THE ESSENTIAL CONNECTION:

Using Evaluation to Identify Programs Worth Replicating

by
Kathryn Furano
Linda Z. Jucovy
David P. Racine
Thomas J. Smith

Replication and Program Strategies, Inc.
One Commerce Square
2005 Market Street, Suite 900
Philadelphia, PA 19103
(215) 557-4482
(215) 557-4485 fax

Table of Contents

THE ESSENTIAL CONNECTION: USING EVALUATION TO IDENTIFY PROGRAMS WORTH REPLICATING

Replication, the act of transporting an effective social program to new sites, is a familiar idea in both public policy and the thinking of philanthropies. Yet, for a variety of reasons, replication remains a neglected strategy.[1] Programs have lacked the resources and know-how to undertake it, and foundations have generally preferred to invest their funding in new programs rather than extensions of existing, successful ones. Nevertheless, confronted today with the parallel realities of increasing social distress and decreasing amounts of public money to address it, foundation officers and public

---

[1] The experience of programs that have won The Ford Foundation Innovations in State and Local Government Awards indicates the challenges facing attempts to replicate in one important arena: the public sector. The winners – ten each year – are selected because they are innovative, address significant concerns, result in proven benefits, and show promise for being successfully replicated. Each winner receives a $100,000 grant to strengthen the program locally and to encourage its replication in other states and communities. In addition, the programs receive extensive national publicity. Still, a study of 26 winners found that six did not replicate at all and nine were replicated in only one-to-five sites. The remaining programs were replicated in larger numbers of sites. Many of the staffs of the award-winning

officials should logically see replication as an important and innovative way to leverage investments in effective social initiatives.

In its simplest form, replication involves two steps: 1) developing credible knowledge about the effectiveness of programs and their potential for broader adoption, and 2) based on that knowledge, reproducing those programs that have been found to "work." Within these neat categories, though, the realities get considerably more complex. Reproduction or transfer can take many forms covering a broad spectrum. At one end of the continuum lie fairly "open" efforts in which practitioners rather freely adapt key program elements to their local situations. At the other end are more "closed" undertakings that virtually clone specific models.

In fact, the term "replication" is used to encompass a wide-ranging set of situations and possibilities. At times, it refers to adopting a single component of a multi-faceted program. For example, operators of juvenile justice programs, increasingly convinced that mentoring should play a key role in efforts to curb recidivism, search for existing models of mentoring that can be replicated within their own programs. At other times, "replication" means furthering the adoption of "concepts." In this case, guiding principles, rather than distinct program

---

programs reported that they did not have the time, funding or expertise to follow through on replication efforts. See Public/Private Ventures. 1994. **"Replicating Exemplary Programs: Lessons from Experience."** Philadelphia: Public/Private Ventures. For a broader, conceptual view of the challenges replication poses see The Conservation Company and Public/Private Ventures. 1993. **Building from Strength: Replication as a Strategy for Expanding Social Programs that Work.** Philadelphia: Replication and Program Services, Inc.

features, are advanced. A good example would be the national Coalition of Essential Schools. It encourages schools to adopt a series of nine principles, which, in turn, become the foundation for building an education reform program that is "unique to each school."

Replication is an investment, and as with all investments, there's an element of risk. The risks in replication can take various forms, too. The program being reproduced might seem attractive - for example, it's had good reviews in the press - but, in fact, is not really accomplishing its aims. Or an essential though not obvious element of the program might get left out of the replication strategy. Or the program itself is quite sound, but its designer might turn out to be a poor manager of a multi-site operation. Defining - and reducing - such risks are integral to the process of thinking through any replication initiative.

To that end, the following pages outline ways in which funders and others interested in successful programs can use evaluation to help make informed decisions about whether to proceed with an investment in replication. Appropriate evaluation develops credible knowledge about a program's effectiveness and transferability. It employs both description (of, for example, program implementation) and measurement of indicators that suggest the extent to and the ways in which a program is achieving its goals and why.

The purpose in this paper, though, is not to provide a primer on evaluation. Ample resources already exist on that topic. Rather, the thrust here is to illuminate, in a practical way, the essential connection between the traditional interest in determining whether a program works and the less developed interest in deciding whether to further that program's reach. While the "science" of this connection is still

in its infancy, enough is known to begin laying down some guidance for those who have to decide whether a particular replication is worth undertaking. At the least, doing so should give readers a better feel than they may have now for the specific challenges that replication entails. Ideally, we hope the following discussion will prompt those with an interest in replication - including those unconvinced of its potential - to engage the subject in a deeper and more concrete way.

