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

Michael J. Weiss
Howard S. Bloom
Thomas Brock

Abstract

Evaluations of public programs in many fields reveal that different types of programs—or different versions of the same program—vary in their effectiveness. Moreover, a program that is effective for one group of people might not be effective for other groups, and 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. 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. © *2014 by the Association for Public Policy Analysis and Management.*

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. Consequently, much is now being learned about the average effects of specific programs for specific groups of people. As this evidence base has grown, so has the realization that different types of programs—or different versions of the same program—vary in their effectiveness. Moreover, a program that is effective for one group of people might not be effective for other groups, and 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 about program effectiveness. With this in mind, a small number of studies have begun 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 it found no 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

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

Journal of Policy Analysis and Management, Vol. 00, No. 0, 1–30 (2014)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.21760

of goals, and activities that were explicitly tied to these goals (Durlak, Weissberg, & Pachan, 2010).

- 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 clients personalized attention—had earnings effects that were much larger than average. In addition, the study found that the context of the programs 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, & Riccio, 2003).

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 (e.g., Abadie, Angrist, & Imbens, 2002; Bitler, Gelbach, & Hoynes, 2006; Bryk & Raudenbush, 1988; Djebbari & Smith, 2008; Friedlander & Robins, 1997; Heckman, 2001; Heckman, Smith, & Clements, 1997; Raudenbush & Liu, 2000), evaluation research and public policy analysis have to date mainly focused on the average effects of programs and paid far less systematic attention to explaining variation in effects.

To address this challenge, 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, treatment fidelity, and program implementation (a glossary of terms used throughout this text is provided at the end of the paper). 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 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 empirical examples to illustrate points. We rely disproportionately on projects conducted by MDRC, an organization best known for conducting randomized control trials in the education, employment, and social service fields, since these are the programs and evaluations that we know best. We expect the issues we raise to apply to many other programs and policies, and hope that readers will think of how this framework relates to their own experiences.

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 use comes from the statistical literature on causal effects and is based on the

concept of *potential outcomes*.[2] Potential outcomes for a person are the outcomes that he 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.[3] In defining this effect, it is assumed that each person has two potential outcomes: (1) that which he would experience if he were assigned to (or offered) the program and (2) that which he would experience if he were not assigned to (or not offered) the program. The first of these is referred to as the *treated outcome*, and the second is referred to as the *untreated counterfactual outcome*, or *counterfactual* for short.

For example, consider the causal effect on earnings of assigning a 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) his future earnings if he were assigned to the program and (2) his future earnings if he were not assigned to the program. This difference represents the *value added* by the program offer over the counterfactual state of the world.

Unfortunately for evaluators, it is not possible to observe both potential outcomes for a person simultaneously because people can only experience one condition (and thus one potential 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 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 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 not—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. It is also possible for strong quasi-experiments to approach the rigor of a randomized trial (Cook, Shadish, & Wong, 2008). Regardless of how these average 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.[4]

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

[2] 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).

[3] In the statistics literature this is referred to as the effect of intent to treat (Angrist, Imbens, & Rubin, 1996).

[4] Often 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.

who differ in their background characteristics, geographic location, and so on. 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 geographic locations, and administrative units. It is then possible to apply the preceding definition of an average effect of a program offer to a client subgroup or a program site. Variation in these average effects across subgroups, sites, or both is what we refer to as variation in program effects, or *effect variation* for short. The second section of this 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 backwards (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 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 program's effects are estimated by comparing *outcomes* for a program group to 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 future earnings, whereas an intermediate outcome[5] might be the rate of client participation in job-search assistance. 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 *program effect*).

Continuing to work right to left in Figure 1, one can see two boxes labeled *treatment received* by those 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 the cause of program effects—a point that cannot be overemphasized and that we return to frequently.

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 (e.g., experiencing an HIV advertising campaign compared with not experiencing such a campaign) affects a mediator (e.g., awareness), and the mediator, in turn, affects the target outcomes (e.g., following safer sex practices). In this way, mediators are part of a causal chain that leads 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 made available with or without access to a program. The link between treatment offered and treatment received represents service *take-up*.

