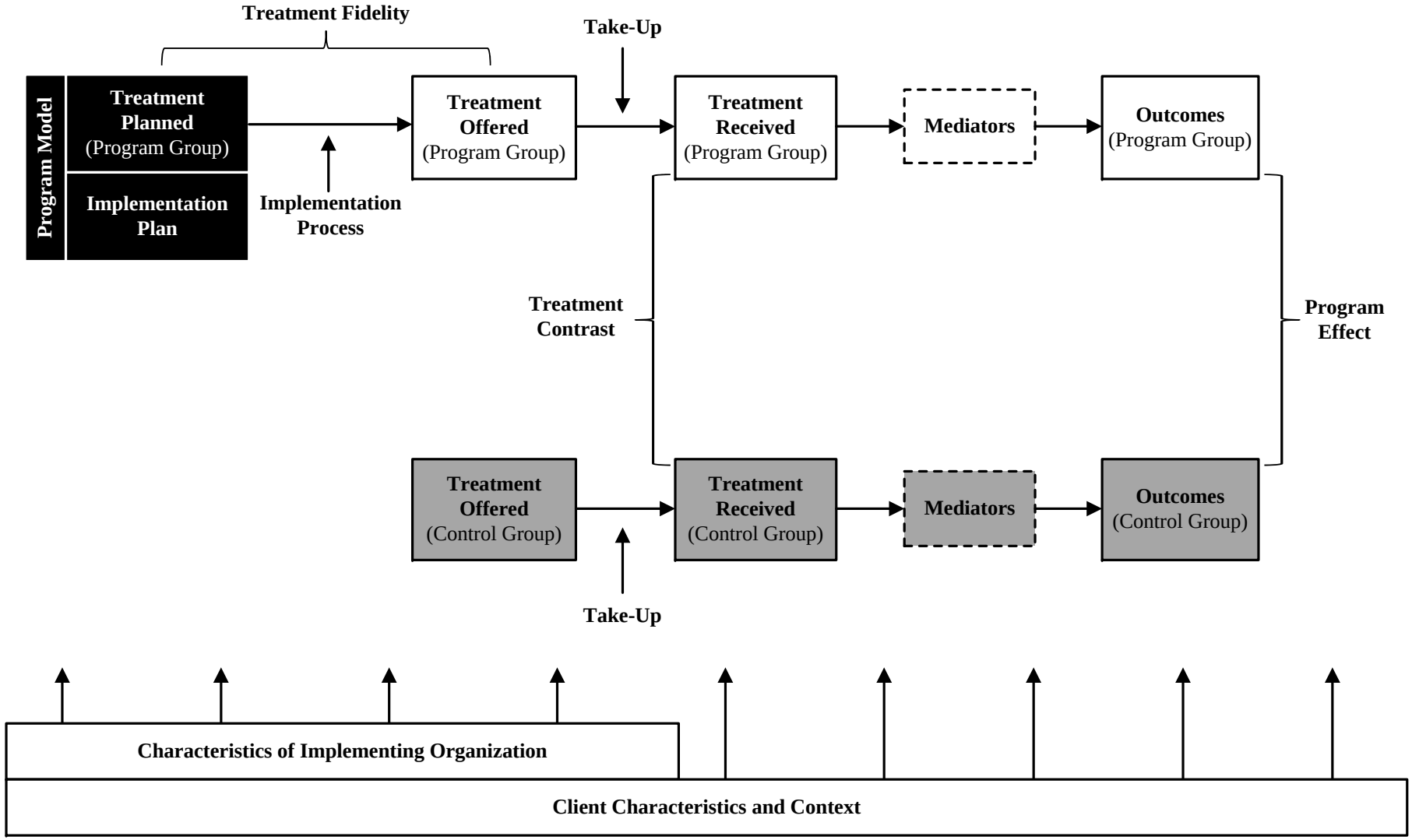
[11]For further discussion of cross-site variation in program effects, see Bloom, Raudenbush, and Weiss (under review). For further discussion of subgroup differences in program effects, see Bloom and Michalopoulos (2011).

[12]In evaluation research this intermediate outcome is sometimes referred to as an "output."

[13]At level of the individual, which cannot be observed, the treatment contrast is the direct cause of the program effect. By contrast, and for technical reasons beyond the scope of this paper, at the group level, which is what would be observed by a randomized trial or strong quasi-experiment, the average treatment contrast is not *always* the cause of the average program effect. For details on why this is the case, see discussions by Reardon and Raudenbush (forthcoming) and Reardon, Unlu, Zhu, and Bloom (2012) about "compliance/effect bias."

Figure 1

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



Between the received treatments and the target outcomes are *mediators*. Many treatment contrasts do not directly cause changes in target outcomes. Rather, the treatment contrast (for example, experiencing an HIV advertising campaign compared with not experiencing such a campaign) affects a mediator (for example, awareness), and the mediator, in turn, affects the target outcomes (for example, number of sexual partners). In this way, mediators are part of a chain of events that lead to program effects.

Continuing upstream, to the immediate left of treatment received is the *treatment offered* with access to the program (for program group members) and without access to the program (for control or comparison group members). Here the treatment offered refers to the services or experiences available with or without access to a program. The link between treatment offered and treatment received represents service *take-up* — an important subject of evaluation research (for example, Bloom and Bloom, 1981; Madrian and Shea, 2001; Bettinger et al., 2009).

The two black boxes at the far left of Figure 1 represent: (1) the treatment or services that are planned or intended for program group members (*treatment planned*) and (2) the plan for implementing the treatment (*implementation plan*). Together, these two boxes represent the *program model*.[14] A program model is sometimes specified by the program's developers.[15] It comprises a blueprint for client services (treatment plan) and a blueprint for implementing these services (implementation plan). A program's implementation plan is adopted (and might be adapted or changed) by an implementing organization to varying degrees based on the context and the site's *organizational characteristics* (see bottom of Figure 1), as well as the specificity of the implementation plan. The result is an enacted *implementation process* that allows a program to be made operational.

This implementation process influences the services that are offered and how they are delivered. These, in turn, influence the treatment that is received by program clients. We refer to the relationship between the treatment that is planned for clients and the treatment that is offered or made available to its clients as *treatment fidelity* — a measure, that is, of fidelity to the intended plan. Other researchers have referred to this construct as "intervention fidelity," "treatment integrity," or "program fidelity," and it is sometimes defined as the difference between planned treatment and received treatment, rather than the difference between planned treatment and offered treatment (for example, Dane and Schneider, 1998; Cordray and Pion, 2006; Hulleman and Cordray, 2009; Carroll et al., 2007; and Durlak and DuPre, 2008). The decision of whether to extend treatment fidelity from planned client services to offered client

[14]Program models may also specify a target population, which is not identified separately in the figure.

[15]As noted later, it is sometimes important to distinguish between the program model that is specified by model developers (for example, a national model in a scale-up effort) and the program model that is specified by a given site (a local model).

services or from planned client services to received client services is not critical; however, appreciating the relationship between services that are planned, offered, and received can be extremely important for program development and implementation and for interpreting the effects of a program.

Finally, at the bottom of Figure 1 are two boxes that represent factors that influence or "moderate" the causal relationships specified in the diagram. The box on top represents characteristics of the local organization responsible for implementing a program. These organizational characteristics are generally hypothesized to moderate many facets of program implementation. This box thus spans the portion of the diagram that involves program implementation. The box on bottom represents characteristics of a program's clients as well as characteristics of its social, physical, economic, financial, and political context. These characteristics are typically hypothesized to moderate every aspect of the program process: from planning and implementation to treatment offered and received, to the program's treatment contrast, and, ultimately, to its effects on client outcomes. This box thus spans the entire diagram. Although for simplicity Figure 1 minimizes the visual appearance of these moderators, this is not meant to indicate that they are less important than other factors.

So far we have defined program effects and program effect variation and given a broad overview of the conceptual framework we've developed to integrate the study of variation in program effectiveness, treatment contrasts, and implementation. The following section examines how variation in one feature of the framework is related to variation in other features. This discussion proceeds from right (downstream) to left (upstream) in the framework and "zooms in" on each component as it is discussed.

Section 2

Proximal Sources of Variation in Program Effects

For decades, policymakers, practitioners, and researchers have hypothesized about factors that influence program effects. To help understand how these influences work, leading scholars like Peter Rossi and Carol Weiss have urged that carefully constructed program theory guide the study of them.[16] Having read many descriptions of program theories, listened to many discussions about them, and tried to construct and use them in practice, we offer the following categorization of sources of variation in program effects as a guide for evaluation practice.

Our discussion begins at the *point of service* for program clients (their treatment receipt with or without access to a program), and continues to the right through to their target outcomes. Thus, when we talk about sources of variation in program effects, we mean proximal sources that come into play from the point of service forward. We do not mean factors that lie further upstream, even though they may determine what the point-of-service experience is like. These more distal factors are discussed later.

From the point of service, all proximal sources of variation in program effects can be grouped into three categories, which we refer to as the "three Cs":

1) Treatment Contrast
2) Client Characteristics
3) Program Context

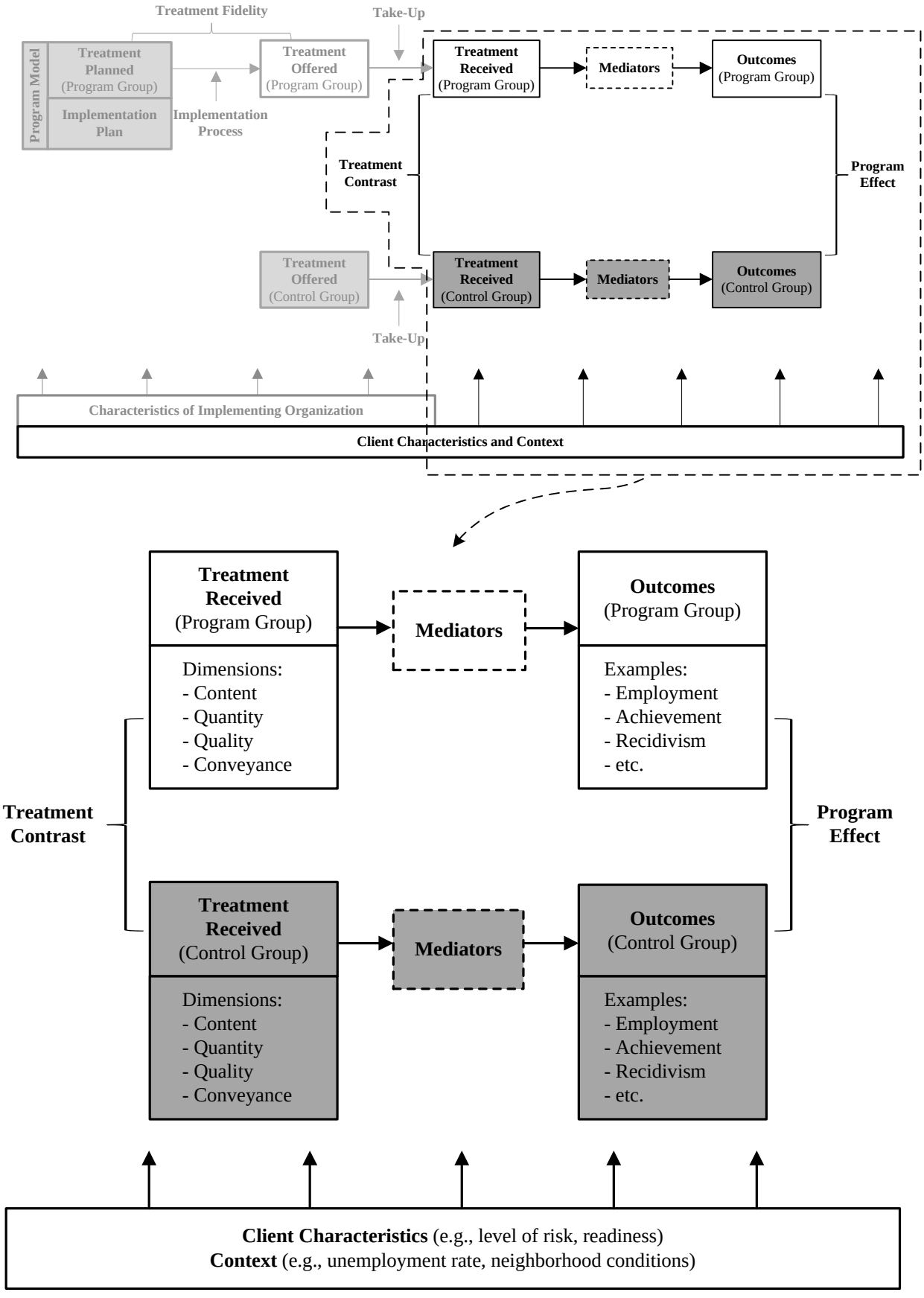
The client treatment contrast mediates or causes program effects, while client characteristics and program context moderate the size and/or sign of these effects.