Because rather different issues are involved in making decisions about "concept" replication, we say little of direct relevance to that topic here. Instead, we focus on programs with relatively clear organizational structures, relatively well-defined, related clusters of activities, and relatively concrete outcomes. Arguably, at least, such programs lend themselves more readily than do other types to reliable replication. Thus, they represent a more straightforward case for showing how evaluation can contribute to informed decision-making by funders and others interested in supporting the expansion of worthwhile social endeavors. "Concept" replication will be the subject of a future contribution in this occasional paper series.

IDENTIFYING REPLICATION-WORTHY PROGRAMS

Few objective methods are currently used to help select programs for replication. Often today, programs that succeed in securing public or private funding to replicate do so more because of marketing initiative, the charisma of their creator, or the reputation of their organizational home than as a result of evidence of their effectiveness and replicability. While this need not mean such programs are, in fact, ineffective or difficult to reproduce, it does suggest that evaluative evidence often gets short shrift in decisions to proceed with replication.[2] Needless to say, decisions not based on some form of systematic evaluation are inherently riskier than those that are.

---

[2] The limited role that evaluation plays in determining what programs should be replicated is suggested by an informal survey Replication and Program Strategies, Inc. did of foundation officials, other nonprofit leaders, and social researchers. Respondents were asked by mail to name programs which they believe are effective and replicable. Of the sixty programs identified, follow-up reviews indicated that fully half were able to offer no systematic evidence of their effectiveness or replicability. Another third had some supporting data, usually basic statistics on program activities or the results of old assessments done when the program was first tested, but no current evaluation looking at outcomes and how the program produced them. Only a sixth of the nominated programs appeared to have been the subject of a recent, formal evaluation.

At the same time, of course, it is impossible to arrive at completely objective validation of any program. Random assignment impact studies, recognized by many researchers as the highest standard of program evaluation, are desirable when they can be done and ought, perhaps, to be the standard against which other evaluation techniques get judged. But such studies are also very specialized and costly. In addition, while rigorous random assignment research can produce credible impact data, the evidence is typically compiled and analyzed over a long period of time - often several years. That time is something social programming, driven as it is by the need to address problems which already exist, cannot always afford. As one leading practitioner has noted, "Program operators need the best available information to make decisions, but none of us has the luxury of waiting for perfect information."[3]

It seems clear that some relatively objective, albeit imperfect, means are needed to guide decisions about worthwhile program investments. What standards or performance criteria, then, can be used to decide, within a reasonable frame of time, whether a program is a good candidate for replication?

Credible evidence will rarely, if ever, prove that a particular model is superior to all other alternatives. But what it must do, at a minimum, is make a well-reasoned and - supported case that:

---

[3] Marion Pines. Quoted in Richard H. De Lone. 1990. **Replication: A Strategy to Improve the Delivery of Education and Job Training Programs.** Philadelphia: Public/Private Ventures, p. 18.

- ✦ the results of the intervention, as measured against program objectives, are moving in the desired direction because of the intervention;
- ✦ the program is as good as or better than other known alternatives seeking to achieve the same purposes;
- ✦ no matter what the results of the program are, it is better than no program at all; and
- ✦ the model has a good chance of working in other settings.

In fact, the evaluation data needed to support a claim that a program works well - that it is "proven" or "successful" and transferable - will differ according to the program's purpose and context. In some cases, participation alone may be an appropriate goal and evaluation measure. This could be the case where society intrinsically values the service offered by the program - for example, providing access to prenatal care or tutoring adult non-readers. These are situations in which providing the service where nothing existed before is often considered a success in itself. In other situations, quantified evidence about short- and perhaps longer-term outcomes may be warranted as a condition of investment. This would apply, for instance, to dropout prevention efforts or job training and placement programs, where the services given are a means to measurable, hoped-for ends (e.g., school completion, job retention), but they are not an end in themselves.

There are, of course, always more subjective factors that must be considered in decisions about a program's replication-worthiness and replicability. These include making judgments about the social or moral importance of the services the program is delivering and the political feasibility of replicating a specific program at a particular time. Some programs lack sufficient significance or are too controversial to merit the level

of investment replication typically entails. Reaching such conclusions, though, is usually more a matter of discernment than objective analysis of evidence.

While acknowledging the existence of more subjective factors and the fact that the kinds and degrees of evidence required to measure a program's worthiness and replicability will vary, the following standards can serve as a useful starting point for evaluations:

- ✦ Clear and plausible aims: The program to be replicated should have an unambiguous and defensible mission - a precise articulation of what the services or activities in the program are intended to do and why. The actual services, then, should logically address that mission. A vague or overly ambitious mission (e.g., to enable all poor people in the community to achieve self-sufficiency) - a not unusual phenomenon in social programming - may be a warning sign that the program is not specifiable to the degree usually needed for replication.