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,

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

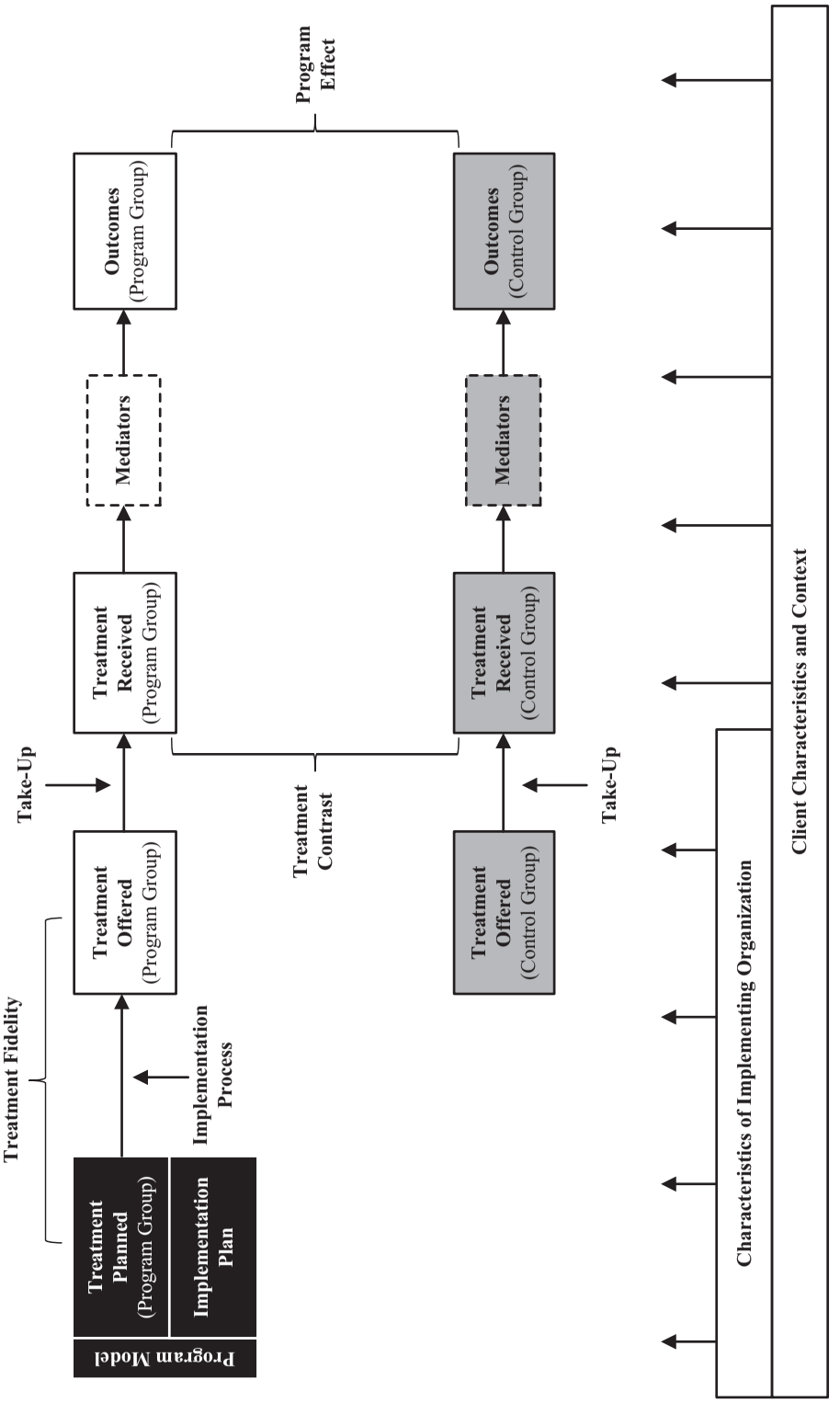

Figure 1. A Conceptual Framework for Studying Variation in Program Effects, Treatment Contrasts, and Implementation.

these two boxes represent the *program model*.[6] A program model is sometimes specified by the program's developers.[7] It comprises a blueprint for client services (treatment plan) and a blueprint for implementing these services (implementation plan). A program model is adopted (and might be adapted or changed) by an implementing organization to varying degrees based on the *context* and the *characteristics of the implementing organization* (see bottom of Figure 1), as well as the specificity of the program model. 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, which in turn influences 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 the planned treatment and the received treatment, rather than the difference between the planned treatment and the offered treatment (e.g., Carroll et al., 2007; Cordray & Pion, 2006; Dane & Schneider, 1998; Durlak & DuPre, 2008; Hulleman & Cordray, 2009). The decision of whether to extend treatment fidelity from the planned client services to the offered client services or from the planned client services to the received client services is not critical; however, appreciating the relationship between services that are planned for, offered to, and received by clients 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 might 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 (e.g., a school or a welfare office). 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 a program's *client characteristics*, as well as characteristics of its 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 to integrate the study of variation in program effectiveness, treatment contrasts, and implementation. The following sections examine how variation in one feature of the framework is related to variation in other features. We organize the discussion by proceeding from right (downstream) to left (upstream) in the framework, and zooming in on each portion of the framework as it is discussed.

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

[6] Program models may also specify a target population, which is not identified separately in the figure.
[7] As noted later, it is sometimes important to distinguish between the program model that is specified by model developers (e.g., a national model in a scale-up effort) and the program model that is specified by a given site (a local model).

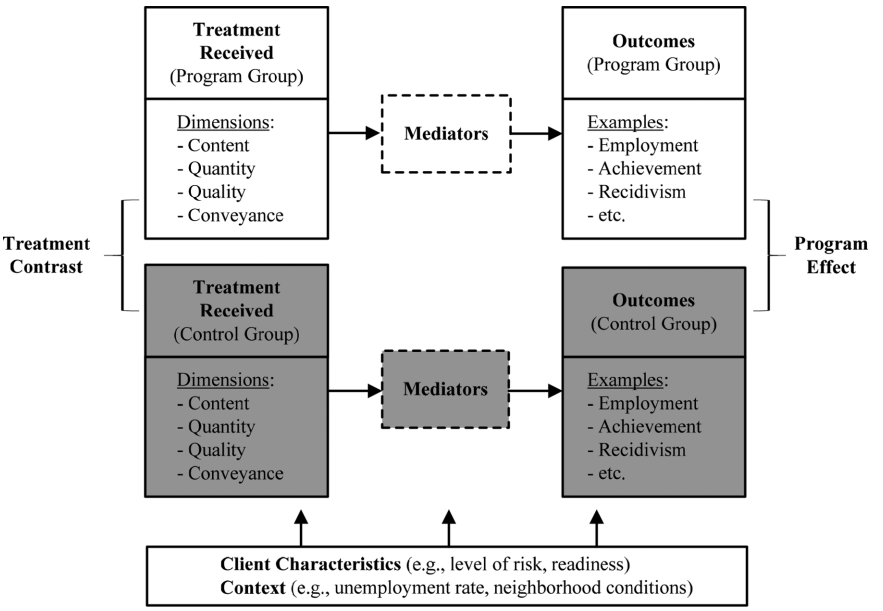

Figure 2. The Treatment Contrast and Effect (Right-Hand Side of Figure 1).

work, leading scholars like Peter Rossi and Carol Weiss have urged that carefully constructed program theory guide program evaluation.[8] Based on their writing and our own experience, we offer the following simple 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 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, we suggest that 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 causes program effects, while client characteristics and program context moderate the size 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

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

moderators. Below we discuss the three proximal sources of variation in program effects in detail.

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 plus any other existing services, and (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[9] 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 client outcomes only by changing their treatment receipt, and thereby producing a treatment contrast. If a treatment contrast does not exist—if clients have the exact same experience in a program as they would have had if they were not 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 11th 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 mid-year.

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 mandatory, 30-minute, mid-year 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 and two optional ones, and 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 produce no treatment contrast for them as well.

Because an individual student can only experience one of these treatment packages, it is not possible to observe an individual treatment contrast. However, it is

[9] 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*.

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.