Figure 2 zooms in on one part of our framework to highlight the relationship between a program's treatment contrast, its mediators, and its effects on client outcomes, including the moderation of these effects by client and contextual characteristics. To make things more concrete this expanded portion of our framework lists examples of target outcomes, includes a heuristic guide to key features of program-related services, and provides examples of potential client and contextual moderators. Below we discuss the three proximal sources of variation in program effects in detail.

[16] See Chen and Rossi (1983) and Weiss (1997).

Figure 2

The Treatment Contrast and Effect



Treatment Contrast

Linking Treatment Contrasts and Program Effects

As has been noted, a program effect is the difference between two potential *target outcomes*: (1) that which occurs with access to program services and any other existing services versus (2) that which occurs with access only to other existing services. A program effect is thus the value-added by one "treatment package" relative to another. Likewise, a treatment contrast[17] is the difference between two potential intermediate outcomes: (1) receipt of program services plus other existing services and (2) receipt of other existing services only. A program can change clients' outcomes only by changing their treatment receipt and thereby producing a treatment contrast. If a treatment contrast does not exist — that is, if clients have the exact same experience in a program as they would have had had they not been in the program — then there cannot be a program effect. Thus a treatment contrast is *necessary* for a program effect to occur. On the other hand, the mere existence of a treatment contrast does not guarantee a program effect. Consequently, a treatment contrast is not *sufficient* for a program effect to occur.

Consider the following hypothetical example of the relationship between treatment contrasts and program effects. A large high school currently requires its eleventh-graders to meet with a guidance counselor for one 30-minute session during the first week of school in order to plan their course schedules and review their educational goals. Any further sessions are optional. To increase graduation rates, the school is considering a new program that would require an additional 30-minute guidance session at midyear, with no special efforts by teachers to induce students to attend additional counseling sessions.

In our proposed framework, the treatment package offered by the new program would be described as including two required counseling sessions and any optional ones that students desire. (This assumes the school successfully makes these services available.) The treatment package offered without the program includes the initial required session and any desired optional ones. The difference in services offered to clients by these two packages is one required 30-minute, midyear counseling session.

Consider how this difference in treatments offered becomes a difference in treatments received (a treatment contrast). Some students might only attend counseling sessions that are required. Under the existing system these students would attend one 30-minute session, whereas under the new program they would attend two 30-minute sessions. Their treatment contrast is thus one additional 30-minute session. Other students might feel a greater need for guidance counseling. Under the existing system they would attend the initial required session

[17]Our discussion of a program's treatment contrast is similar to the discussion of what Cordray and Pion (2006) and Hulleman and Cordray (2009) refer to as a program's "achieved relative strength."

and two optional ones; under the new program they would attend the two required sessions and one optional session. Hence, the new program would produce no treatment contrast for them. Still other students might be so resistant to guidance counseling that they would not attend any sessions under either system. The new program would thus produce no treatment contrast for them.

Because an individual student can experience only one of these treatment packages, it is not possible to observe an individual treatment contrast. However, it is possible to observe services received by a group of students assigned at random to a new program and services received by another group assigned at random to the existing system. The average difference in treatment received by the two groups is an unbiased estimate of their average treatment contrast. For example, if students assigned to the new program attended 2.6 sessions on average and students assigned to the existing system attended 1.5 sessions on average, their average treatment contrast would be an additional 1.1 sessions.[18]

The subsequent difference in graduation rates for the two groups is an unbiased estimate of the average causal effect on a target outcome of the new program relative to the existing system. If within two years, 67 percent of the students assigned to the new program graduated and 66 percent of the students assigned to the existing program did so, the best estimate of the program's effect on graduation rates is an increase of 1 percentage point.[19]

Consider a second large high school testing the same program. Its teachers make a concerted effort to get students who are assigned to the new program to attend as many counseling sessions as possible but make no special effort for other students. Students assigned to the new program attend 5.5 counseling sessions and students assigned to the existing system attend 1.3 sessions, for an average service contrast of 4.2 counseling sessions. Subsequent graduation rates are 73 percent for the former group and 62 percent for the latter group, for an average program effect of 11 percentage points.

At a third large high school testing the new program, teachers vigorously promote attendance at guidance counseling sessions for all students (perhaps because control over the evaluation here was weak). Students assigned to the new program attend 4.0 sessions on average and students assigned to the existing system attend 3.8 sessions on average, for an average service contrast of 0.2 sessions. Subsequent graduation rates are 70 percent for the first group and 69 percent for the second group, for a program effect of 1 percentage point.

[18]If service contrasts were observed at multiple program sites it would be possible to directly estimate cross-site variation in their service contrasts.

[19]One could be confident about the findings for this high school and those for the two others described below if each were based on a large sample of students.

The substantial variation across schools in their average treatment contrast and average program effects and the strong positive association between these two factors provides prima facie evidence of the effectiveness of additional guidance counseling for eleventh-graders. This evidence would be even stronger if it were based on a large number of schools. On the other hand, if the second high school, which had a large treatment contrast, experienced a negligible program effect, this lack of a pattern of results would provide prima facie evidence that guidance counseling for eleventh-graders is not by itself necessarily effective, other things being equal. In this way the relationship between variation across schools in their treatment contrast and variation across schools in their effects can provide *suggestive* evidence about the effect of the intended treatment on the target outcome.

The preceding example illustrates the potential value of using a multisite trial to compare *natural* cross-site variation in program treatment contrasts with natural cross-site variation in program effects.[20] If such a trial produced prima facie evidence that increasing the treatment contrast increased program effects, one might attempt to confirm this hypothesis by assigning clients at random to *planned* variation in service contrasts.

Identifying a Treatment Contrast

A treatment contrast comprises at least four important dimensions:

Content: What services are provided?
Quantity: How much of each service is provided?
Quality: How well is each service provided?
Conveyance: How, when, and by whom is each service provided?[21]

Although these dimensions can overlap, we believe that they provide a useful checklist. We are thus less concerned about splitting hairs over categorizing each program component into these four dimensions and more concerned about identifying the contrast as comprehensively as possible. We cannot overemphasize that the treatment contrast reflects the *difference* in services experienced under the program and counterfactual condition, so content, quantity, quality, and conveyance must be measured for program group members as well as for the counterfactual condition (that is, for control group members in an experiment). Below we discuss each dimension in more detail and with examples.

[20]Bloom, Hill, and Riccio (2003) provide a detailed example of such an analysis based on an unusually large data set for three large-scale multisite trials.

[21]Hong and Raudenbush (forthcoming) provide a theoretical framework for considering the causal influence of agents who provide program services (for example, teachers, therapists, and employment counselors).

Content

By treatment content we mean the features, components, or ingredients of a service package that are a program's basic building blocks. For example, services to increase student academic achievement in preschool through postsecondary educational settings or in after-school or out-of-school programs might include new curricula (see Klein et al., 2008; Agodini and Harris, 2010). Examples from other program areas include: cognitive behavioral therapy for correctional inmates to reduce their violent behavior (Lipsey, Landenberger, and Wilson, 2007); informational campaigns to educate teenagers about the risks of alcohol and substance abuse, unprotected sexual activity, or smoking (for example, Flay et al., 2004); home visiting services to teach low-income mothers and pregnant women how to improve the health and development of their infants and young children (Olds et al., 2007; Duggan et al., 2007; Paulsell, Avellar, Martin, and Del Grosso, 2010; Gomby, 2005); and financial incentives for students and their parents to promote better health care and increased educational engagement (Riccio et al., 2010).[22]

A concrete example of the content of one element of a treatment contrast comes from a random assignment evaluation of the City University of New York's Accelerated Study in Associate Programs (ASAP). One component of this multifaceted program is "comprehensive advisement." In a survey administered to program and control group members, study participants were asked what topics had been covered during advising sessions (for example, academic goals, course selection, requirements for graduation, job opportunities, or personal matters.). While simplistic, these questions reveal the proportion of students who covered each topic as part of ASAP and approximately what proportion would have covered each topic in the absence of ASAP (Scrivener, Weiss, and Sommo, 2012).

Quantity

The notion of the "quantity" of treatment received, or how much services are received, is often described in terms of "dose" or "exposure" to the treatment (Dane and Schneider, 1998; Cordray and Pion, 2006; Hulleman and Cordray, 2009). Similarly, we define treatment quantity in terms of the *prevalence*, *frequency*, *intensity*, and *duration* of services that are received.

Treatment prevalence is the percentage of clients who receive that treatment (also sometimes described as a program's "reach"). For example, if 90 percent of students in our hypothetical high school counseling program received some counseling, this is the prevalence of counseling services under the treated condition. If 80 percent of the students received

[22]To address the reality that most programs involve multiple services, based on engineering principles and fractional factorial randomized trials, researchers are developing systematic ways to choose an effective service mix (for example, Collins, Murphy, Nair, and Strecher, 2005; Box, Hunter, and Hunter, 1978).

counseling under the existing system, this is the prevalence of counseling under the untreated counterfactual condition or status quo.

The frequency of services represents how often they are received during any given period (a day, a week, a month, a year). For example, tutoring in English language arts might be provided twice a week, cognitive behavioral therapy might be provided daily, activities to prevent risk behaviors might be provided monthly, or home visits to new mothers might be provided several times a month.

By "treatment intensity" we mean the length of a typical service session. According to this definition, other things being equal, a client service consisting of 15-minute sessions is less intensive than a client service consisting of 60-minute sessions. In our high school counseling example, each session was 30 minutes long. This provides a certain sustained amount of time for students and counselors to interact and to explore issues that concern them. Note that by treatment intensity we do not mean the emotional or psychological intensity of an interaction between clients and service providers. We categorize this construct under treatment quality and discuss it below.[23]

Lastly, the duration of services received is the total period of time during which they are received. For example, in Reading First, enhanced reading classes are supposed to continue from kindergarten through third grade (Gamse et al., 2009).

Quality

Treatment quality is perhaps the most elusive dimension of a treatment package. In general, treatment quality refers to how effectively the critical elements of a program are delivered to clients. The basic idea is that quality services create effective interactions between clients and service providers, promote a high level of client engagement and responsiveness, accurately convey the information they are supposed to convey, stimulate deep personal reflection by clients (especially for therapeutic interventions), get delivered on a timely and predictable basis, and so on. The assessment of treatment quality thus tends to be more subjective than the assessment of other program features. In fact, treatment quality is often largely "in the eye of the beholder." This is so much the case that widely used assessments of treatment quality rely on direct observation by trained raters.

For example, the Classroom Assessment Scoring System (CLASS) (Pianta et al., 2008) assesses the quality of socioemotional and instructional interactions between teachers and students. Originally created for preschools, it was later extended to elementary and high

[23]Again, we care less about where service features are categorized and more about whether they are identified.

schools. The preschool CLASS has 10 subscales from 3 domains: (1) *emotional support* (positive climate, negative climate, teacher sensitivity, and regard for student perspectives), (2) *classroom organization* (behavior management, productivity, and instructional formats), and (3) *instructional support* (concept development, quality of feedback, and language modeling). In an evaluation designed to improve the quality of teacher-student interactions along these dimensions (for example, MyTeaching Partner), CLASS's measure of quality could be an appropriate indicator of the treatment contrast (Allen et al., 2011).

It is important to note that for studies of intervention effects on program service quality, quality measures are sometimes the target outcome of interest. They are therefore the basis for estimates of program effects. In other words, they are treated as if they are ends in themselves. However, for studies of program effects on client outcomes, quality measures can reflect a dimension of the treatment contrast and thus are potential mediators of program effects on client outcomes. In other words, they are treated as means to an end.