- ✦ Essential program features that are transferable: A replicable program will have an identifiable, coherent set of interrelated program elements. These might include, among other things, the target group, recruitment strategies, staffing patterns, the specific services offered and their duration. At the same time, the program cannot be based on factors that are so idiosyncratic or site-specific that it is essentially non-reproducible elsewhere.

  Idiosyncracy is sometimes the case with small programs that develop in response to a particular need in a specific, local setting. There, the personality of the program's founder or first director, or characteristics of the setting or

participants, might be contributing to the program's success in ways that cannot be adequately reenacted in other communities. Idiosyncratic qualities may also limit the replicability of programs of a very different kind: large, comprehensive ones. Such programs may often simply be too complex to be transferred in any fairly precise way to other sites.

- Descriptive data indicating program success: An attractive replication candidate should be evaluable in a manner that will produce evidence about its operational coherence and effectiveness. The program ideally will have data on both in-program indicators (i.e., coherence) and post-program outcomes (i.e., effectiveness).

  In-program indicators should provide evidence that the program is reaching those people for whom and in the way intended. These indicators might include, for example, rates of participant enrollment and retention; information concerning the demographic characteristics of participants; and data on their use of services and involvement in activities.

  Data on post-program outcomes should, at a minimum, yield evidence that the program attains "face validity" in achieving its operational objectives. Depending upon the aims of the program, such data might include, for example, job placements and 30- and 60-day retention rates; or attainment of GEDs; or placement of homeless families in permanent housing within a specified period of time; or increased access to home financing for low-income people; or greater first trimester use of prenatal care.

- Feasible requirements: The cost and logistical requirements of the program have to be reasonable. Costs must be perceived or shown to be less, or certainly no greater than, the expected benefits of the program. Even when this condition is generally met, programs with unusually high price tags will, in all likelihood, be quite difficult to sustain over the long haul as they face competition from lower cost alternatives and new demands on funders' limited resources. Logistical or operational needs also must be feasible for the type of program it is. The match between these needs and what the program tries to do should be clear and straightforward. "Ornate" models are often not only costly but hard to replicate as a practical matter.

If replication is to become a significant form of social investment, it will have to rely on evidence and criteria that fall short of "definitive." At the same time, this doesn't mean decisions about replication have to be based primarily on anecdotes about program success. If we are willing to accept - as reason dictates we should - some imprecision, a program's appropriateness for replication can be assessed adequately by using shorter-term performance measures and other, more qualitative evidence. The following sections discuss the two most prominent of these approaches: implementation reviews or studies and outcomes data.

IMPLEMENTATION REVIEWS

A careful look at the implementation of a program provides valuable information regarding its historical development, how it operates, and why it operates the way it does. An implementation review should identify the essential features of the model, the extent to which the program's activities reflect fidelity to that model, and - where actual program elements diverge from those defined by the model - how these alternative strategies might be achieving the desired objectives and why.

Understanding the Model

The theoretical model provides the backdrop against which the reality of a program is viewed. It sets forth why and how the program is supposed to work as a change strategy. Even programs which are not explicitly based on theory reflect implicit, theory-like assumptions that serve to justify or explain what they seek to do. Testing the adequacy of a program's underlying assumptions or theory of change is, or surely should be where it is not, one of the functions of program evaluation.[4]

---

[4]For an extended discussion of the case for theory-based evaluation see Carol Hirschon Weiss, "*Nothing as Practical as Good Theory: Exploring Theory-Based Evaluation for Comprehensive Community Initiatives for Children and Families.*" In J.P. Connell, A.C.

While an understanding of the model and its essential features will develop gradually as the program is being evaluated, a first step is to build a preliminary base of knowledge by reading any available materials that describe or shed light on the model - including grant proposals, reports to funders, and relevant research summaries. It is also desirable to look at models of other programs with similar aims (e.g., other mentoring programs for high school youth, other initiatives that provide outreach services to pregnant women) to define what makes this particular program different and conceivably more effective. Differentiating the program from "competing" approaches will, for obvious reasons, often be a critical aspect in establishing its worthiness for replication.

Understanding the model requires being able to identify both the essential program features that should be built into any replication of it and those features that are more pliable or even optional. Few issues are more important when trying to transport a program to new settings than distinguishing between the "fixed" and the "variable."

Take the case of an after-school program for high school students that includes job training and rehabilitation of housing in the community. This specific program might include all of the following features:

- A partnership between the school, the community-based program that oversees the program, and a community development corporation that provides access to the housing that participants work to rehabilitate.

Kubisch, L.B. Schorr, and C.H. Weiss (Eds.). 1995. **New Approaches to Evaluating Community Initiatives**. Washington, D.C.: The Aspen Institute.