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 67 percent of the students assigned to the new program graduated within two years compared to 66 percent of the students assigned to the existing program, the best estimate of the program's effect on graduation rates is an increase of 1 percentage point.

Consider a second high school testing the same program. Its teachers make a concerted effort to get students 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 estimated average treatment 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—a much larger impact.

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 11th 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 some evidence that additional guidance counseling for 11th 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.[10] If such a trial produced suggestive 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: By what delivery mode, when, and by whom is each service provided?

Although these dimensions can overlap, we believe that they provide a useful checklist to run through for each program component. We are thus less concerned

[10] Bloom, Hill, and Riccio (2003) provide a detailed example of such an analysis.

about splitting hairs over categorizing each program component into these four dimensions, and more concerned about identifying the contrast as comprehensively as possible. Because the treatment contrast reflects the difference in services experienced under the program and counterfactual condition, the four dimensions of content, quantity, quality, and conveyance must be measured for program group members as well as for the counterfactual condition (i.e., control group members in an experiment). Below we discuss each dimension in more detail and with examples.

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 might include new curricula (e.g., see Agodini & Harris, 2010; Klein et al., 2008). Examples from other program areas include cognitive behavioral therapy (CBT) for correctional inmates to reduce their violent behavior (Lipsey, Landenberger, & Wilson, 2007); informational campaigns to educate teenagers about the risks of alcohol and substance abuse, unprotected sexual activity, or smoking (e.g., Flay et al., 2004; Trenholm et al., 2008); home visiting services to teach low-income mothers and pregnant women how to improve the health and development of their infants and young children (Duggan et al., 2007; Gomby, 2005; Olds et al., 2007; Paulsell et al., 2010); and financial incentives for students and their parents to promote better health care and increased educational engagement (Riccio et al., 2010).

A concrete example of the content 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 ASAP is comprehensive advisement. In a survey administered both to program and control group members, study participants were asked what topics were covered in advising sessions (e.g., course selection and personal matters). While simplistic, these questions reveal the proportion of students who covered each topic in the presence or absence of ASAP, a key element of the treatment contrast (Scrivener, Weiss, & 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 (Cordray and Pion, 2006; Dane and Schneider, 1998; 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 the treatment (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 counseling under the existing system, this is the prevalence of counseling under the status quo.

The frequency of services represents how often they are received during any given period (a day, a week, a month, a year, etc.). For example, CBT might be provided daily, tutoring in English might be provided twice a week, 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.

Lastly, the duration of services received is the total period of time during which they are received. For example, in the Reading First program, 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 well 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, and are delivered on a timely and predictable basis. The assessment of treatment quality thus tends to be more subjective than the assessment of other program features.

The Classroom Assessment Scoring System (CLASS) (Pianta et al., 2008) is an example of an assessment of the quality of socio-emotional and instructional interactions between teachers and students. The preschool version of CLASS has three key domains: (1) emotional support, (2) classroom organization, and (3) instructional support. In an evaluation designed to improve the quality of teacher–student interactions along these dimensions (e.g., MyTeaching Partner), CLASS's measure of quality could be an appropriate indicator of the treatment contrast.

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 on-line experiences, or through hardcopy 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 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, local daycare, 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 daycare 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(s) of the treatment contrast that does not correspond with the core components of the intervention, but is nonetheless important to measure. For example, the introduction of a CBT program in a group home for delinquent boys will mean less of some other activities that would be happening in the business-as-usual condition. The treatment contrast in such a scenario would include increased CBT and a reduction of, for example, group therapy.

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 84 new Small Schools of Choice, or SSCs, in New York City public high schools. An initial study found that SSCs increased high school graduation rates by about 9.5 percentage points for disadvantaged students of color (Bloom & Unterman, forthcoming); 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 Department of Education conducts at all high schools.

Limitations on data collection can make it difficult to quantify the 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 attempting to measure, for both program and control group members, the key services that 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. Oftentimes, however, services do not immediately and directly cause changes in target outcomes. Between treatment receipt and the target outcomes lay mediators that are part of a hypothesized causal chain of events that yield program effects. A mediator of a program effect is a mechanism through which the program causes its effect on client outcomes. For example, an advertising campaign designed to reduce HIV transmission must first be received, then raise awareness, then cause behavioral changes such as a reduction in the number of sexual partners, increased condom usage, or both—ultimately leading to the target outcome, a reduction in HIV transmission. 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, number of sexual partners, condom usage) under the treated and counterfactual conditions, in addition to the treatment contrast and target outcome. Importantly, *causal* mediation analysis can be very complex, requiring strong assumptions that are difficult to prove (Bullock, Green, & Ha, 2010; Imai, Tingley, & Keele, 2009). By collecting data on the hypothesized mediators under the treated and counterfactual conditions—in addition to the treatment contrast and target outcome—researchers can advance this type of work.

Thus far we have discussed the first of three proximal sources of variation in program effects: the treatment contrast, which causes program effects. Now we turn to our second source of variation in program effects, client characteristics, which moderates program effects. When we discuss a client characteristic as a moderator of the program effect, we have in mind a characteristic that predicts differences in program effects, but is not itself affected by the program.[11]

[11] In statistical terms, this means that the value of a moderator must be determined exogenously to the program being studied.

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 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. 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 of interest to many fields, including welfare-to-work, medical research, and education (Friedlander, 1993; Gueron & Pauly, 1991; Kemple, Snipes, & Bloom, 2001; Michalopoulos & Schwartz, 2000; Rothwell, 2005). There are three competing hypotheses: (1) that programs work best for the participants who are the most disadvantaged, since they have the greatest margin for improvement; (2) that programs work best for the participants who are the least disadvantaged, since they might best be able to utilize program services; or (3) that programs work best for 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 in order to improve.

Indicators of disadvantage or risk vary 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 or drinking.

A growing body of research points to the importance of noncognitive skills in predicting a variety of outcomes for children and adults. Roberts (2009) defines noncognitive skills as "relatively enduring patterns of thoughts, feelings and behaviors that reflect the tendency to respond in certain ways under certain contexts." They include the ability of an individual to (1) avoid distractions while working on long-term goals (e.g., learning a new skill, quitting smoking, advancing in a career); (2) deal productively with stressors, including setbacks and failures; and (3) interacting well with others. The empirical literature suggests that conscientiousness—the tendency to be dependable, persistent, and organized—has the highest predictive power of positive life outcomes among different types of noncognitive skills. Emotional stability also has a high predictive power on educational attainment and health outcomes (Borghans et al., 2012).