Conveyance

Our final dimension is the manner in which services are conveyed. By this we mean the extent to which services are provided to clients individually or in groups and the extent to which services are provided by individuals, such as teachers or counselors, in person, over the telephone, through electronic interactions such as e-mail or other online experiences, or through hard-copy written materials. Many programs consist of a mix of approaches, and it is important to document these approaches for both the treated and untreated conditions because different delivery approaches may affect the way that clients react to a program or intervention.

Issues When Measuring the Treatment Contrast

The following are some issues to consider when measuring the treatment contrast.

Direct Versus Indirect Program Treatment Received

Services received through a program can include those that it provides directly as well as those to which it refers or recommends clients and which are thus provided indirectly. One such example is a college advising program that refers students to other available services, such as the college's tutoring center or library, local day care, and so on. The guidance itself is a critical part of the services that make up the treatment contrast. It is also clear that an evaluation of the referral program should, if possible, collect data on tutoring, library, and day care usage if they are important parts of the chain of events that are hypothesized to yield improved target outcomes.

Displacement of Services

Relatedly, the implementation of a program often coincides with the displacement of services that would be received in the absence of the program. This creates a dimension (or dimensions) of the treatment contrast that does not correspond with the core components of the intervention but is (or are) nonetheless important to measure. For example, the introduction of a cognitive behavioral therapy (CBT) program in a group home for delinquent boys will mean that less of some other activities would be happening in the practice-as-usual condition. The treatment contrast in such a scenario would include increased CBT and a reduction, for example, of group therapy.

Generic or Specific[24] Service Measures

When measuring the treatment contrast, there is sometimes a tension between measuring highly specific program features and more generic aspects of the intervention. Program developers may be most interested in understanding whether clients experience the specific planned program features. However, when measuring a program's treatment contrast to predict variation in its effects, more generic measures might be more appropriate. Consider an evaluation of a program that provided a scholarship and support services targeted to Latino males at a community college in Arizona. The program featured a component called *Pláticas* (conversations) that were held for program group members on a variety of different specific topics, including self-identity, misplaced pride,[25] cultural awareness, and failure[26] (Patel and Valenzuela, forthcoming). Examining whether control group students had conversations on those exact topics and then describing differences as the entire treatment contrast could result in overalignment of program features. Instead, one might extract the broader principle and study the extent to which students (both program and control) had a safe place to discuss personal issues with other Latino males facing similar personal and academic challenges.

Striking a balance of appropriate alignment (not too fine grained but not too broad) is more art than science and may best be accomplished when based on a firm grasp of the principles and theory of the studied intervention. For example, in our Pláticas example, if the program theory was that students respond particularly well to the specific content discussed in the Pláticas (and the theory was not about having a safe place to discuss issues that affect Latino males in community college), then it may be more appropriate to describe the treatment contrast in terms of the specific content.

---

[24]Terminology borrowed from Lipsey and Holdzkom (2008).

[25]The program coordinator introduced this topic to discuss moments when students might allow pride to prevent them from reaching out to others when they needed help.

[26]While students discussed the idea of failure and how it played out in their lives, toward the end of a Plática on this topic, the program coordinator shifted the conversation toward discussing strategies that students could put in place to feel more successful in school and in their personal lives.

Data Collection

Measuring the treatment contrast can be a daunting task, especially for multisite trials in which control group members can receive services from many different organizations that are geographically dispersed. In situations such as these, direct observation might be infeasible. However, it is sometimes possible to collect basic information about services received by program and control group members through follow-up surveys of a random subsample. This is how Bloom, Hill, and Riccio (2003) measured the percentage of program and control group members who received the three main service components of welfare-to-work programs: basic education, job-search assistance, and vocational training.

In addition, administrative records can sometimes be used to measure a service contrast. This approach is being explored for an ongoing study of 105 new small public high schools of choice, or SSCs, in New York City. An initial report from the study found that SSCs increased high school graduation rates by about 8.6 percentage points for disadvantaged students of color (Bloom and Unterman, 2012); the study is now focused on factors that predict variation in these effects. Because control group members attended over 200 high schools, data on the SSC treatment contrast are being obtained from the annual teacher surveys that the New York City Department of Education conducts at all high schools.

Limitations on data collection can make it extremely difficult to quantify a program treatment contrast. Nonetheless, since the treatment contrast is the proximal cause of program effects, it is essential to understand variation in treatment contrasts to understand variation in program effects. We advise researchers to start by developing a logic model, and then attempt to measure, for both program group members and control group members, the key services that the researchers believe will drive impacts.

Mediators

When we refer to the "treatment received" and the "treatment contrast," we are generally referring to the program-related services that clients experience. Most services, however, do not immediately and directly cause changes in target outcomes. Between treatment receipt and the target outcomes lie mediators that are part of a hypothesized causal chain of events that yield program effects.

For example, an advertising campaign designed to reduce HIV transmission might first be received, then raise awareness, knowledge, and fear, which in turn, might cause behavioral changes such as a reduction in the number of sexual partners and/or increased condom usage, ultimately leading to a reduction in HIV transmission (the target outcome). In this case, the treatment contrast would refer to the received content, quantity, quality, and conveyance of the advertising campaign (compared with no campaign or an alternative campaign). A well-

designed (and well-funded) evaluation would measure the hypothesized mediators (awareness, knowledge, fear, number of sexual partners, condom usage) under the treated and counterfactual conditions, in addition to the treatment contrast and target outcome.

- - - - - - -

Thus far we have discussed the first of three sources of variation in program effects — the treatment contrast, which mediates or causes program effects. Now we turn to our second source of variation in program effects, client characteristics, which moderates program effects.

Client Characteristics as Moderators

One important evaluation question to consider is "Who does the program help — all eligible participants or only particular types of individuals?" Program effects may be large for some people and small or null for others, even when there is a consistently robust treatment contrast. Clients have been characterized in many ways for studying such variation. Below we illustrate two such ways.

Client Risk

There may be good reason to expect a program to have different effects for clients with different levels of risk in terms of a study's outcome of interest. Thus, one may wonder, "To what extent do individuals who would fare worst/best in the absence of a program benefit most/least from it?" This question is motivated by popular hypotheses from many fields, including welfare-to-work, medical research, and education (Gueron and Pauly, 1991; Friedlander, 1993; Michalopoulos and Schwartz, 2000; Rothwell, 2005; Kemple, Snipes, and Bloom, 2001).

The intuition behind such hypotheses is that programs may work best for: (1) the participants who are the most disadvantaged, since they have the greatest margin for improvement; (2) the participants who are the least disadvantaged, since they might best be able to use program services; or (3) the participants who are between these two extremes, since they have the best mix of room for improvement and the ability to capitalize on program services.

Specific indicators of risk vary widely across program areas. For example, in education research these indicators are often measures of prior academic achievement (standardized test scores) or school engagement (rates of attendance). In welfare-to-work research they are often measures of prior income, employment, welfare receipt, or education. In public health research they are often measures such as age, weight, and blood pressure or measures of risk behavior such as smoking, drinking, unprotected sexual activity, or poor eating habits.

Client Readiness

Most social and educational programs are intended for clients who, without them, would fare poorly in terms of health, well-being, education, employment, or other life chances. In other words, these programs are intended for clients who "need" them. But in order for clients to benefit from a program, especially a voluntary one, it is often hypothesized that they must be ready, willing, and able to participate (Rollnick, Kinnersley, and Stott, 1993). Specifically, clients must be committed to actively engaging the services that are offered, able to cope with the personal, logistical, and intellectual demands of these services, and have the capacity and resources to spend the time necessary to participate. In short, they must be *ready* for the program.[27]

For parents of young children, client readiness might mean a commitment to keeping appointments and following prescribed regimens for their children's health and well-being. For school-age children, it might mean their parents' ability to get them to school or other program activities regularly and on time. For dropout youth, readiness might mean a willingness to conform to a work-like schedule. For participants in a recovery program for alcohol or substance abuse, readiness might mean a willingness to admit having a serious problem and needing help. For families in a program that subsidizes rents in the private housing market, readiness might mean the ability to navigate the market to find housing that qualifies.

The hypothesis that client readiness moderates program effects makes common sense. However, some research on interventions for low-income youth and families suggests that psychological readiness can be overwhelmed by unpredictable personal circumstances such as the failure of housing, child care, or transportation arrangements that are needed for program participation. Further empirical research is thus needed to better understand the influence of client readiness. In addition, there are few existing measures of this construct (Rollnick, Heather, Gold, and Hall, 1992; Miller and Tonigan, 1996; and Duckworth, 2007, provide examples). Further measurement development is thus also needed.

Client Characteristics Implied by Theory or Policy Significance

Although many client characteristics have been hypothesized to influence program effects, the best guide for choosing them for a given study is the theory of action for the program being tested. For example, when evaluating an adult education program that relies largely on technology to improve instruction, one might hypothesize that, all other things being equal, program effects will vary by age because younger students are more familiar with

[27]Organizational readiness for change is an important factor beyond client readiness to change. For example, if a new program relies on service deliverers (such as teachers, nurses, coaches), their readiness to change may be critical to program implementation.

technology. In general, a strong program theory is a good place to start when selecting client characteristics that might moderate program effects.

In addition, it may be worth examining whether program effects vary by types of clients who are of particular interest to policymakers. For example, if policymakers are particularly concerned about high unemployment rates among veterans, examining a job training program's effects on veterans would make sense.

Program Context as a Moderator

Another major category of factors that can moderate the effects of a program on client outcomes is the broader *context*, or environment, in which the program operates. If the same client were able to experience the same treatment contrast in two different contexts, he or she might nonetheless experience two different program effects.

Consider a youth employment and training program that operates when the unemployment rate is 20 percent versus one that operates when the rate is 5 percent. At 20 percent unemployment, the program might have no effect because there are too few job openings for it to make a difference. On the other hand, if its close relationship with local employers allows the program to have access to job openings that would not typically be available to clients, it might make a difference. With 5 percent unemployment, it might be easier for the program to find jobs for its clients. But it might also be so easy for clients to find jobs on their own that the program could not add much value.

Thus, in theory it is not clear in what way the unemployment rate would influence the effects of an employment and training program. Bloom, Hill, and Riccio (2003) find that, other things being equal, employment and training programs for welfare applicants or recipients have larger effects when unemployment rates are low than when they are high. However, more empirical work is needed to fully understand the influence of this contextual factor.

It is often difficult to distinguish between (1) contextual factors that *directly* moderate the effect of a program by moderating the effect of its treatment contrast on client outcomes and (2) contextual factors that *indirectly* moderate the effects of a program by influencing the treatment contrast itself (or a later mediator). For example:

1) *Direct moderation*: Imagine the effect of receiving a polio vaccine in two contexts: (a) the United States in the 1950s and (b) the United States in the twenty-first century. Receiving the vaccine (versus not receiving it) would have a much larger effect during the 1950s, when the virus was spreading rapidly, than in the twenty-first century, when it is virtually nonexistent. In

this example, the treatment contrast is exactly the same in both contexts, but the context moderates the effect of the treatment contrast.

2) *Indirect moderation*: Imagine an after-school program that is designed to reduce crime by "keeping kids off the streets." It might have no added value when there are many alternative after-school activities already available but substantial added value when it is the only option available. Here the context changes the treatment contrast and, in doing so, leads to impact variation.

Section 3

Connecting the Treatment Contrast to Program Implementation and Service Fidelity

The previous section focused on the right-hand side of Figure 1, which highlights the proximal sources of variation in program effects. The present section focuses on the left-hand side of the figure, which represents the distal factors that produce the treatment received by program group members — one half of the treatment contrast. Variation in these upstream factors can yield variation in the treatment received (and treatment contrasts), and thus variation in program effects. Some of these factors (for example, program take-up) can vary within and between sites; consequently they are important factors to consider when explaining variation in program effects, both within and between sites. Other factors vary primarily between sites (for example, organizational or site characteristics like "leadership"), and therefore are most useful to consider when explaining variation in program effects between sites only. All of these factors are upstream from the treatment contrast.