- A structured linkage between students' in-school curriculum and after-school work experience. The linkage might include both academic skills and common themes: for example, a connection to community service that is part of students' in-school learning as well as their after-school program.
- A teacher/facilitator who is both in the classroom and at the worksite - thus, strengthening the school and after-school linkage. In this case, that person might be a carpenter who participates in the in-school carpentry class and teaches carpentry skills in the after-school program.
- A life skills component that includes employment maturity training, such as being on time for work.
- A staff person who is a member of a trade union and thus provides contacts with organized labor and potential connections with apprenticeship programs.
- A low instructor-to-student (1 to 4) ratio at the worksite. This means the program must have the capacity to ensure that there are a sufficient number of appropriate instructors available to accommodate the number of youth who are working on any given afternoon.

A key challenge of the implementation review is discerning which of these features are essential, meaning they must be present if the program is to succeed in other locations, and which can be adapted or treated as options. In this case, essential features might be the structured linkage between the in-school and after-school experiences (on the assumption that cumulative learning is more effective), along with having one staff member who connects both experiences (on the assumption that this facilitates continuity for and building trust with students). A more adaptable feature might be the

partnership - the first element above - which could be modified if, for example, a community development corporation operated the program itself. Similarly, having a staff person who is a union member is a feature that could vary among sites, depending on the strength of construction trade unions in each particular location.

Beyond this first step of gathering information about the "essential" model, implementation reviews may employ several data collection methods which build upon and mutually reinforce one another.

Examination of Program Documents And Records

A program's paperwork - the records it keeps, how it keeps them, and what the data in them say - is both its history and a map of its workings. While the term "paperwork" resonates with notions of bureaucracy and endless stacks of materials, the "truth" about a program will often reside here as much as anywhere.

While the kinds of records that a program keeps will vary - depending, among other things, on its mission and goals - an examination of records will typically focus on the following kinds of data:

- Administrative data – Administrative records may consist of: 1) the program budget, including its funding sources, how it allocates its money, and why the money is allocated in this way; 2) job descriptions; 3) correspondence, memos, press clippings and anything else that is revealing about the program's relationship with the community; and 4) recruitment and other promotional materials.

- Participant data – Individual participant files, along with compilations of participant data, will normally include information about: 1) who the program is serving, 2) their activity in the program, and 3) the results of that activity. Records that delineate the first type of information include application forms, eligibility information, and demographic data. Such data allow evaluators to see who is actually enrolling in the program and if, in fact, the program is attracting the model's target audience. Data should also be available that track actual involvement in the program. This would typically be found in attendance records, along with staff evaluations of participants and other performance reviews.

- Other relevant data on services and accomplishments – For example, a program that provides education and training services to community-based groups about urban environmental hazards should have records that provide details about materials disseminated, the number of education and training sessions provided, attendance at these sessions, and any follow-up contacts with the community-based groups that have been involved.

How a program collects and stores information is also a relevant facet of program implementation: it speaks directly to the level of importance placed on having a permanent, reviewable record of program activity and performance. This is pertinent not only to accountability but also to the program's ability to develop a factual grasp of why and how it works. If data are not being adequately collected and stored, that is reason at least to question whether the program's operators

have access to appropriate means for fully understanding what they are doing.

Of course, a program that is technologically advanced, for example, with computerized participant databases and budgetary spreadsheets, is likely to have more capacity to reflect activity on paper than a program that is limited by its resources to using legal pads and a calculator. Small programs, because they may lack staff and technology, are often restricted in their ability to collect, store, and retrieve data. In this case, surveys (see below) can sometimes be used as part of an evaluation to capture needed data.

Observations

Observations provide the picture that complements the words, or data, revealed through review of records. Site visits to a program provide the opportunity to examine firsthand several important qualities not usually evident on paper:

- the interaction between staff and participants (for example, between facilitator and students in a GED classroom, between the carpenter/instructor and youth at the worksite, between an outreach worker and the pregnant women she is visiting);
- interaction among participants in the program;
- participants' degree of engagement in the program (how, for example, they go about asking questions, or the extent to which they show initiative in solving problems); and
- the overall physical environment and tangible resources for the program.

Observations at program headquarters can also be revealing about a program's management structure. Site

visitors can note, for example, the layout of office space and what that implies about staff relationships, how staff use their time, and how staff interact with one another. These may all be factors that turn out to bear on the program's replicability. For example, an office that appears to be disorganized or where reporting relationships are unclear may indicate an aspect of the program which would need to be more tightly structured before replication is pursued.

Observations should be made based on an understanding of what the program model seeks to achieve and how that work is to be accomplished. The site visitor should have in mind, to the extent they've been identified, the theories and assumptions that underlie the program, and should test these in the course of observing actual operations. Taking the example used earlier of the after school program for high school students, an observer could look at how the expected continuity between classroom and worksite is, in fact, being achieved or if it is being achieved. Some factors, such as this one, can often be adequately comprehended only by observing them first-hand.