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 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 client types that are of particular interest to policymakers. For example, if policymakers are concerned about high unemployment rates among veterans, examining a job training program's effects on veterans would make sense.

While moderator analyses based on client characteristics can allow for a description of the type of individuals in a study who benefited more or less from a treatment, such analyses typically do not uncover why groups were differentially affected. Moreover, if a moderator analysis reveals that, for example, a preschool program has a positive effect on boys and zero effect on girls, the finding may be misattributed to gender. It is possible that the program is effective for children who have exhibited aggressive behavior in the past and has no effect for children who have not exhibited aggressive behavior in the past. If gender is correlated with past aggression, then the moderation could manifest itself through the more easily observed gender. Such misattribution can result in poor theory building. It may also result in imprecise targeting of resources. In the above example, resources would be better spent if the intervention were targeted to aggressive children rather than to boys. However, it would still be more efficient to target boys than to be gender neutral, even if the reason for targeting boys were misunderstood.

Program Context as a Moderator

A third 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 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 for clients—unless the program's relationship with local employers allows it to have access to job openings that clients would not otherwise find. On the other hand, with 5 percent unemployment, a program might have an easier time placing clients in jobs, but it could be that clients would just as easily find work on their own. In this case, the program would 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. An extensive analysis of welfare and employment data from the Current Population Survey led to a similar conclusion (Herbst, 2008). More empirical work is needed to fully understand the influence of this contextual factor.

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 contrast), and thus contribute to variation in program effects. Some of these factors (e.g., 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 (e.g., 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.

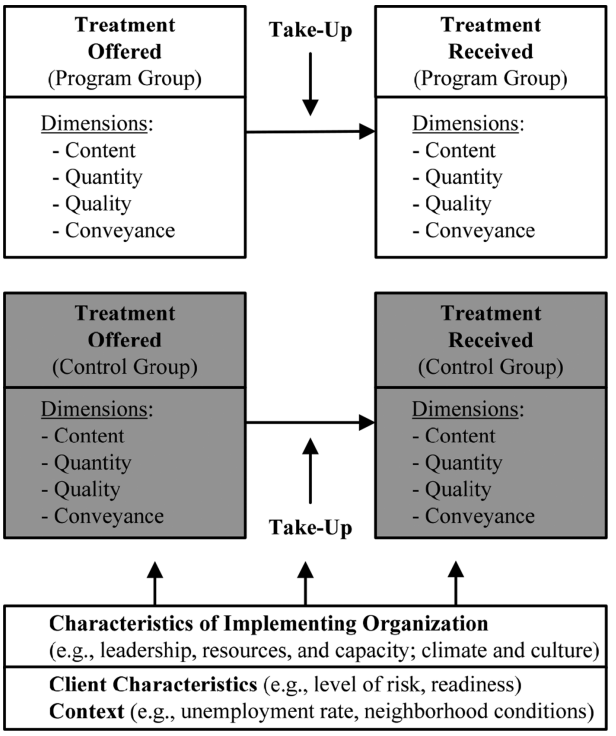

Figure 3. Treatment Offered and Treatment Received (Middle of Figure 1).

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 do not show up (i.e., if take-up is low), then the treatment contrast will be diminished, as will the chance that the program produces 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 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 these exist on a continuum. Similarly, there is a range of client engagement and responsiveness to those services (an element of service quality).

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 developed to explain

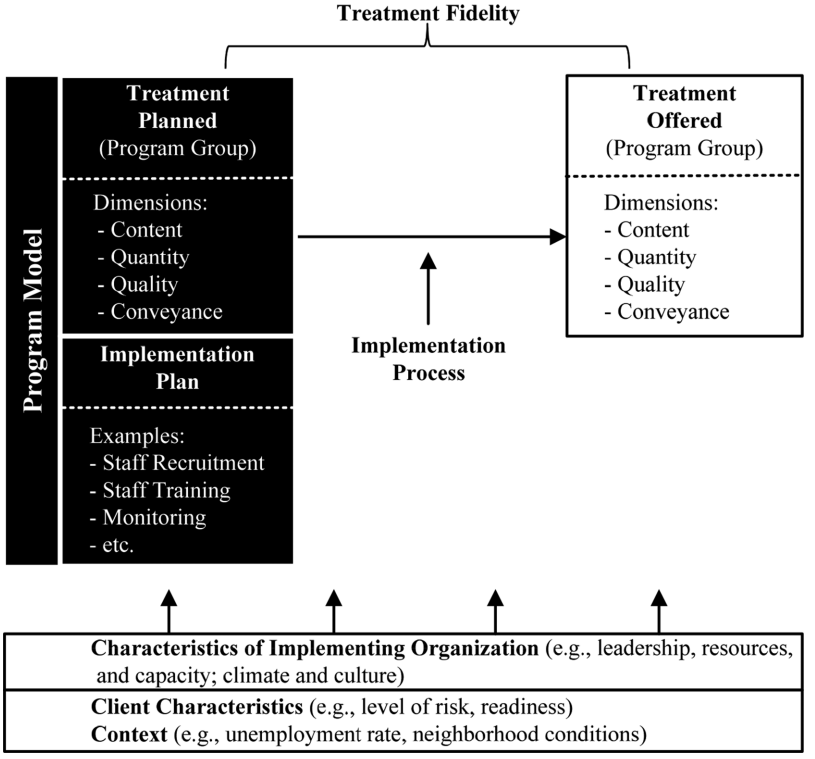

Figure 4. Connecting the Treatment Offered to Program Implementation and Treatment Fidelity (Left-Hand Side of Figure 1).

why some programs succeed in delivering services as planned and others fail to do so (see, e.g., Damschroder et al., 2009; Dane & Schneider, 1998; Durlak & DuPre, 2008; Elmore, 1985; Mazmanian & Sabatier, 1989; Van Meter & Van Horn, 1975). To facilitate our discussion of these issues, Figure 4 zooms in to the portion of our proposed framework where they are 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(s).[12] 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.