From Treatment Offered to Treatment Received: Client Take-Up

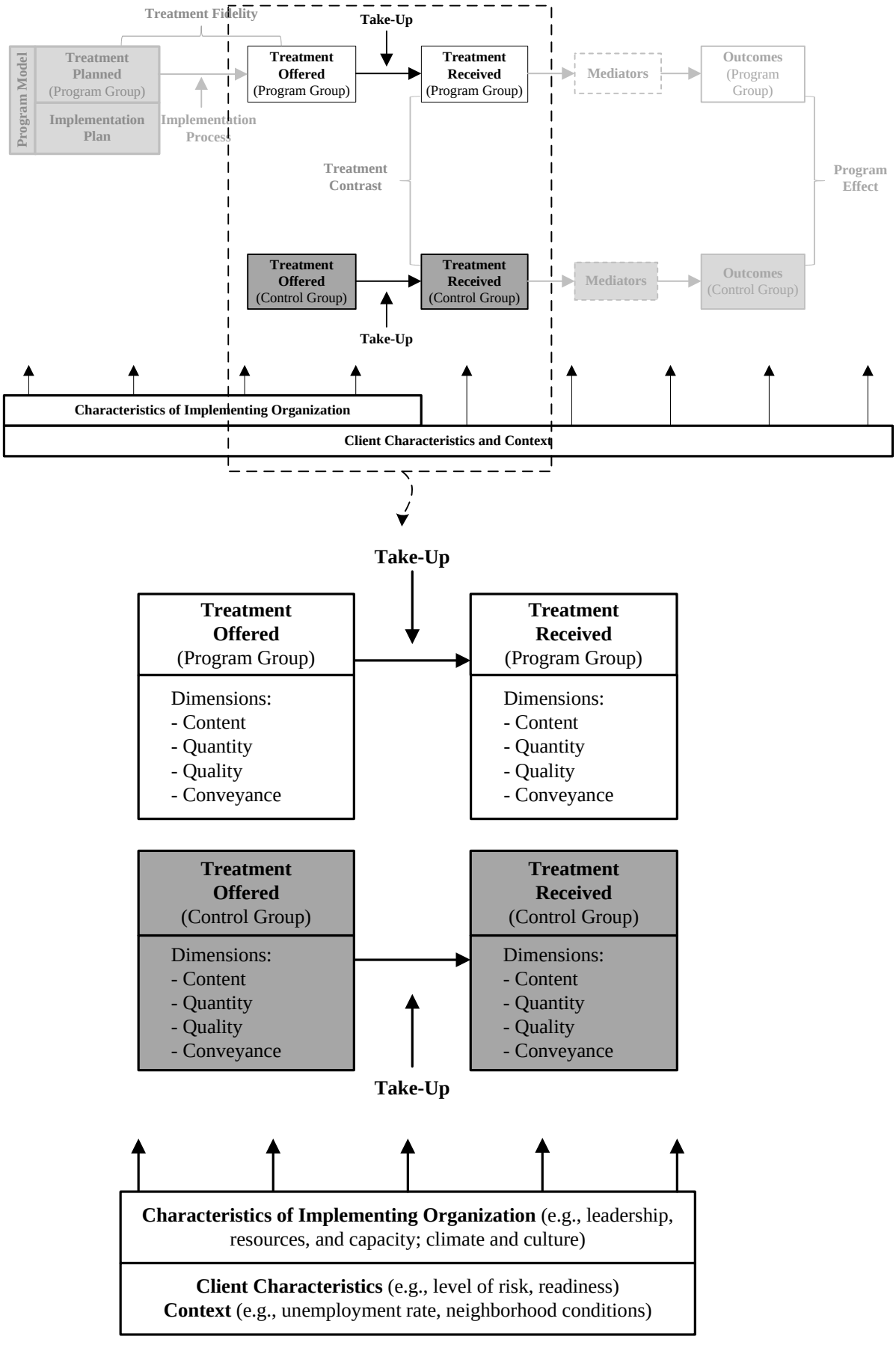
Continuing to work from right to left in Figure 1, we distinguish between treatments that are *received* by clients and treatments that are *offered*, or made available, to them. (As before, treatments are represented by their content, quantity, quality, and conveyance.) By definition, the link between treatments offered and treatments received is client "take-up." A treatment can be delivered exactly as designed in terms of the services made available to clients, but if clients don't show up (that is, if take-up is low), then the treatment contrast will be diminished, as will the chance that the program will produce its desired effects. To the extent that take-up varies across sites or types of clients, this can lead to variation in treatment contrasts and, ultimately, to variation in program effects.

Figure 3 helps focus on this part of our framework by zooming in to provide some additional detail. Although take-up is typically thought of as binary, our focus on the quantity and quality of services made available (offered) and received makes clear that the amount of services offered and received (service quantity) exists on a continuum, as is also true of how engaged and responsive clients are to those services (an element of service quality).

We focus on this link because it is potentially an important determinant of a program's treatment contrast and because there are explicit steps that program developers can take to strengthen that link. Several strategies that have been used to increase take-up of services include *incentives* or *inducements* (Skoufias, 2005; Scrivener, Weiss, and Sommo, 2012; Patel

Figure 3

Treatment Offered and Treatment Received



and Valenzuela, forthcoming), *requirements* (Scrivener, Sommo, and Collado, 2009; Scrivener, Weiss, and Sommo, 2012), *removal of potential barriers* (Bettinger et al., 2009; Madrian and Shea, 2001), and *outreach* (Castleman, Arnold, and Wartman, 2012). These examples all come from the evaluation literature, and in some cases the strategy was intended to increase take-up of an intervention being evaluated, while in others the strategy was the intervention itself and the target outcome was participation.

In either case, these approaches illustrate different strategies to increase program take-up that have been tested and thus may represent good options for program developers to include *as part of their program services*. If such efforts are exhausted and take-up remains low, then the program services might need to be modified to make them more engaging, accessible, or culturally appropriate. It may also be possible that the services reflect a fundamental misunderstanding of what clients want or need and should thus be retooled.

From Treatment Planned to Treatment Offered: Program Implementation

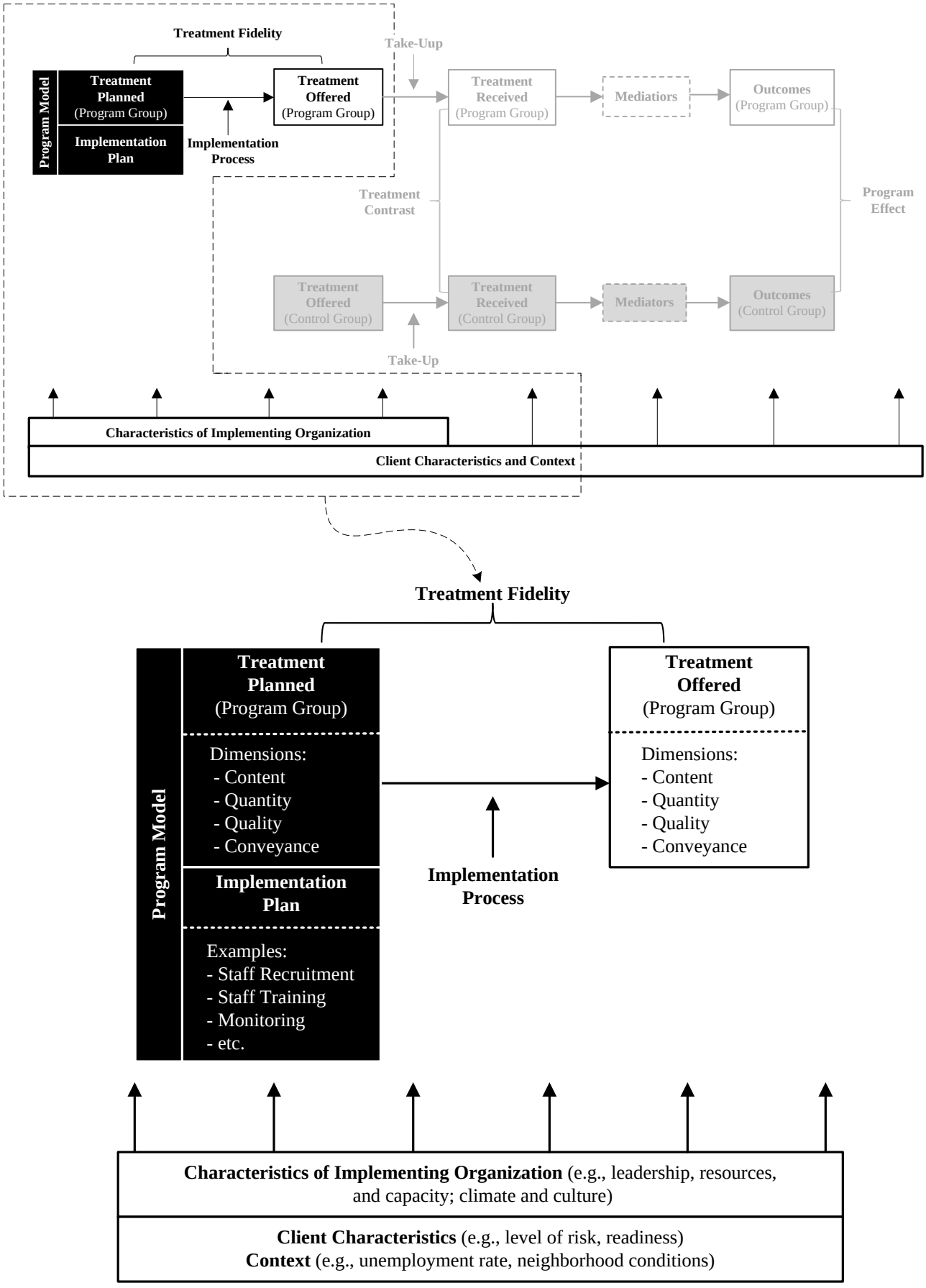
The connection between the treatment that is planned for clients and the treatment that is offered to and received by clients is where implementation, or "the process of putting a defined practice or program into practical effect" (Fixsen et al., 2005, p. 82), comes into play. There have been countless studies of the implementation of specific programs, and many different frameworks have been proposed for guiding researchers who study factors that explain why some programs succeed in delivering services as planned and others fail to do so (see, for example, Damschroder et al., 2009; Durlak and DuPre, 2008; Dane and Schneider, 1998; Elmore, 1985; Mazmanian and Sabatier, 1989; Van Meter and Van Horn, 1975). To facilitate our discussion of these issues, Figure 4 zooms in to the portion of our proposed framework where it is represented.

Below we consider the implementation process for a program at a given location or site. To the extent that the factors examined vary across sites, one might expect variation in their implementation processes. For simplicity, we conceptualize site-level program implementation as a function of three main factors that are represented in Figure 4: (1) the *treatment planned* as part of the program model, (2) the *implementation planned* as part of the program model, and (3) *the characteristics of the implementing organization or organizations*.[28] Program models vary markedly in the specificity of their treatment and implementation plans. In addition, implementing organizations vary widely in the degree to which they are amenable to, and capable of, adopting a given program model versus adapting it to meet local conditions and preferences.

[28]Two other key factors represented in our framework, but not described here, are the clients' characteristics and the broader context (beyond the implementing organization).

Figure 4

Connecting the Treatment Offered to Program Implementation and Treatment Fidelity



The treatment plan describes what a program is expected to offer to clients. (This can be thought of in terms of content, quantity, quality, and conveyance.) The treatment plan can vary widely (if not wildly) in its specificity, complexity, and explicitness. At one extreme, the treatment plan may be little more than a statement of the *problem* that a program is supposed to address, the *population* that it is supposed to serve, and the *principles* by which it is supposed to be run. At the opposite extreme, the treatment plan may provide a set of prescriptions for the content, quantity, quality, and conveyance of program services — detailing, for example, what program staff and clients will be doing on any given hour or day of the week, and establishing clear guidelines for what services are to be offered and by whom.

The implementation plan is the set of instructions on how the treatment plan is to be realized. Several broad-based literature reviews of implementation research conducted in the health, mental health, social services, juvenile justice, education, employment services, and substance abuse treatment fields have led to "meta-frameworks" that try to spell out the processes and factors that contribute to successful program implementation and treatment fidelity (Cordray and Pion, 2006; Hulleman and Cordray, 2009; Dane and Schneider, 1998). Three elements common to these frameworks are (1) staff recruitment and selection, (2) staff training, and (3) staff monitoring and/or supervision and supports. As with treatment plans, implementation plans vary widely in their level of specificity.