Observations can provide an illuminating "snapshot" of the activity of a program. Yet, at the same time, the picture produced by observations may not be representative. While much valuable data can be derived by seeing a program "in action," there are inherent limitations as well.

Timing obviously can influence what is observed. While site visits should take place during a "typical day" in the program, observers must, nonetheless, operate under the assumption that what they see and hear on a one-day visit is simply what takes place during that particular day. If, for example, an essential program component is a 1-5 ratio of staff to youth, but there is a 1-10 ratio on the day of the visit, it is

important to know why. (Is this an intentional change from the model? Is it the result of a temporary staffing problem?). Similarly, if fewer than the expected number of participants show up, site visitors will want to know the reasons. Conversely, if everything seems to be working perfectly on the day of the site visit, that doesn't necessarily mean the program runs smoothly all the time.

Thus, it becomes important to try to place the day of a site visit in the larger context of ongoing program activities. Site visitors need to ask, for example, what a typical day is in the program and how long the participants they're observing have been involved. Indeed, it's a good idea, if affordable and otherwise feasible, to try to make several visits spaced over time. It is important as well to remember that the presence of an outsider - in this case, the site visitor - always has some effect on the dynamics of what is going on in a program during her or his visit.

Finally, it should be noted that observation will often expose the evaluator to more intangible aspects of the program which are less apt to show up in paper records or through other cognitively oriented methods of gathering information. For instance, an observer may see that participants seem to enjoy being around one another and with program staff, but she cannot quite figure out the reasons for this. Being sensitive to the presence and play of intangible factors when observing a program up close should alert the site visitor to issues that may merit further inquiry. A program that appears to work for reasons which are not well understood will not normally be a good candidate for replication.

Interviews

Interviewing moves evaluation beyond the static confines of what has been written about a program and reinforces or expands upon what has been learned through observation. In addition, through their interactions with staff, participants and other service recipients, interviewers can gain insight into how the program is perceived by the people most involved with it.

The program director: An interview with the program's director or head provides an opportunity to clarify the program's mission and to discuss issues such as budget, planned future growth, internal management capacity, and the program's reputation in the community. If the director has been with the program since its beginning, this is also a good opportunity to discuss implementation challenges that the program might have faced when it was starting up. Getting a handle on start-up challenges may yield insights of significance to launching the program elsewhere. At the same time, it is also worth emphasizing that directors often have less knowledge about day-to-day program operations than do other staff.

Staff: In talking with staff who are interacting directly with participants or other service recipients, interviewers will want to discuss actual job responsibilities, as opposed to what is written in job descriptions, along with such issues as staff participation in making program decisions, staff development needs and opportunities, job satisfaction, and commitment to the program. Staff interviews also furnish an opportunity to learn about staff members' perceptions of the program and their ideas concerning how elements of the model might be effectively altered. The information gathered from

interviewing staff will be useful input, not only to evaluation, but to any later effort undertaken to replicate the program, where carefully and clearly defining staffing requirements will often be critical to success.

Participants and other service recipients: Conversations with program participants or other service recipients can be particularly revealing about the extent to which a program is intentionally fulfilling its aims and is doing so in ways that can be reproduced if it were to be replicated. Determining why a participant enrolled in a program, for example, speaks not only to individual motivation, but also to the success of a particular recruitment effort or the clarity with which the program's purpose is conveyed to the public. In addition to asking why participants are in this program rather than others, the interviewer can discuss how long they have been in the program, why they are still in it, what they do in it, and how it could be different. In order to gain an understanding of participants' involvement in and grasp of program goals, the evaluator can also ask about specific program elements. Why, for example, does the program have you doing calisthenics every morning? Or how does the community benefit from your work fixing up this house? Or what has been your experience as a resident gaining access to the program's services? Comparing various participants' and recipients' responses to such questions should help to reveal overall patterns about the functioning of the program as a planned intervention.

Collaborating agencies: Staff of partnering agencies can offer a different perspective on the program within its larger community context - a view of the program which might shed

some new light on replicability, such as idiosyncracies of the model or whether it depends upon special resources that cannot easily be found in other places. Collaborating agencies will also have some perspective on whether the program operator has the capacity to function effectively as a replicator. Is the organization easy to collaborate with? Can staff effectively communicate to others about the program's goals and components? Does the organization follow through on its commitments? Does it have the capacity both to operate the existing program and to teach it to others? Can the organization withstand "being in the fishbowl" and endure the scrutiny that will be required when it demonstrates the model for adoption in other communities?

While interviews and observations will not provide the most "scientific" data concerning a program's effectiveness, they are a vital way of seeing how closely the program in practice reflects the program as planned. Good interviewers and observations can balance objectivity, judgment, and interpretation in using the evidence to determine how and why a program works.