The treatment plan describes what a program is expected to offer to clients. At one extreme, the treatment plan may simply be a statement of the problem that a program is supposed to address, the population that it is supposed to serve, and the

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

principles by which it is supposed to be run. At the opposite extreme, the treatment plan may provide prescriptions for the content, quantity, quality, and conveyance of 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. As with treatment plans, implementation plans vary widely in their level of specificity. 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 metaframeworks that try to spell out the processes and factors that contribute to successful program implementation and treatment fidelity (Damschroder et al., 2009; Durlak & Dupre, 2008; Fixsen et al., 2005). Each of the frameworks emphasizes the importance of staff training and the need for organizational supports that will help staff carry out their responsibilities.

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.[13] This model, now operating in over 1,400 U.S. elementary schools, was developed over many years and has been evaluated extensively (e.g., 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 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.[14] The Knowledge is Power Program (KIPP) charter management organization is another example of an educational reform with a high degree of specificity (Angrist et al., 2012).

Frequently, however, program developers offer only general guidelines on the treatment and implementation plans and expect practitioners to fill in the details. For example, consider the recently evaluated Building Strong Families (BSF) program, which taught relationship skills to unmarried parents who were expecting or recently had a baby. All BSF program sites had three components that they were expected to implement, including group sessions on relationship skills, individual support from family coordinators, and assessment and referral to support services. Outside of these elements, the sites could choose how and where they recruited couples, the curriculum used in the group education component, and how they provided the family coordinator and referral services. Not surprisingly, an evaluation of BSF revealed considerable variation in how the programs were operated (Wood 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 (Cordray & Pion, 2006; Dane & Schneider, 1998; Dusenbury et al., 2003; Hulleman & Cordray, 2009).[15] Fixsen

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

[14] A blank copy of this contract was provided to the authors by Dr. Nancy Madden, president and co-founder of *Success for All*.

[15] For a definition of this term and related terms used elsewhere in the literature, see the prior section.

et al. (2005) argue that implementation is facilitated by having a program or practice that is well defined and clearly operationalized. An earlier generation of 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, e.g., Mazmanian & Sabatier, 1989).[16] While these arguments are compelling, there may be limits to the degree of specificity that is desirable in a treatment or implementation plan. Some researchers (e.g., Bardach, 1980; Hjern & Hull, 1982) argue that organizations responsible for carrying out program implementation have agendas, needs, priorities, and goals that 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 (1983) and Kezar (2011) have suggested that policymakers or model developers and program operators need to learn from each other and be willing to compromise on implementation—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. At this point, there is insufficient data to know how much specificity or flexibility is optimal. Researchers can shed light on these issues by noting the level of specificity within a service or implementation plan and determining the degree to which it is associated with treatment fidelity (and ultimately, treatment contrasts and effects).

Characteristics of the Implementing Organization

As indicated above, local service-providing organizations—schools, welfare offices, health care providers, and others—typically have some say over how an implementation plan is carried out, and can influence the quality of implementation through their actions and inactions. A number of researchers have identified characteristics that may make some organizations more predisposed to implementing an intervention well (e.g., Chase, 1979; Damschroder et al., 2009; Durlak & 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

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

checklist or be turned into a rating system for assessing the strength of program leadership during the implementation process.

Because implementation of a new policy or program often takes time, stability of leadership may also be important. To the extent that there is frequent turnover at the top of an organization, there will likely be less attention to executing the implementation plan. For example, in a five-year evaluation of the Achieving the Dream initiative—a foundation-funded effort designed to help community colleges undertake a multiyear institutional improvement process—nearly half of the colleges involved experienced at least one change in presidential and senior administrator leadership, and several colleges experienced greater turnover. 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 with implementation (Kisker et al., 2002).

Sufficient Organizational Resources and Capacity. Implementation requires adequate fiscal resources for organizations to hire staff, acquire space, and purchase whatever goods and services are needed to operate a new program. Money alone does not guarantee implementation success, however. Mazmanian and Sabatier (1989, p. 26) suggest that there is likely "a threshold level of funding" necessary to achieve programmatic objectives, and that while the probability of achieving those objectives may rise with increased funding, there may also be a point of saturation, beyond which additional dollars provide no further value.

The dedication and skills of an organization's staff are other factors to consider. For example, in a study of school improvement in Chicago, Bryk et al. (2010) identified *professional capacity* as one of five core conditions related to school improvement.[17] 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. 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—not just among the leadership. 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.

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

[17] The other core conditions were inclusive leadership, 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.

divisions of an organization or may be influenced by external events or conditions (such as an election cycle or a temporary drop in revenue.) 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). For example, an organization may be more receptive to implementing a new program if it has 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. Incentives and rewards aligned with the intervention (e.g., bonuses for organizations and staff that adopt a new practice) may also help.

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. An organization might be expected to do better at implementation when its culture is rated as proficient (e.g., one in which clients' needs are placed first) as opposed to rigid or resistant (e.g., one in which staff have little 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 (e.g., characterized by cooperation and a sense of accomplishment) rather than stressful (e.g., 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 education and social service programs to determine whether 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 may be more likely to implement an intervention successfully if there is an external overseer who is specifically charged with monitoring this implementation (Fixsen et al., 2005; Mazmanian & 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 be performed by a strong program officer at a foundation, a government board, an intermediary organization, or a national office that owns a 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 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.[18]

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 operate (staff recruitment,

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

selection, training, monitoring, supervision, along with various supports). These same elements are reflected in the enacted implementation process that occurs for any given program at any given site.[19] The next questions to ask are the following: 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, for centuries (Cordray & Pion, 2006). Cordray and Pion (2006) provide a history and survey of this issue in the field of evaluation research, which they begin with work in the 1970s by Lee B. Sechrest and his colleagues (e.g., Sechrest & 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 are 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 school students from 13 science classrooms taught by eight teachers at three high schools).[20] This provides 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.[21]

One problem that often arises when this work is discussed is that the treatment contrast is conflated with treatment fidelity. This 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 a dimension of treatment fidelity. We do not. To avoid this ambiguity, our proposed framework (Figure 1) depicts and labels the treatment contrast and treatment fidelity, visually displaying their relationship and distinctiveness. To the extent that treatment fidelity has the potential to influence program effects, it does so through one half of the treatment contrast—the treatment received by the program group. However, it is unrelated to the other half of the treatment contrast—the treatment received by control or comparison group members. Thus, theoretically a program can be implemented

[19] Although not depicted in the present framework, Hulleman, Rimm-Kaufman, and Abry (2013) conceptualize *implementation fidelity* as answering the following 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.