One program that is well known for the specificity of both its treatment plan and implementation plan is Success for All, a school reform model for improving reading achievement for elementary school students.[29] This model, now operating in over 1,400 U.S. elementary schools, was developed over many years and has been evaluated extensively (for example, Borman et al., 2007). It has an unusually detailed treatment plan that includes specific curricular materials and prescribed instructional practices and activities that are designed to be developmentally appropriate for students in kindergarten through sixth grade. It also has an unusually detailed implementation plan that includes materials and plans for on-site and off-site teacher professional development, teacher monitoring, feedback, and coaching, and a wide range of activities that are designed to convey the essential ingredients of the program. One indication of the degree of extensive detail in these plans is Success for All's 27-page, single-spaced contract with participating schools that thoroughly outlines the program's elements and requirements.[30]

This high degree of specificity in Success for All stands in stark contrast to six learning communities programs evaluated by MDRC. These particular programs operated in community colleges throughout the United States and were designed to improve outcomes for students in

[29]See http://www.successforall.org/elementary/.

[30]A blank copy of this contract was provided to the authors by Nancy Madden, president and cofounder of Success for All.

developmental (remedial) education (Visher et al., 2012). They involved linking a developmental English or math course with one or two other courses at the college, and placing students into groups that take these courses together — in part to encourage curricular integration and also to foster greater peer support (Tinto, 1998). Community college administrators and faculty are often drawn to learning communities after reading about them or hearing presentations at conferences or professional meetings. There was, however, no national organization that monitored the way the learning communities were taught or, in MDRC's evaluation, that certified that particular standards were met. There was also no master plan or guidebook to inform faculty how they should teach in learning communities. Perhaps not surprisingly, MDRC's evaluation of these learning communities found significant variation in their content and quality, both within and across colleges (Visher et al., 2012).

It stands to reason that the more fully specified a program treatment plan is, the more influence it will have on the services that are offered by a site. Likewise, the more fully specified a program implementation plan is, the more influence it will have on how a program is implemented at a site — and, most likely, on the services that are offered. All of this affects treatment fidelity (Dane and Schneider, 1998; Dusenbury, Brannigan, Falco, and Hansen, 2003; Cordray and Pion, 2006; Hulleman and Cordray, 2009).[31] Fixsen et al. (2005) argue that implementation is facilitated by having a practice or program that is well defined and clearly operationalized. An earlier generation of implementation researchers concerned with the implementation of federal and state policies advanced a similar argument in calling for greater specificity and coherence in program statutes in order to increase the odds that policymakers' intentions would be realized (see, for example, Mazmanian and Sabatier, 1989).[32]

While these arguments are compelling, there may be limits to the degree of specificity that is desirable in a treatment or implementation plan. Researchers (for example, Bardach, 1980; Hjern and Hull, 1982) have long recognized that organizations responsible for carrying out program implementation have agendas, needs, priorities, and goals of their own and that these may not fully align with those of treatment planners or program model developers. Too much specificity in the service or implementation plan could backfire and lead to resistance on the part of implementing organizations, when what is most needed is their buy-in. Browne and Wildavsky (1984) and Kezar (2011) have suggested that policymakers (or model developers) and program operators need to learn from each other and be willing to compromise — a process they refer to as *mutual adaptation.* From this perspective, policymakers or model developers may be better off articulating broad goals and strategies and allowing program implementers to figure out the details.

[31]For a definition of this term and related terms used elsewhere in the literature, see Section 1.

[32]Similar points are made by Blakely et al. (1987), Dobson and Shaw (1984), and Kazdin (1986), as cited in Dane and Schneider (1998).

Researchers can shed light on these issues by noting the level of specificity within a service or implementation plan and determining whether it is associated with treatment fidelity (and, ultimately, treatment contrasts and effects).

Characteristics of the Implementing Organization

A simple but essential truth about program implementation is that it happens *inside* local organizations. Damschroder et al. (2009) define implementation as "the constellation of processes intended to get an intervention into use within an organization" (p. 3). A number of researchers have identified characteristics that may make some organizations more predisposed to implementing an intervention well (for example, Chase, 1979; Damschroder et al., 2009; Durlak and DuPre, 2008). Four major factors include (1) strong leadership, (2) sufficient resources and capacity, (3) supportive climate or culture, and (4) involvement of an outside monitor or "fixer."

While this list is not exhaustive, it offers a starting point for understanding differences across program sites — differences that may help to explain variations in treatments received and ultimately in program effects.

Strong Leadership

The execution of a program implementation plan requires a strong leader — or leadership team — that is fully committed and willing to see implementation through. Durlak and DuPre (2008) identify the role of leadership as setting priorities, building consensus, and investing the requisite time and skills to manage the implementation process. Fixsen et al. (2005) outline five key tasks that leaders must perform: (1) initiating and shepherding the organization through the steps outlined in the implementation plan; (2) setting goals and deadlines and communicating them clearly throughout the organization; (3) assigning individuals or teams to specify the details of activities, processes, and tasks to put the implementation plan into effect; (4) inspiring and motivating staff; and (5) recruiting and retaining the right staff to deliver the planned services. These tasks may form a checklist or be turned into a rating system for assessing the strength of program leadership during the implementation process.

Some of the implementation literature stresses the value of a "program champion" who can rally and maintain support for the intervention and negotiate solutions to implementation problems that arise (see, for example, Durlak and DuPre, 2008). Ideally, program champions are situated at or near the top of an organization so that they can use their influence to garner resources needed for program implementation. Midlevel staff can also function as champions so long as they are able to dedicate substantial time to the program and

have strong support from program administrators and colleagues. For example, MDRC's evaluation of learning communities in community colleges found that committed faculty leaders, including a paid coordinator, were essential to managing and scaling up these initiatives. As coordinators clarified expectations and offered support to faculty teaching in learning communities, faculty responded by changing their teaching practices (Visher, Schneider, Wathington, and Collado, 2010).

Effective relationship-building — both inside and outside the organization — is arguably the most important aspect of strong leadership. Internally, leaders need to get staff on board and working together to carry out the implementation plan. Externally, organizational leaders need to build relationships with funders, community leaders, and others to ensure that the program has the financial and political support it needs. Bryk et al. (2010) paid close attention to these issues in their study of elementary school reform in Chicago. After comparing the organizational conditions of 100 schools that had made improvements in reading and math scores with 100 schools that had not, they identified inclusive leadership as one of the primary factors that substantially influenced the dynamics of teaching and learning in the classroom. Specifically, schools were more likely to show improvements if their principals regularly reached out to and involved faculty, parents, and community leaders in setting goals and making major decisions about how instruction would be delivered and how their school would be operated. They also found that school improvement was associated with the principal's and other school leaders' knowledge of the immediate community, as well as the amount of time they spent in the community and the closeness of their partnerships with people or institutions there.

Finally, leadership strength and commitment can be markedly influenced by the stability of a leadership team. To the extent that there is frequent turnover at the level of the president, director, or other key positions, it is less likely that an organization will maintain its focus on the implementation process and do a good job of putting a new program into place. For example, in a five-year evaluation of the Achieving the Dream initiative — a foundation-funded effort designed to help community colleges undertake a five-step institutional improvement process — nearly half of the colleges involved experienced at least one change in presidential and senior administrator leadership, and several colleges experienced multiple changes. Not surprisingly, these institutions tended to make only moderate or weak progress toward carrying out the program (Zachry Rutschow et al., 2011). Similarly, an evaluation of early Head Start programs identified leadership changes as a factor that sometimes set back or stalled progress in implementation (Kisker, Paulsell, Love, and Raikes, 2002).

Sufficient Organizational Resources and Capacity

Although organizations require adequate fiscal resources to hire staff, acquire space, and purchase whatever goods and services are needed to carry out an implementation plan, money alone does not guarantee implementation success. In this regard, Mazmanian and Sabatier (1989) suggest that there is "a threshold level of funding" (p. 26) necessary to achieve programmatic objectives, and add that while the probability of achieving those objectives may rise with increased funding, there might be a point of saturation beyond which additional dollars provide no further value. Understanding what different organizations spend on program implementation — and how spending levels correlate with other indicators of the quality of services offered or received — may help identify this threshold and help explain variation in program effects.

The dedication and skills of an organization's staff is another factor to consider. In the aforementioned study of school improvement in Chicago, Bryk et al. (2010) identified "professional capacity" as one of five core conditions related to school improvement.[33] Their measure of professional capacity included factors such as teachers' experience living and working in the community served by the school (on the theory that more experience is associated with deeper commitment to the school and better understanding of the population it serves) and quality of undergraduate education, averaged across teachers within a school. They also took into account the frequency and quality of professional development offered by the school, including the presence of a coherent and sustained professional development program. The importance of professional development as a predictor of implementation success is a major theme in implementation research (Fixsen et al., 2005). In a review of research on factors affecting the program implementation process in the youth field, for example, Durlak and DuPre (2008) emphasize the importance of training and technical assistance that enhances staff's sense of self-efficacy, offers emotional support, and encourages local (rather than top-down) problem solving.

Finally, research suggests that program implementation processes benefit from staff stability at all levels of an organization — not just the top. Organizations that experience rapid turnover at the lower and middle levels must divert attention to hiring new workers and bringing them up to speed, and will not benefit from the accumulated knowledge or interpersonal relationships that are formed when staff members stay in their positions for extended periods. In their review of program implementation in the health care field, Damschroder et al. (2009) conclude that the more stable implementation teams are, the more likely implementation will be successful. As an example, a national evaluation of the early Head Start program found that low

[33]The other core conditions were: inclusive leadership (described above), strong parent-school-community ties, a student-centered learning climate (an indicator of school safety and ways of managing disruptive behavior by students), and the structure and integration of curricula across grade levels.

turnover among center leaders and staff was an important factor in helping programs reach full implementation more quickly (Kisker, Paulsell, Love, and Raikes, 2002).

Supportive Organizational Culture and Climate

The culture and climate of an organization are defined by its institutionalized norms, values, and belief systems. Though the terms are often used interchangeably, "culture" is sometimes considered to be more permanent and enduring, while "climate" may be more variable across divisions of an organization or may be influenced by external events or conditions (for example, an election cycle or a series of budget cuts.) From an implementation standpoint, what matters about organizational culture and climate is what Damschroder et al. (2009) refer to as its "absorptive capacity for change" (p. 8). In other words, organizational norms, values, and belief systems may influence an organization's ability to carry out an implementation plan and deliver program services as intended. For instance, a health education program designed to prevent unwanted pregnancy would probably look and feel quite different as implemented by a faith-based social service organization than by Planned Parenthood, given that the organizations start out with very different attitudes toward human sexuality and contraception.

Damschroder et al. (2009) suggest that an organization may be more conducive to implementing an intervention if its culture or climate is characterized by several features.[34] First is a belief among staff that the situation *without* the planned intervention is intolerable and requires change. Second is the perception of a "tangible fit" between the norms and values associated with the intervention and the organization's norms and values. Third is a sense among staff that the intervention is a priority for the organization. Fourth is the presence of organizational incentives and rewards aligned to the intervention (for example, bonuses or pay raises for staff who adopt the new practice). Fifth is regular reinforcement of programmatic goals in staff communications. Last is what the authors term a "positive learning climate": that is, a sense that it is safe to try new things, that it is okay to make mistakes so long as there is an effort to learn from them, and that there is adequate time and space for reflection.

One tool for measuring organizational culture, climate, and work attitudes is the Organizational Social Context measurement system (OSC) (Glisson et al., 2008). The OSC is based on a theory that successful program implementation depends as much on the social processes within an organization as on the technical processes embodied in its treatment plan or implementation plan. For example, an organization might be expected to do better at implementation when its culture is rated as proficient (for example, one in which clients' needs are placed first) as opposed to rigid or resistant (for example, one in which staff have little

[34]Using a somewhat different rubric and terminology, Durlak and DuPre (2008) identify many of the same factors as important influencers.

discretion or flexibility in their daily work and show little interest in making changes). Similarly, an organization would seem more likely to carry out an implementation plan if its staff members described their office climate as functional and engaged (for example, characterized by cooperation and a sense of accomplishment) rather than stressful (for example, reflecting multiple ambiguous goals, an inability to get necessary things done, and exhaustion from the work that is required). Though the OSC was developed for mental health organizations, it has been demonstrated to have strong psychometric properties and may be adaptable for use in evaluations of other education and social service programs. Specifically, researchers may use the OSC to determine whether indicators of organizational culture and climate are associated with stronger program implementation and, ultimately, with program effectiveness.