Surveys

Pencil and paper surveys can produce information that is similar to what can be learned through interviews with staff and participants. However, unlike the latter, surveys lack the open-endedness and potential for nuance that person-to-person conversations allow. Nonetheless, surveys are an efficient tool for accumulating large amounts of data. Thus, if a program has not collected participant data that seems important in evaluating it - information on economic status, for

example - then a survey of participants may be a useful instrument for gathering that information.

Looking Beyond The Model

Using the model alone as a yardstick to measure how successfully a program has been implemented or is operating does not take into consideration the adjustments that program managers must make to fit the model to prevailing circumstances. In this regard, the model is less a rigid requirement than a map to be applied in the field as context dictates. Contextual factors - both external (for example, political climate, competing programs, positive or negative community response) and internal (staffing changes, funding limitations, technical capacity) - affect how closely program implementation can mirror what is planned. A program's variations from the model may, in fact, be of significant interest in any one of a number of ways.

Variations may occur in any facet of the model, and they provide an opportunity to ask questions and learn. If staff's actual jobs differ in important ways from their formal job descriptions, why is that? Is it a result of an internal circumstance in the program, such as another employee being on leave, or has the job evolved to meet changing realities, and if so, are these realities likely to be at play in other communities as well? The latter case would suggest that the model itself might need to be refined before it is replicated elsewhere.

Program features might have been purposefully adapted so that they are no longer entirely consistent with the model. For example, while Big Brothers/Big Sisters of America advocates that adult volunteers commit themselves to a minimum of a one-year relationship with the youth they mentor, some local

sites - one located near a military base, another in a college community - have experimented with allowing shorter-term relationships. The purpose of this adaptation was to attract more volunteers from among the military personnel and college students located nearby and to match them with young people who would otherwise have a longer wait for a mentor. The success of the experiments suggests not that the model per se should be altered, but that this particular feature, which might previously have been considered inflexible, can be adjusted to enhance or preserve program performance in certain contexts.

In fact, it will often be unnecessarily limiting to decide not to replicate a program because of variations from the model. A model's value can be measured, in part at least, by how adaptable and improvable it is. Unless a program is going to be replicated in almost identical contexts - an unlikely situation - it must have some degree of flexibility built into it. While the goals and objectives that drive a program should be consistent from place to place, the manner by which those aims are achieved will, to some extent, almost inevitably require some variation if success is to be reproduced and the program is to evolve wherever it exists. A key challenge in replication efforts comes, then, in trying to identify the types of variations that should be encouraged or tolerated.[5] Adaptations that are identified when conducting program implementation studies provide a way to begin to address this challenge. Indeed, these

---

[5]There is some evidence that adaptations which supplement the program model may contribute more to improved effectiveness than those which change the model itself. In a study of seven nationally disseminated social programs, Blakely and his colleagues found that local revisions to the model were unrelated to effectiveness while local additions tended to enhance effectiveness. See Craig H. Blakely, et al. (1987). *"The Fidelity-Adaptation Debate: Implications for the Implementation of Public Sector Social Programs."* American Journal of Community Psychology, 15(3): 253-268.

adaptations may enhance the potential for successful replication by broadening the range of contexts into which the program can be effectively transplanted.

PROGRAM OUTCOMES

While implementation studies reveal how a program operates and the extent to which those operations seem to be an effective approach to addressing the program's aims, knowledge of outcomes is also important in reaching judgments about replication-worthiness and replicability. In some cases, the outcomes may be fairly straightforward and not difficult to quantify: How many children received immunizations as a result of the program? How many pregnant women received first trimester prenatal care? How many adults remained in literacy classes until they became functional readers? In many cases, though, outcomes are more difficult to define and measure. What, for example, is a "good" outcome of a job training and placement program for out-of-school youth? While placement in a job is the obvious first measure, the next measure must focus on how long the former participant in the program retains that job. Is 30 days an adequate measure? 60 days? 6 months? Further, the nature of the job itself may become an issue. Is it a low-wage, dead-end job, or one with the potential for future training and advancement?

Programs that attempt to generate multiple benefits raise additional questions. A program that combines housing rehabilitation with job training and employability skills

development for in-school youth has the potential for several positive outcomes. How many houses have been rehabilitated, and what have been the economic and social benefits for residents of the community where those previously abandoned buildings are located? How many youths remained in school and graduated, but might have dropped out were it not for their participation in the program? How many of them moved onto jobs in the construction trades after graduation, or entered apprenticeship programs, or enrolled in college or took some other apparently positive step after high school graduation?