[20] 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.

[21] For example, Dusenbury et al. (2003) provide an extensive review of research on implementation fidelity in school-based drug abuse programs.

with perfect treatment fidelity and yet have little or no treatment contrast (if in the counterfactual condition the same services are received). Conversely, a program can be implemented with low treatment fidelity and yet have a large treatment contrast.

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 and 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 (e.g., 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. One place to start is to consider whether the services planned by program sites vary in terms of their content, quantity, quality, and conveyance. If sites do not plan the same services, then it is unlikely they will offer the same services. Cross-site variation in services offered 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 (e.g., 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.

A program's treatment plan, implementation process, and take-up determine the services received by its clients, 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

Table 1. Examples of measures for elements of the conceptual framework.

Construct	Possible measures
Treatment planned, treatment offered, and treatment 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 services are provided.*	• Prevalence, frequency, intensity, and duration (e.g., 30-minute tutoring sessions, offered 5 days per week, lasting 30 minutes each)
Quality *How well services are 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 *By what mode, when, and by whom services are 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])
Program context	
Location type	• Rural/urban/suburban
Economic indicators	• Unemployment rate • Average household income
Safety	• Crime rate
Sociodemographic variables	• Ethnic/racial composition • Percent foreign-born • Percent 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

Table 1. (*Continued*)

Construct	Possible measures
Characteristics of the implementing organization	
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 to 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

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 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 in an environment where similar services are available, and thus has little chance of 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 (which is substantial) in math and French, 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 (Dumitrescu et al., 2011; Levy et al., 2009). 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.[22] Although BRIGHT and IMAGINE schools both intended to provide higher quality education than the alternative, the difference between

[22] 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 and 74 percent.

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 describe differences in program effects between the efficacy and effectiveness tests. It can also help describe differences in program effects across sites, contexts, and client types in the effectiveness test.

Concluding Thoughts: Closing the Loop

This paper began with two examples of evaluation research—a meta-analysis of 69 after-school programs and a secondary analysis of data collected from 59 welfare-to-work programs—that demonstrated it is possible to learn about more than simply the average effect of a program and the extent to which it was implemented faithfully. These studies share several important features. First, both capitalize on separate estimates of program effects for multiple sites (after-school programs and 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 (e.g., active learning strategies and a strong and consistent employment message), client characteristics (participating children and welfare recipients), and program contexts (after-school programs 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 & Patel, 2012). Similarly, the National Evaluation of Welfare-to-Work Strategies conducted side-by-side tests of work first and human capital development programs in the same welfare offices to learn which approaches would be most effective in reducing welfare payments and increasing employment and earnings among welfare recipients (Hamilton, 2002).

The preceding exemplar studies should provide encouragement for researchers to pursue similar studies of other social and educational programs. It should also encourage policymakers and practitioners who want to understand why programs produce the effects that they do, and what options they might consider to increase program effectiveness. 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.

Glossary

Term	Definition
Context	The broader environment in which a program operates.
Implementation plan	The instructions or guidelines for how to set up and operate a program.
Implementing organization	The organization that delivers a program's treatment or services.
Mediator	A mechanism by which a program causes its effects on client outcomes.
Moderator	A variable (e.g., a characteristic of the clients, the implementing organizations, or the contexts) that predicts variation in the effect of the program, but is not itself affected by the program.
Program implementation	"The process of putting a defined practice or program into practical effect" (Fixsen et al., 2005, p. 82).
Program model	The treatment or services planned for a program's clients *and* the plan for implementing these services.
Take-up	The link between the treatment or services made available to clients and the treatment or services they receive.
Treatment dimensions	
Content	*What* client services are provided.
Quantity	*How much* client services are provided.
Quality	*How well* client services are provided.
Conveyance	*By what mode, when, and by whom* client services are provided.
Treatment	
Planned	The services that a program is expected to offer its clients.
Offered	The services that are made available to clients.
Received	The services that are received by clients.
Treatment contrast	The difference between treatment or services *received* with and without access to the program.
Treatment fidelity	The difference between the *treatment planned* for clients and the *treatment offered* or made available to clients.

MICHAEL J. WEISS is a Senior Associate, MDRC, 16 East 34th Street, New York, NY 10016 (e-mail: michael.weiss@mdrc.org).

HOWARD S. BLOOM is Chief Social Scientist, MDRC, 16 East 34th Street, New York, NY 10016 (e-mail: howard.bloom@mdrc.org).

THOMAS BROCK is Commissioner, National Center for Education Research, Institute of Education Sciences, 555 New Jersey Avenue, NW, Washington, DC 20208 (e-mail: Thomas.Brock@ed.gov).

This paper was written when Thomas Brock was employed by MDRC, and it does not necessarily reflect the views of the United States, the U.S. Department of Education, or the Institute of Education Sciences, where he is currently employed.

ACKNOWLEDGMENTS

The ideas, writing, editing, and review of this paper involved input from 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.

REFERENCES

Abadie, A. K., Angrist, J. D., & Imbens, G. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. Econometrica, 70, 91–117.

Agodini, R., & Harris, B. (2010). An experimental evaluation of four elementary school math curricula. Journal of Research on Educational Effectiveness, 3, 199–253.

Angrist, J. D., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91, 444–455.

Angrist, J. D., Dynarski, S. M., Kane, T. J., Pathak, P. A., & Walters, C. R. (2012). Who benefits from KIPP? Journal of Policy Analysis and Management, 31, 837–860.

Bardach, E. (1980). The implementation game: What happens after a bill becomes a law. Cambridge, MA: MIT Press.