Involvement of an Outside Monitor or "Fixer"

Finally, a theme from some implementation research is that an organization is more likely to implement a program model successfully if there is an external overseer who is specifically charged with monitoring this implementation (Fixsen et al., 2005; Mazmanian and Sabatier, 1989). The function could be imagined as one of compliance or one of technical assistance and support (or some combination). Bardach (1980), for example, emphasized the role of a "fixer" in ensuring the implementation of a California mental health reform program. In this case, a prominent state legislator played the role, but it could also be performed by a strong program officer at a foundation, a government board, an intermediary organization, or a national office that "owns" a particular program model. The Nurse-Family Partnership (NFP) program offers a good example of this. After three randomized control trials demonstrated the effectiveness of NFP in improving the health outcomes of low-income single mothers and their children, an NFP National Service Office was set up to make sure that other health agencies would replicate the model precisely (Olds et al., 2007). The National Service Office now works to educate policymakers, clinicians, and the public about the research behind the model and provides technical assistance to providers who use it.[35]

The Program Implementation Process and Program Treatment Fidelity

Our previous discussion of the program implementation plan outlined several key elements that are often put in place to enable a program to happen (for example, staff recruitment, selection, training, monitoring, and supervision, along with various supports). These same elements are reflected in the enacted implementation process that occurs for any given program at any given

[35]See also http://www.nursefamilypartnership.org/about.

site.[36] The next questions to ask are: To what extent does the treatment offered to clients faithfully reflect the treatment that was planned? In other words, to what degree did the treatment offered to clients demonstrate *fidelity* to what was intended for them? In addition, how does treatment fidelity vary across sites and to what extent does it help to predict variation in treatment contrasts and program effects?

The issue of fidelity to a plan has been discussed by social scientists for decades, and in other settings it has been discussed for centuries. For example, according to Cordray and Pion's (2006) survey of the literature on treatment strength and integrity (fidelity), the origins could be "traced back to the earliest moments of civilization. For example, major German beer makers proudly advertise that their product is brewed in accordance with the German Purity Law of 1516!" (p. 115). However some authors chart the trajectory of research on treatment integrity as beginning in the 1950s, with research on psychotherapy (Bond et al., 2000; Moncher and Prinz, 1991).

Cordray and Pion (2006) begin their history and survey of this issue in the field of evaluation research with work in the 1970s by Lee B. Sechrest and his colleagues (for example, Sechrest and Redner, 1979; Sechrest et al., 1979). They credit this work as laying the foundation for a rigorous conceptual framework for studying treatment strength and integrity. Sechrest and his colleagues carefully distinguished between *treatment strength* — the type and amount of services prescribed for a program (planned treatment) — and *treatment integrity* — the extent to which treatment services are delivered as planned (treatment fidelity). Cordray and Pion (2006) then describe the evolution of research on these issues and extend it to include the concept of a program's "achieved relative strength," which we refer to as the program's "treatment contrast."

Hulleman and Cordray (2009) then apply these concepts to study why the effect of a motivation-based educational program as measured by a tightly designed laboratory experiment was much larger than its effect as measured by a multisite field experiment. They find that the treatment contrast was much larger in the laboratory experiment (for 107 undergraduate students at a single university) than it was in the field experiment (for 182 high

[36]Although not depicted in the present framework, Hulleman, Rimm-Kaufman, and Abry (2013) conceptualize *implementation fidelity* as answering the question: To what extent does the enacted implementation process reflect the planned implementation process? For example, if frontline staff members are supposed to attend a five-day training, did they? The distinction between *implementation fidelity* and *treatment fidelity* may be very helpful to program developers. If treatment fidelity is less than desired, this may be due to a lack of fidelity to the implementation plan or an inadequate implementation plan — each of which suggests different improvement strategies. This is similar to the distinction between theory failure and implementation failure that has been described by Raudenbush (2008) and Rossi, Lipsey, and Freeman (2004), just at a different location in the conceptual framework.

school students from 13 science classrooms taught by 8 teachers at 3 high schools).[37] This provides strong evidence of the relationship between a program's treatment contrast and its effects on client outcomes. Other researchers have also studied this relationship and interest in it is growing rapidly.[38]

One problem that often arises when this work is discussed is that the treatment contrast (which represents the difference between treatment received by program group members and treatment received by control group members) is often conflated with treatment fidelity (which represents the difference between treatment offered to or received by program group members and the treatment that was planned for them). This frequently occurs because both have been referred to as components of fidelity. In fact, some researchers consider "differentiation" (which is essentially the treatment contrast) to be one dimension of treatment fidelity — we do not. To avoid this ambiguity, our proposed framework (Figure 1) depicts and labels *treatment contrast* and *treatment fidelity*, visually displaying their relationship and thus highlighting the important distinction. Crucially, to the extent that treatment fidelity has the potential to influence program effects, it does so through one half of the treatment contrast. A further distinction that we make is between treatments offered or made available by a program and treatments received by its clients; the difference between treatments offered and treatments received is accounted for by client take-up. By focusing explicitly on this step in the causal chain between a program's treatment plan and its effects on client outcomes, we believe that researchers, practitioners, and policymakers can obtain important information for predicting variation in program effects and guiding decisions about how to improve program performance.

[37]They also found that variation in the intervention's achieved relative strength in the field experiment was positively correlated with its variation in educational effects.

[38]For example, Dusenbury, Brannigan, Falco, and Hansen (2003) provide an extensive review of research on "implementation fidelity" in school-based drug abuse programs.

Section 4

Conclusion

This final section considers how our proposed framework is relevant for program creation, improvement, and evaluation. To facilitate the discussion, Table 1 lists elements of the framework with illustrative measures for each element. The section concludes by returning to the research examples that were presented initially to illustrate promising ways to study sources of variation in program effects.

Relevance of the Framework

Researchers who study policy implementation or who focus on understanding how programs operate or improve usually focus on a program's treatment plan, implementation plan, and its local organization, all of which are located upstream in our framework (Figure 1). When a program model specifies a clearly defined set of services for a clearly defined target population, then program monitoring, development, and research can focus on the extent to which the treatment that is offered aligns with the treatment that was planned (treatment fidelity). If treatment fidelity is inadequate, emphasis can be placed on understanding failures in enacting the program's implementation plan (for example, was training for treatment providers offered as planned?) or improving, creating, or clarifying the implementation plan.

If the treatment offered by a program varies appreciably across implementing organizations, it seems worth unpacking the sources of this variation. An obvious place to start is to consider whether the services planned for clients by sites vary in terms of their content, quantity, quality, and conveyance. If sites do not plan the same services, it is unlikely they will offer the same services. Cross-site variation in services offered by a program may also depend on the interaction between the program's implementation plan and its implementing organizations. This interaction leads to the implementation process that is enacted by each site. Unlike corporate franchises, many social programs have no formal implementation plan. Instead, local organizations often must figure things out for themselves, which can produce substantial variation, for better or for worse. However, some programs (for example, Success for All) have a tightly specified treatment and implementation plan, which presumably produces less variation in the services offered.

Table 1 lists some factors to consider when examining a program's implementation plan (its clarity, specificity, adaptability, and monitoring) and when assessing its local implementing organizations (their leadership, resources, capacity, climate, culture, and external monitoring). These factors can influence the treatment offered by a program and thus its treatment fidelity.

Table 1

Examples of Measures for Elements of the Conceptual Framework

Construct	Possible Measures
Services Planned, Services Offered, and Services Received	
The components, features, and activities of a program that clients are intended to experience	
Content *What services are provided?*	• Type of services offered (e.g., instruction emphasizing phonemic awareness)
Quantity *How much of each service is provided?*	• Prevalence, frequency, intensity, and duration (e.g., 30-minute tutoring sessions, offered 5 days per week, lasting 30 minutes each)
Quality *How well is each service provided?*	• Interactions between staff and clients (e.g., CLASS is a system for observing and assessing the quality of interactions between teachers and students) • Ratings of program setting and resources (e.g., Early Childhood Environmental Rating Scale-Revised Edition [Harms et al., 1998] is used to assess the quality of child care services provided by preschools and Head Start centers)
Conveyance *How, when, and by whom is each service provided?*	• Services provided to clients individually (versus in groups) • Services provided in person, over the telephone, by electronic means such as e-mail, or through hard-copy written materials
Client Characteristics	
Risk	• Measure of prior academic achievement (e.g., standardized test scores) • Measures of employment, prior income, welfare receipt • Measures of age, weight, blood pressure; risk behaviors such as smoking or drinking
Readiness	• Extent to which clients feel ready to make a change • Extent to which clients persevere in their goals (e.g., "grit" scale, measuring client's perseverance and passion for long-term goals [Duckworth, 2007])

(continued)

Table 1 (continued)

Construct	Possible Measures
Program Context	
Location type	• Rural/urban/suburban
Economic indicators	• Unemployment rate • Average household income
Safety	• Crime rate
Sociodemographic variables	• Ethnic/racial composition • Percentage foreign-born • Percentage of adults holding a high school and college degree
Implementation Plan and Implementation Process *"The process of putting a defined practice or program into practical effect"* *(Fixsen et al., 2005, p. 82)*	
Clarity and specificity	• The degree to which program planners are explicit about their implementation plan (i.e., with respect to staff recruitment, training, monitoring, and support)
Adaptability	• The degree to which an intervention can be adapted, tailored, or reinvented to meet local needs (Damschroder et al., 2009). There is an inherent tension between adaptability and fidelity, and there is disagreement on whether and how much adaptability should be permitted in an implementation process.
Monitoring	• The degree to which there are plans to monitor the delivery of services

(continued)

Table 1 (continued)

Construct	Possible Measures
Organizational Characteristics	
The characteristics of the service delivery organization	
Leadership	• Presence of "program champion" • Inclusive leadership scale (Bryk et al., 2010) - Involvement of staff in goal setting, planning - Involvement of community members in goal setting, planning • Stability of leadership
Resources and capacity	• Program cost per client served • Frequency and quality of professional development for staff • Professional capacity scale (Bryk et al., 2010) - Staff experience with, and commitment to, community - Quality of staff education/training
Climate and culture	• Organizational social context scales (Glisson et al., 2008) - Proficient or resistant culture - Functional or stressful culture
External monitoring/"fixing"	• Presence of external overseer • Level of outside technical assistance

A program's treatment plan, implementation process, and take-up determine the treatment received by its program group members — which is one half of the program's treatment contrast and therefore one half of the proximal cause of the effects on client outcomes. The other half of the treatment contrast is the treatment received by control or comparison group members; this represents the treatment that would have been received by program group members in the absence of the program. To understand variation in program effects, one must understand both halves of the treatment contrast.

Most researchers who conduct summative evaluations of program effects are well aware of the role played by a program's treatment contrast in producing its effects. Likewise most researchers who conduct formative evaluations of program implementation are well aware of the role played by local organizations and program implementation processes in producing its treatment fidelity. The central goal of this paper is to encourage these stakeholders and others to consider the broader picture of how these factors fit together. For example, practitioners or

policymakers who are overly focused on program implementation and treatment fidelity may not recognize that their program is located near other similar services and thus has a small margin for producing a net gain.

A striking example of this is provided by two recent studies from the international development literature. One study evaluated the effects of BRIGHT schools for villages in the African country of Burkina Faso. The other study evaluated the effects of IMAGINE schools for villages in bordering Niger. Both programs constructed quality schools for village children. However, BRIGHT schools had estimated effects of about 0.4 standard deviations in math and French (which is substantial) whereas IMAGINE schools had negligible estimated effects.

While many factors might be responsible for this difference, perhaps the most compelling explanation is the remarkable difference in alternative services that were available. Only 60 percent of the BRIGHT comparison group villages had a preexisting school, while 99 percent of the IMAGINE control group villages had a preexisting school (Levy, Sloan, Linden, and Kazianga, 2009; Dumitrescu, Levy, Orfield, and Sloan, 2011). The resulting service contrast for BRIGHT schools was a 20 percentage point program and comparison group difference in school enrollment rates versus a 4 percentage point difference for IMAGINE schools.[39] Although BRIGHT and IMAGINE schools both intended to provide higher quality education than the alternative, the difference between nothing and something was greater than the difference between something and something intended to be of higher quality.