Time figures prominently in defining and measuring outcomes. It does so in two ways. First of all, programs need sufficient time to demonstrate whether they can produce the outcomes expected of them. While it is often difficult to know how much time is enough, the two to three years typically allowed by funding sources may not be adequate in many cases. For most kinds of social programs, committed as they are to the uncertain business of changing human behavior, either longer testing periods or more realistic expectations about near term outcomes would probably be desirable. At the least, assessing outcomes in light of the time a program has had to operate should reveal important information about its suitability for replication.

Time is also important in gauging the full effects of programs. In many cases, the hope is that a time-limited intervention will have lasting effects on those involved in it. Indeed, occasionally at least, short-term positive outcomes for participants provide a logical theoretical basis for believing that the intervention might result in more enduring gains. This is true, for example, of dropout-prevention programs, since high school graduates have a greater likelihood of becoming

employed than do dropouts. In other cases, however, short-term outcomes do not serve as reliable indicators that the program will produce long-term benefits.[6] Still, data on such outcomes may be the only way available to evaluate whether the program is achieving its aims.

To ensure that data are as revealing as possible, it will be helpful to consider:

- ✦ the precise definition of positive outcomes for the program's participants;
- ✦ the length of time, after the program ends, during which former participants are tracked to see how they're doing (if they are or can be tracked at all);
- ✦ the source(s) and apparent validity of the data; and
- ✦ comparisons of the program's outcomes with outcomes of similar programs with similar aims.

It will be especially important to relate available data on outcomes, limited though their quality may be, to whatever is learned from a review of implementation. Outcomes, in themselves, represent only a partial picture. Simply knowing that a program is associated with particular results says little about how and why those results occurred. A program may appear to produce great outcomes, but until these can be

---

[6] One well documented example of the gap between short-term positive outcomes and long-term impacts is the Summer Training and Education Program (STEP) research demonstration, initiated by Public/Private Ventures in 1984. The 14- and 15-year-old participants in this two-summer remediation, work and life skills intervention showed significant short-term gains in reading, math and life skills. However, once these youth left STEP and returned to their regular school and life routines, they experienced the same school dropout, teenage pregnancy and employment rates as youth who did not participate in the intervention. See Gary Walker and Frances Vilella-Velez. 1992. **Anatomy of a Demonstration**. Philadelphia: Public/Private Ventures.

satisfactorily explained in terms of actual program operations, any conclusions reached about "cause and effect" should be treated with skepticism. A careful analysis of the possible causal connections between implementation and outcomes will provide insight into which aspects of the program and its operating environment seem most important in producing the results obtained. It will also at least hint at whether factors unrelated to the program as designed (e.g., unrecorded or seemingly irrelevant participant characteristics) may be contributing to outcomes. While it will usually be quite difficult (even in random assignment approaches) to establish such detailed causal connections with a high degree of scientific precision, evaluation should, at a minimum, generate a logical explanation for how the program does or doesn't get results. Absent a convincing explanation, the program is not likely to be a compelling candidate for replication.

All in all, of course, the reality is that many programs will have collected only minimal amounts of valid information on outcomes in any event. All programs are likely to have anecdotal evidence of their effects, and they are likely also to have data on negative and positive "exits" from the program. But because of the constraints of time, resources, and know-how, they are less likely to be collecting data on participants' experiences after the program. As a result, developing better instruments for measuring outcomes may become a key task in the early stages of replication, when attention must be paid to establishing common ways to gauge effectiveness.

BUILDING IN MONITORING AND EVALUATION

Planned replication is typically a time- and resource-intensive process, involving repeated rounds of training, on-site assistance, and overall support for the emerging program network. While usually less costly on a per-site basis than the design and implementation of a new program model, replication still requires, in most cases, a significant financial investment. The more funders can know about how a program works and why, the more assurance they can have in their decision to invest in its replication.

Credible evidence about a program's transferability and replication-worthiness is not acquired at one final point; rather, it accumulates over time. Thus, when funders support the development and implementation of new programs, it makes sense to anticipate the possibility of future replication. When done right, this includes examining the replicability of program design as well as building in ongoing assessments to measure the program's worthiness.

At times, of course, the crafting of design and assessment tools occurs almost as a byproduct of program development. One example of this is the community schools initiative operated by the Children's Aid Society and its partners in a low-income area of New York City. Their community school

concept aims to transform public schools into full-service community centers that are open all day, year-round to everyone in the community, including children, teens, parents and other adults. Along with a nontraditional academic curriculum, the schools have on-site health clinics, child care, extended-day programs, adult education, parent workshops and summer camps.

The Children's Aid Society began operating one community school (a middle school) in 1992 and a second (an elementary school) in 1993, and evidence has begun to indicate that the approach is effective, at least in the short term. In preliminary evaluations, the two schools have had the highest attendance rates in their district, improved reading and math scores, and no serious incidence of violence. There is also anecdotal evidence that the schools have helped build a sense of community in the neighborhood. While the long-term impacts of the schools will be examined in a ten-year longitudinal study that is expected to begin in 1995, the initiative has already attracted significant attention outside of New York City, and the Children's Aid Society has received funds to help other schools and community-based organizations replicate the approach.