Bitler, M. P., Gelbach, J. B., & Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. American Economic Review, 96, 988–1012.

Blakely, C. H., Mayer, J. P., Gottschalk, R. G., Schmitt, N., Davidson, W. S., Roitman, D. B., & Emshoff, J. G. (1987). The fidelity adaptation debate: Implications for the implementation of public sector social programs. American Journal of Community Psychology, 15, 253–268.

Bloom, H. S., Hill, C. J., & Riccio, J. A. (2003). Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments. Journal of Policy Analysis and Management, 22, 551–575.

Bloom, H. S., & Unterman, R. (forthcoming). Can small high schools of choice improve education prospects for disadvantaged students? Journal of Policy Analysis and Management.

Borghans, L., Diris, R., Heckman, J. J., Kautz, T., & Weel, B. (2012). Fostering non-cognitive skills to promote lifetime success. Unpublished manuscript.

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A., Madden, N., & Chambers, B. (2007). Final reading outcomes of the National Randomized Field Trial of Success for All. American Educational Research Journal, 44, 701–731.

Browne, A., & Wildavsky, A. (1983). Implementation as mutual adaptation. In J. L. Pressman & A. Wildavsky (Eds.), Implementation: How great expectations in Washington are dashed in Oakland (pp. 206–231). Berkeley, CA: University of California Press.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. American Journal of Education, 97, 65–108.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2010). Organizing schools for improvement: Lessons from Chicago. Chicago, IL: University of Chicago Press.

Bullock, J. G., Green, D. P., & Ha, S. E. (2010). Yes, but what's the mechanism (don't expect an easy answer). Journal of Personality and Social Psychology, 98, 550–558.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. Implementation Science, 2, 40. Retrieved February 18, 2014, from http://www.biomedcentral.com/content/pdf/1748-5908-2-40.pdf.

Chase, G. (1979). Implementing a human services program: How hard can it be? Public Policy, 27, 385–420.

Chen, H.-T., & Rossi, P. (1983). Evaluating with sense: The theory-driven approach. Evaluation Review, 7, 283–302.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which observational studies produce the same results as experiments. Journal of Policy Analysis and Management, 274, 724–750.

Cordray, D. S., & Pion, G. M. (2006). Treatment strength and integrity: Models and methods. In R. R. Bootzin & P. E. McKnight (Eds.), Strengthening research methodology: Psychological measurement and evaluation (pp. 103–124). Washington, DC: American Psychological Association.

Damschroder, L. J., Aron, D. C., Rosalind, E. K., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. Implementation Science, 4, 50. Retrieved February 18, 2014, from http://www.implementationscience.com/content/pdf/1748-5908-4-50.pdf.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? Clinical Psychology Review, 18, 23–45.

Djebbari, H., & Smith, J. (2008). Heterogeneous impacts in PROGRESA. Bonn, Germany: The Institute for the Study of Labor (IZA).

Dobson, K. S., & Shaw, B. F. (1984). The use of treatment manuals in cognitive therapy: Experience and issues. Journal of Consulting and Clinical Psychology, 56, 673–680.

Duggan, A., Caldera, D., Rodriguez, K., Burrell, L., Rohde, C., & Crowne, S. S. (2007). Impact of a statewide home visiting program to prevent child abuse. Child Abuse and Neglect, 31, 801–827.

Dumitrescu, A., Levy, D., Orfield, C., & Sloan, M. (2011). Impact evaluation of Niger's IMAGINE program. Washington, DC: Mathematica Policy Research.

Durlak, J. A., & DuPre, E. P. (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, 327–350.

Durlak, J. A., Weissberg, R. P., & Pachan, M. (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, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. Health Education Research, 18, 237–256.

Elmore, R. F. (1985). Forward and backward mapping: Reversible logic in the analysis of public policy. In K. Hanf & T. A. J. Toonen (Eds.), Policy implementation in federal and unitary systems: Questions of analysis and design (pp. 33–70). Dordrecht, Netherlands: Martinus Nijhoff.

Fisher, R. A. (1935). The design of experiments. London, UK: Oliver and Boyd.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (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, B. R., Graumlich, S., Segawa, E., Burns, J. L., Holliday, M. Y., & 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, 377–384.

Friedlander, D. (1993). Subgroup impacts of large-scale welfare employment programs. Review of Economics and Statistics, 75, 138–143.

Friedlander, D., & Robins, P. K. (1997). The distributional impacts of social programs. Evaluation Review, 21, 531–553.

Gamse, B. C., Jacob, R. T., Horst, M., Boulay, B., & Unlu, F. (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, C., Landsverk, J., Schoenwald, S., Kelleher, K., Hoagwood, K. E., Mayberg, S., & Green, P. (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, 98–113.

Gomby, D. S. (2005). Home visitation in 2005: Outcomes for children and parents. Working Paper 7. Retrieved February 18, 2014, from http://www.readynation.org/docs/ivk/report_ivk_gomby_2005.pdf.

Gueron, J. M., & Pauly, E. (1991). From welfare to work. New York, NY: Russell Sage Foundation.

Hamilton, G. (2002). Moving people from welfare to work: Lessons from the National Evaluation of Welfare-to-Work Strategies. New York, NY: MDRC.

Heckman, J. J. (2001). Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture. Journal of Political Economy, 109, 673–748.

Heckman, J. J. (2005). The scientific model of causality. Sociological Methodology, 35, 1–97.

Heckman, J. J., Smith, J., & Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. Review of Economic Studies, 64, 487–535.

Herbst, C. M. (2008). Do social policy reforms have different impacts on employment and welfare use as economic conditions change? Journal of Policy Analysis and Management, 27, 867–894.

Hjern, B., & Hull, C. (1982). Implementation research as empirical constitutionalism. European Journal of Political Research, 10, 105–115.

Holland, P. (1986). Statistics and causal inference. Journal of the American Statistical Association, 81, 945–960.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. Journal of Research on Educational Effectiveness, 2, 88–110.

Hulleman, C. S., Rimm-Kaufman, S. E., & Abry, T. D. S. (2013). Whole-part-whole: Construct validity, measurement, and analytical issues for fidelity assessment in education research. In T. Halle, A. Metz, & I. Martinez-Beck (Eds.), Applying implementation science in early childhood programs and systems (pp. 65–93). Baltimore, MD: Paul H. Brookes Publishing Co.