Evaluation researchers can use our proposed framework to see this bigger picture and also as a guide or checklist for data collection efforts. To the extent that it is possible, researchers should be aware of and collect data on a program's treatment plan, offer, and receipt and the implementation plan and enacted implementation process, and they should understand the theory for why and how the program is expected to produce its intended effects.

For a study of program efficacy (its potential effects under favorable conditions), researchers should collect the preceding information as well as data on the observed treatment contrast and observed program effects.

For a study of program effectiveness (its actual effects in multiple locations under normal operating conditions), researchers ought to attempt to collect the preceding information, especially data on the program treatment contrast. This can be used to help explain any observed differences in program effects between the efficacy and effectiveness tests and any observed differences in program effects across sites, contexts, and client types in the effectiveness test.

[39]For the BRIGHT study about 55 percent of program group members and 35 percent of comparison group members enrolled in a school. Corresponding results for the IMAGINE study were 79 percent and 74 percent.

Concluding Thoughts: Closing the Loop

This paper began with three examples of evaluation research that demonstrated that it is possible to learn about more than simply the average effect of a program and the extent to which it was implemented faithfully. The first example is a research synthesis or meta-analysis of reported results from previous research. The second example is a single multisite randomized trial. The third example is a secondary analysis of primary data from three previous multisite randomized trials.

These studies share several important features. First, they capitalize on separate estimates of program effects for multiple sites (after-school programs, schools, or welfare offices). This made it possible to conduct *exploratory* research on why program effects were larger for some sites than others. Second, each study was based on *natural variation* in program features and treatment contrasts (for example, active learning strategies, clear behavior standards, and a strong and consistent employment message), client characteristics (for example, participating children, school students, and welfare recipients), and program contexts (for example, after-school programs, schools, and labor markets). By examining the extent to which variation in these factors predicted variation in program effects, each study was able to provide *suggestive* evidence about what made the observed programs work, for whom they worked, and under what conditions they worked.

In addition, a few studies have sought to rigorously confirm hypotheses about the influence of specific program features by randomizing planned variation in these features. For example, after finding that performance-based scholarships improved academic outcomes for low-income parents attending two community colleges in Louisiana (Richburg-Hayes et al., 2009a), a new project was launched to test the effects of such scholarships in different contexts, for different clients, and with randomly assigned variation in the timing, duration, and amount of the scholarships and the criteria for receiving them (Richburg-Hayes et al., 2009b; Ware and Patel, 2012).

The preceding exemplar studies should provide encouragement for: (1) researchers to pursue similar studies of other social and educational programs, (2) research funders to build agendas that support this kind of research, and (3) practitioners and policymakers to rely on this type of research and demand more of it. We hope that our proposed conceptual framework can help focus this research on the key questions that it should address and integrate its findings in ways that best inform the design, implementation, and improvement of future programs.

References

Abadie, Alberto, Joshua Angrist, and Guido Imbens. 2002. "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings." *Econometrica* 70, 1: 91-117.

Agodini, Roberto, and Barbara Harris. 2010. "An Experimental Evaluation of Four Elementary School Math Curricula." *Journal of Research on Educational Effectiveness* 3, 3: 199-253.

Allen, Joseph P., Robert C. Pianta, Anne Gregory, Amori Yee Mikami, and Janetta Lun. 2011. "An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement." *Science* 333, 6045: 1034-1037.

Angrist, Joshua D., Guido Imbens, and Don Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91, 434: 444-455.

Bardach, Eugene. 1980. *The Implementation Game: What Happens After a Bill Becomes a Law.* Cambridge, MA: MIT Press.

Bettinger, Eric, Bridget Long, Philip Oreopoulos, and Lisa Sanbonmatsu. 2009. *The Role of Simplification and Information in College Decisions: Results from the H&R Block FAFSA Experiment*. NBER Working Paper. Cambridge, MA: The National Bureau for Economic Research.

Bitler, Marianne P., Jonah B. Gelbach, and Hillary W. Hoynes. 2006. "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments." *American Economic Review* 96: 4: 988-1012.

Blakely, Craig H., Jeffrey P. Mayer, Rand G. Gottschalk, Neal Schmitt, William S. Davidson, David B. Roitman, and James G. Emshoff. 1987. "The Fidelity Adaptation Debate: Implications for the Implementation of Public Sector Social Programs." *American Journal of Community Psychology* 15, 3: 253-268.

Bloom, Howard S., and Susan E. P. Bloom. 1981. "Household Participation in the Section 8 Existing Housing Program: Evaluating a Multistage Selection Process." *Evaluation Review* 5, 3: 325-340.

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., and Charles Michalopoulos. 2011. "When Is the Story in the Subgroups? Strategies for Interpreting and Reporting Intervention Effects on Subgroups." *Prevention Science* 14, 2: 179-188.

Bloom, Howard S., Stephen W. Raudenbush, and Michael Weiss. Under review. *Estimating Variation in Program Impacts: Theory, Practice, and Applications*. New York: MDRC.

Bloom, Howard S., and Rebecca Unterman. 2012. "Sustained Positive Effects on Graduation Rates Produced by New York City's Small Public High Schools of Choice." Policy Brief (January). New York: MDRC.

Bond, Gary R., Lisa Evans, Michelle P. Salyers, Jane Williams, and Hea-Won Kim. 2000. "Measurement of Fidelity in Psychiatric Rehabilitation." *Mental Health Services Research* 2, 2: 75-87.

Borman, Geoffrey D., Robert E. Slavin, Alan Cheung, Anne Chamberlain, Nancy Madden, and Bette Chambers. 2007. "Final Reading Outcomes of the National Randomized Field Trial of Success for All." *American Educational Research Journal* 44, 3: 701-731.

Box, George E. P., J. Stuart Hunter, and William G. Hunter. 1978. *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. New York: Wiley.

Browne, Angela, and Aaron Wildavsky. 1983. "Implementation as Mutual Adaptation." Pages 206-231 in Jeffrey L. Pressman and Aaron Wildavsky (eds.), *Implementation: How Great Expectations in Washington Are Dashed in Oakland*. Berkeley: University of California Press.

Bryk, Anthony S., and Stephen W. Raudenbush. 1988. "Toward a More Appropriate Conceptualization of Research on School Effects: A Three-Level Hierarchical Linear Model." *American Journal of Education* 97, 1: 65-108.

Bryk, Anthony S., Penny Bender Sebring, Elaine Allensworth, Stuart Luppescu, and John Q. Easton. 2010. *Organizing Schools for Improvement: Lessons from Chicago*. Chicago: University of Chicago Press.

Carroll, Christopher, Malcolm Patterson, Stephen Wood, Andrew Booth, Jo Rick, and Shashi Balain. 2007. "A Conceptual Framework for Implementation Fidelity." *Implementation Science* 2, 1: 40.

Castleman, Benjamin L., Karen D. Arnold, and Katherine Lynk Wartman. 2012. "Stemming the Tide of Summer Melt: An Experimental Study of the Effect of Post-High School Summer Intervention on Low-Income Students' College Enrollment " *The Journal of Research on Educational Effectiveness* 5, 1: 1-18.

Chase, Gordon. 1979. "Implementing a Human Services Program: How Hard Can It Be?" *Public Policy* 27, 4: 385-420.

Chen, Huey-Tsyh, and Peter Rossi. 1983. "Evaluating with Sense: The Theory-Driven Approach." *Evaluation Review* 7, 3: 283-302.

Collins, Linda M., Susan A. Murphy, Vijay N. Nair, and Victor J. Strecher. 2005. "A Strategy for Optimizing and Evaluating Behavioral Interventions." *Annals of Behavioral Medicine* 30, 1: 65-73.

Cook, Thomas D., William R. Shadish, and Vivian C. Wong. 2008. "Three Conditions Under Which Observational Studies Produce the Same Results as Experiments." *Journal of Policy Analysis and Management* 274, 4: 724-50.

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.

Damschroder, Laura J., David C. Aron, Rosalind E. Keith, Susan R. Kirsh, Jeffery A. Alexander, and Julie C. Lowery. 2009. "Fostering Implementation of Health Services Research Findings into Practice: A Consolidated Framework for Advancing Implementation Science." *Implementation Science* 4: 50.

Dane, Andrew. V., and Schneider, Barry. H. 1998. "Program Integrity in Primary and Early Secondary Prevention: Are Implementation Effects Out of Control?" *Clinical Psychology Review* 18, 1: 23-45.

Djebbari, Habiba, and Jeffrey Smith. 2008. *Heterogeneous Impacts in PROGRESA*. Bonn, Germany: The Institute for the Study of Labor (IZA).

Dobson, Keith S., and Brian F. Shaw. 1984. "The Use of Treatment Manuals in Cognitive Therapy: Experience and Issues." *Journal of Consulting and Clinical Psychology* 56, 5: 673-680.

Dorsett, Richard, and Philip K. Robins. 2011. *In-Work Support for Lone Parents: Using the UK ERA Demonstration to Examine Cross-Office Variation in Effectiveness*. Working Paper 105. Sheffield, UK: Department for Work and Pensions.

Duckworth, Angela L., Christopher Peterson, Michael D. Matthews, and Dennis R. Kelly. 2007. "Personality Processes and Individual Differences: Grit: Perseverance and Passion for Long-Term Goals." *Journal of Personality and Social Psychology* 92, 6: 1087-1101.

Duggan, Anne, Debra Caldera, Kira Rodriguez, Lori Burrell, Charles Rohde, and Sarah Shea Crowne. 2007. "Impact of a Statewide Home Visiting Program to Prevent Child Abuse." *Child Abuse and Neglect* 31, 8: 801-827.

Dumitrescu, Anca, Dan Levy, Cara Orfield, and Matt Sloan. 2011. *Impact Evaluation of Niger's IMAGINE Program*. Washington, DC: Mathematica Policy Research.

Durlak, Joseph A., and Emily P. DuPre. 2008. "Implementation Matters: A Review of Research on the Influence of Implementation on Program Outcomes and the Factors Affecting Implementation." *American Journal of Community Psychology* 41, 3-4: 327-350.

Durlak, Joseph A., Roger P. Weissberg, and Molly Pachan. 2010. "A Meta-Analysis of After-School Programs That Seek to Promote Personal and Social Skills in Children and Adolescents." *American Journal of Community Psychology* 45: 294-309.

Dusenbury, Linda, Rosaland Brannigan, Mathea Falco, and William B. Hansen. 2003. "A Review of Research on Fidelity of Implementation: Implications for Drug Abuse Prevention in School Settings." *Health Education Research* 18, 2: 237-256.

Elmore, Richard F. 1985. "Forward and Backward Mapping: Reversible Logic in the Analysis of Public Policy." Pages 33-70 in Kenneth Hanf and Theo A. J. Toonen (eds.), *Policy*

Implementation in Federal and Unitary Systems: Questions of Analysis and Design. Dordrecht, Netherlands: Martinus Nijhoff.

Fisher, Ronald A. 1935. *The Design of Experiments.* London: Oliver and Boyd.

Fixsen, Dean L., Sandra F. Naoom, Karen A. Blase, Robert M. Friedman, and Frances Wallace. 2005. *Implementation Research: A Synthesis of the Literature*. Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, the National Implementation Research Network.

Flay, Brian R., Sally Graumlich, Eisuke Segawa, James L. Burns, Michelle Y. Holliday, and Aban Aya Investigators. 2004. "Effects of Two Prevention Programs on High-Risk Behaviors among African American Youth: A Randomized Trial." *Archives of Pediatrics and Adolescent Medicine* 158, 4: 377-384.

Friedlander, Daniel. 1993. "Subgroup Impacts of Large-Scale Welfare Employment Programs." *The Review of Economics and Statistics* 75, 1: 138-143.

Friedlander, Daniel, and Philip K. Robins. 1997. "The Distributional Impacts of Social Programs." *Evaluation Review* 21, 5: 531-553.

Gamse, Beth C., Robin Tepper Jacob, Megan Horst, Beth Boulay, and Fatih Unlu. 2009. *Reading First Impact Study Final Report*. NCEE 2009-4039. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Glisson, Charles, John Landsverk, Sonja Schoenwald, Kelly Kelleher, Kimberly Eaton Hoagwood, Stephen Mayberg, and Philip Green. 2008. "Assessing the Organizational Social Context (OSC) of Mental Health Services: Implications for Research and Practice." *Administration and Policy in Mental Health and Mental Health Services Research* 35, 1-2: 98-113.

Gomby, Deanna S. 2005. *Home Visitation in 2005: Outcomes for Children and Parents*. Working Paper 7. Committee for Economic Development: Invest in Kids Working Group.

Gueron, Judith M., and Edward Pauly. 1991. *From Welfare to Work.* New York: Russell Sage Foundation.

Harms, Thelma, Richard M. Clifford, and Debby Cryer. 1998. *Early Childhood Childhood Environment Rating Scale-Revised Edition (ECERS-R)*. New York: Teachers College Press.

Heckman, James J. 2001. "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture." *Journal of Political Economy* 109, 4: 673-748.

Heckman, James J. 2005. "The Scientific Model of Causality." *Sociological Methodology* 35, 1: 1-97.

Heckman, James J., Jeffrey Smith, and Nancy Clements. 1997. "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies* 64, 4: 487-535.

Hjern, Benny, and Chris Hull. 1982. "Implementation Research as Empirical Constitutionalism." *European Journal of Political Research* 10: 105-115.

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

Hong, Guanglei, and Stephen W. Raudenbush. Forthcoming. "Heterogeneous Agents, Social Interactions, and Causal Inference."

Hulleman, Chris S., and David S. Cordray. 2009. "Moving From the Lab to the Field: The Role of Fidelity and Achieved Relative Intervention Strength." *Journal of Research on Educational Effectiveness* 2, 1: 88-110.

Hulleman, Chris S., Sara E. Rimm-Kaufman, and Tashia D. S. Abry. 2013. "Whole-Part-Whole: Construct Validity, Measurement, and Analytical Issues for Fidelity Assessment in Education Research." Pages 65-93 in T. Halle, A. Metz, and I. Martinez-Beck (eds.), *Applying Implementation Science in Early Childhood Programs and Systems*. Baltimore, MD: Paul H. Brookes Publishing Co.

Kazdin, Alan E. 1986. "Comparative Outcome Studies of Psychotherapy: Methodological Issues and Strategies." *Journal of Consulting and Clinical Psychology* 54, 1: 95-105.

Kemple, James J., Jason C. Snipes, and Howard S. Bloom. 2001. *A Regression-Based Strategy for Defining Subgroups in a Social Experiment*. New York: MDRC.

Kezar, Adrianna. 2011. "What Is the Best Way to Achieve Broader Reach of Improved Practices in Higher Education?" *Innovations in Higher Education* 36: 235-247.

Kisker, Ellen Eliason, Diane Paulsell, John M. Love, and Helen Raikes. 2002. *Pathways to Quality and Full Implementation in Early Head Start Programs*. Princeton: Mathematica Policy Research.

Klein, Alice, Prentice Starkey, Douglas Clements, Julie Sarama, and Roopa Iyer. 2008. "Effects of a Pre-Kindergarten Mathematics Intervention: A Randomized Experiment." *Journal of Research on Educational Effectiveness* 1, 3: 155-178.

Lake, Robin, Melissa Bowen, Allison Demeritt, Moira McCullough, Joshua Haimson, and Brian Gill. 2012. *Learning from Charter School Management Organizations: Strategies for Student Behavior and Teacher Coaching*. Seattle, WA: Center on Reinventing Public Education and Mathematica Policy Research.

Levy, Dan, Matt Sloan, Leigh Linden, and Harounan Kazianga. 2009. *Impact Evaluation of Burkina Faso's BRIGHT Program*. Washington, DC: Mathematica Policy Research.

Lipsey, Mark W., and David Holdzkom. 2008. "Workshop on Evaluating State and District Level Interventions." Institute of Education Sciences, Washington, DC, April 24, 2008.

Lipsey, Mark, W., Nana A. Landenberger, and Sandra J. Wilson. 2007. *Effects of Cognitive Behavioral Programs for Criminal Offenders*. Nashville, TN: Center for Evaluation Research and Methodology, Vanderbilt Institute for Public Policy Studies.

Madrian, Brigitte C., and Dennis F. Shea. 2001. "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior." *The Quarterly Journal of Economics* 116, 4: 1149-1187.

Mazmanian, Daniel A., and Paul A. Sabatier. 1989. *Implementation and Public Policy*. Lanham, MD: University Press of America.

Michalopoulos, Charles, and Christine Schwartz. 2000. *What Works Best for Whom: Impacts of 20 Welfare-to-Work Programs by Subgroup*. Washington, DC: U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation and Administration for Children and Families, and U.S. Department of Education.

Miller, William R., and J. Scott Tonigan. 1996. "Assessing Drinkers' Motivation for Change: The Stages of Change Readiness and Treatment Eagerness Scale (SOCRATES)." *Psychology of Addictive Behaviors* 10, 2: 81-89.

Moncher, Frank J., and Ronald J. Prinz. 1991. "Treatment Fidelity in Outcome Studies." *Clinical Psychology Review* 11, 3: 247-266.

Neyman, Jerzy. 1923. "Statistical Problems in Agricultural Experiments." *Journal of the Royal Statistical Society* 2, 2: 107-180.

Olds, David L., Harriet Kitzman, Carole Hanks, Robert Cole, Elizabeth Anson, Kimberly Sidora-Arcoleo, Dennis W. Luckey, Jr., Charles R. Henderson, John Holmberg, Robin A. Tutt, Amanda J. Stevenson, and Jessica Bondy. 2007. "Effects of Nurse Home Visiting on Maternal and Child Functioning: Age-9 Follow-Up of a Randomized Trial." *Pediatrics* 120, 4: e832-e845.

Patel, Reshma, and Ireri Valenzuela. Forthcoming. *Early Findings from the Performance-Based Scholarship Demonstration in Arizona.* New York: MDRC.

Paulsell, Diane, Sarah Avellar, Emily Sama Martin, and Patricia Del Grosso. 2010. *Home Visiting Evidence of Effectiveness Review: Executive Summary*. Princeton, NJ: Mathematica Policy Research.

Pianta, Robert C., Andrew J. Mashburn, Bridget B. Hamre, Jason T. Downer, Oscar A. Barbarin, Donna Bryant, and Diane M. Early. 2008. "Measures of Classroom Quality in Prekindergarten and Children's Development of Academic Language and Social Skills." *Child Development* 79, 3: 732-749.

Quandt, Richard. 1972. "A New Approach to Estimating Switching Regressions." *Journal of the American Statistical Association* 67, 338: 306-310.

Raudenbush, Stephen W. 2008. "Advancing Education Policy by Advancing Research on Instruction." *American Education Research Journal* 45, 1: 206-230.

Raudenbush, Stephen W., and Xiaofeng Liu. 2000. "Statistical Power and Optimal Design for Multisite Randomized Trials." *Psychological Methods* 5, 2: 199-213.

Reardon, Sean, and Stephen Raudenbush. Forthcoming. "Under What Assumptions Do Site by Treatment Instruments Identify Average Causal Effects?" *Sociological Methods and Research*.

Reardon, Sean, Fatih Unlu, Pei Zhu, and Howard S. Bloom. 2012. *Bias and Bias Correction in Multisite Instrumental Variables Analysis of Heterogeneous Mediator Effects*. New York: MDRC.

Riccio, James, Nadine Dechausay, David Greenberg, Cynthia Miller, Zawadi Rucks, and Nandita Verma. 2010. *Toward Reduced Poverty Across Generations: Early Findings from New York City's Conditional Cash Transfer Program*. New York: MDRC.

Richburg-Hayes, Lashawn, Thomas Brock, Allen LeBlanc, Christina Paxson, Cecilia Elena Rouse, and Lisa Barrow. 2009a. *Rewarding Persistence: Effects of a Performance-Based Scholarship Program for Low-Income Parents*. New York: MDRC.

Richburg-Hayes, Lashawn, Paulette Cha, Monica Cuevas, Amanda Grossman, Reshma Patel, and Colleen Sommo. 2009b. *Paying for College Success: An Introduction to the Performance-Based Scholarship Demonstration*. New York: MDRC.

Rollnick, Stephen, Nick Heather, Ruth Gold, and Wayne Hall. 1992. "Development of a Short 'Readiness to Change' Questionnaire for Use in Brief, Opportunistic Interventions Among Excessive Drinkers." *British Journal of Addiction* 87, 5: 743-754.

Rollnick, Stephen, Paul Kinnersley, and Nigel Stott. 1993. "Methods of Helping Patients with Behaviour Change." *British Medical Journal* 307, 6897: 188-190.

Rossi, Peter A., Mark W. Lipsey, and Howard E. Freeman. 2004. *Evaluation: A Systematic Approach.* Thousand Oaks, CA: Sage.

Rothwell, Peter M. 2005. "Subgroup Analysis in Randomised Control Trials: Importance, Indications and Interpretation." *Lancet* 365: 176-186.

Roy, Andrew D. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3, 2: 135-146.

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 for Causal Effects: The Role of Randomization." *Annals of Statistics* 6, 1: 34-58.

Rubin, Don. 1986. "Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers?" *Journal of the American Statistical Association* 81, 396: 961-962.

Scrivener, Susan, Colleen Sommo, and Herbert Collado. 2009. *Getting Back on Track: Effects of a Community College Program for Probationary Students*. New York: MDRC.

Scrivener, Susan, Michael J. Weiss, and Colleen Sommo. 2012. *What Can a Multifaceted Program Do for Community College Students? Early Results from an Evaluation of Accelerated Study in Associate Programs (ASAP) for Developmental Education Students*. New York: MDRC.

Sechrest, Lee B., and Robin Redner. 1979. *Strength and Integrity of Treatments in Evaluation Studies*. Washington, DC: National Criminal Justice Reference Service.

Sechrest, Lee B., Stephen G. West, Melinda A. Phillips, Robin Redner, and William Yeaton. 1979. "Some Neglected Problems in Evaluation Research: Strength and Integrity of Treatment." Pages 15-35 in Lee B. Sechrest, Stephen G. West, Melinda A. Phillips, Robin Redner, and William Yeaton (eds.), *Evaluation Studies Review Annual*. Thousand Oaks, CA: Sage.

Skoufias, Emmanuel. 2005. *PROGRESA and Its Impacts on the Welfare of Rural Households in Mexico*. Research Report 139. Washington, DC: International Food Policy Research Institute.

Tinto, Vincent. 1998. *Learning Communities and the Reconstruction of Remedial Education in Higher Education. Paper prepared for the Ford Foundation and U.S. Department of Education Conference on Replacing Remediation in Higher Education.* Stanford, CA: Stanford University.

Van Meter, Donald S., and Carl E. Van Horn. 1975. "The Policy Implementation Process: A Conceptual Framework." *Administration and Society* 6, 4: 445-488.

Visher, Mary, Emily Schneider, Heather Wathington, and Herbert Collado. 2010. *Scaling Up Learning Communities: The Experience of Six Community Colleges*. New York: MDRC.

Visher, Mary G., Michael J. Weiss, Evan Weissman, Timothy Rudd, and Heather D. Wathington. 2012. *The Effects of Learning Communities for Students in Developmental Education: A Synthesis of Findings from Six Community Colleges*. New York: National Center for Postsecondary Research.

Ware, Michelle, and Reshma Patel. 2012. *Does More Money Matter? An Introduction to the Performance-Based Scholarship Demonstration in California*. New York: MDRC.

Weiss, Carol. 1997. "How Can Theory-Based Evaluation Make Greater Headway." *Evaluation Review* 21, 4: 501-524.

Zachry Rutschow, Elizabeth, Lashawn Richburg-Hayes, Thomas Brock, Genevieve Orr, Oscar Cerna, Dan Cullinan, Monica Reid Kerrigan, Davis Jenkins, Susan Gooden, and Kasey Martin. 2011. *Turning the Tide: Five Years of Achieving the Dream in Community Colleges*. New York: MDRC.

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 City and Oakland, California, 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 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 proactively 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 ex-offenders and people with disabilities, 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.