Imai, K., Tingley, D., & Keele, L. (2009). A general approach to causal mediation analysis. Psychological Methods, 15, 309–334.

Kazdin, A. E. (1986). Comparative outcome studies of psychotherapy: Methodological issues and strategies. Journal of Consulting and Clinical Psychology, 54, 95–105.

Kemple, J. J., Snipes, J. C., & Bloom, H. S. (2001). A regression-based strategy for defining subgroups in a social experiment. New York, NY: MDRC.

Kezar, A. (2011). What is the best way to achieve broader reach of improved practices in higher education? Innovations in Higher Education, 36, 235–247.

Kisker, E. E., Paulsell, D., Love, J. M., & Raikes, H. (2002). Pathways to quality and full implementation in early Head Start programs. Princeton, NJ: Mathematica Policy Research.

Klein, A., Starkey, P., Clements, D., Sarama, J., & Iyer, R. (2008). Effects of a pre-kindergarten mathematics intervention: A randomized experiment. Journal of Research on Educational Effectiveness, 1, 155–178.

Levy, D., Sloan, M., Linden, L., & Kazianga, H. (2009). Impact evaluation of Burkina Faso's BRIGHT program. Washington, DC: Mathematica Policy Research.

Lipsey, M. W., Landenberger, N. A., & Wilson, S. J. (2007). Effects of cognitive behavioral programs for criminal offenders. Nashville, TN: Center for Evaluation Research and Methodology, Vanderbilt Institute for Public Policy Studies.

Mazmanian, D. A., & Sabatier, P. A. (1989). Implementation and public policy. Lanham, MD: University Press of America.

Michalopoulos, C., & Schwartz, C. (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.

Neyman, J. (1923). Statistical problems in agricultural experiments. Journal of the Royal Statistical Society, 2, 107–180.

Olds, D. L., Kitzman, H., Hanks, C., Cole, R., Anson, E., Sidora-Arcoleo, K., Luckey, D. W., Henderson, C. R., Holmberg, J., Tutt, R. A., Stevenson, A. J., & Bondy, J. (2007). Effects of nurse home visiting on maternal and child functioning: Age-9 follow-up of a randomized trial. Pediatrics, 120, e832–e845.

Paulsell, D., Avellar, S., Martin, E. S., & Del Grosso, P. (2010). Home visiting evidence of effectiveness review: Executive summary. Princeton, NJ: Mathematica Policy Research.

Pianta, R. C., Mashburn, A. J., Hamre, B. B., Downer, J. T., Barbarin, O. A., Bryant, D., & Early, D. M. (2008). Measures of classroom quality in prekindergarten and children's development of academic language and social skills. Child Development, 79, 732–749.

Quandt, R. (1972). A new approach to estimating switching regressions. Journal of the American Statistical Association, 67, 306–310.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. Psychological Methods, 5, 199–213.

Riccio, J., Dechausay, N., Greenberg, D., Miller, C., Rucks, Z., & Verma, N. (2010). Toward reduced poverty across generations: Early findings from New York City's conditional cash transfer program. New York, NY: MDRC.

Richburg-Hayes, L., Brock, T., LeBlanc, A., Paxson, C., Rouse, C. E., & Barrow, L. (2009a). Rewarding persistence: Effects of a performance-based scholarship program for low-income parents. New York, NY: MDRC.

Richburg-Hayes, L., Cha, P., Cuevas, M., Grossman, A., Patel, R., & Sommo, C. (2009b). Paying for college success: An introduction to the performance-based scholarship demonstration. New York, NY: MDRC.

Roberts, B. W. (2009). Back to the future: Personality and assessment and personality development. Journal of Research in Personality, 43, 137–145.

Rothwell, P. M. (2005). Subgroup analysis in randomised control trials: Importance, indications and interpretation. Lancet, 365, 176–186.

Roy, A. D. (1951). Some thoughts on the distribution of earnings. Oxford Economic Papers, 3, 135–146.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Education Psychology, 66, 688–701.

Rubin, D. (1978). Bayesian Inference for causal effects: The role of randomization. Annals of Statistics, 6, 34–58.

Scrivener, S., Weiss, M. J., & Sommo, C. (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, NY: MDRC.

Sechrest, L. B., & Redner, R. (1979). Strength and integrity of treatments in evaluation studies. Washington, DC: National Criminal Justice Reference Service.

Sechrest, L. B., West, S. G., Phillips, M. A., Redner, R., & Yeaton, W. (1979). Some neglected problems in evaluation research: Strength and integrity of treatment. In L. B. Sechrest, S.

G. West, M. A. Phillips, R. Redner, & W. Yeaton (Eds.), Evaluation studies review annual (pp. 15–35). Thousand Oaks, CA: Sage.

Tinto, V. (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.

Trenholm, C., Devaney, B., Fortson, K., Clark, M., Bridgespan, L. Q., & Wheeler, J. (2008). Impacts of education on teen sexual activity, risk of pregnancy, and risk of sexually transmitted diseases. Journal of Policy Analysis and Management, 27, 255–276.

Van Meter, D. S., & Van Horn, C. E. (1975). The policy implementation process: A conceptual framework. Administration and Society, 6, 445–488.

Ware, M., & Patel, R. (2012). Does more money matter? An introduction to the performance-based scholarship demonstration in California. New York, NY: MDRC.

Weiss, C. (1997). How can theory-based evaluation make greater headway. Evaluation Review, 21, 501–524.

Wood, R. G., McConnell, S., Moore, Q., Clarkwest, A., & Hsueh, J. (2012). The effects of building strong families: A healthy marriage and relationship skills education program for unmarried parents. Journal of Policy Analysis and Management 31, 228–252.

Zachry Rutschow, E., Richburg-Hayes, L., Brock, T., Orr, G., Cerna, O., Cullinan, D., Reid, M. K., Jenkins, D., Gooden, S., & Martin, K. (2011). Turning the tide: Five years of achieving the dream in community colleges. New York, NY: MDRC.