Since the community schools are part of a citywide school system, and since all students in the city schools take standardized tests, student evaluation has been included in the Children's Aid Society program. The positive short-term outcomes on these tests, along with the political attractiveness of a comprehensive, fully integrated system of child and family services, have made the community schools an appealing candidate for replication.

Sometimes, the potential for replication is part of the thinking of program designers from the very beginning. This

was the case with the Summer Training and Education Program (STEP), developed by Public/Private Ventures in the mid-1980s and now operating in more than a hundred sites around the country. STEP was created under a fairly unusual set of circumstances and with a significantly larger investment of money than is typical, but the process, in its essential form, would appear to be applicable to other program design efforts.

The STEP model was developed to be compatible with the U.S. Department of Labor's Summer Youth Employment Program, which annually has placed hundreds of thousands of young people in publicly subsidized jobs across the country. Through STEP, participants' work experience would be combined with intensive remedial instruction focused on reading and math, along with life skills. STEP was designed with clear aims and explicit program components based on previous research. The immediate thrust was to stop summer learning loss among 14- and 15-year-old at-risk youth and generate increases in conventional test scores in reading and math. In addition, the program aimed to improve youths' knowledge and behavior regarding pregnancy prevention, good health practices, and other life skills. The model's components, designed to work within the Summer Youth Employment Program, focused directly on achieving these aims. They included: 1) a fixed minimum number of hours for each key program element; 2) detailed remediation and life skills curricula; 3) the requirement of a two-summer program, with the second year structured to parallel the first; and 4) a relatively low-key "school-year support" component to ensure that there would be ongoing contact with STEP youth between their first and second summers.

STEP also had a second, knowledge-building aim: to learn whether those expected increases in performance and knowledge, and changes in behavior, in fact, occurred and endured over time. Evaluation was an integral part of STEP from the beginning. The research included a random assignment impact study to assess the short-term and long-term impacts of the program on participants; and an implementation analysis that used data from application forms, questionnaires, and program records, as well as case studies and observations to assess the feasibility of implementing the model in various settings and on a large scale.

Since STEP was a summer program, it was able to produce pre- and post-program data quickly. The early data provided relatively strong evidence of short-term effectiveness. Those results, along with the transferable nature of the design - particularly the structured allocation of program hours and the scripted curriculum - and its tailored fit to an established federal funding stream were key factors in the replication of STEP. When the opportunity came to foster the adoption of the program in locations beyond the demonstration sites, those factors, along with the managerial capacity Public/Private Ventures had developed to assist new sites, made it possible to act swiftly and effectively.

CONCLUSION

The STEP experience is an unusual but nonetheless instructive example: designing programs so that they are potentially replicable and so that their replicability and worthiness can be assessed in an ongoing way from the start is a wise move for funders who are making investments in new programs and strategies. Doing so can yield good evidence of programs' operational strengths and weaknesses and make for a smooth, efficient transition into replication for those programs that pass evaluation muster.

Yet, it must also be recognized that not all outstanding programs can or will emerge in this deliberate fashion. Allowance has to be made for "diamonds in the rough" - for high performing programs designed and carried out by local people for local purposes but which show promise of being workable in other communities. Though such programs usually lack the detailed, systematic evidence that is attainable when the potential for replication is contemplated from the outset, they can and should still be evaluated. Taking the time to scrutinize their accomplishments and operations is the only way to determine whether the "promise" they represent is real and transferable.

Regardless of a program's origins - whether it stems from a high-profile demonstration effort or emerges at the local

level - funders need to be reasonably confident that investing in its replication will, indeed, reproduce the program's original success in new settings. Evaluation builds the factual case for determining whether such confidence is warranted or not. While, in the final analysis, deciding whether to fund a replication is a judgment call, the better informed that judgment is, the more likely the final decision will prove to be the right one. And more right decisions of this kind should lead, if the logic holds, to greater support for social programs that really do work.

* * *

For funders who already invest in program evaluation, we hope this brief presentation of methods and issues has suggested fruitful ways in which questions about replication can be incorporated more specifically in the evaluation process. Of course, much remains to be learned about the conditions needed for replication to succeed. Nevertheless, there is good reason, as the preceding material tries to make clear, to give "replication potential" a more fulsome meaning as an evaluation criterion than has generally been the case.

For funders whose commitment to evaluation remains tentative, our hope is that the prospect of seeing programs they fund get adopted more broadly will entice them to at least experiment with the evaluation approach outlined here. While every social program grant a funder makes need not be subject to rigorous evaluation, surely those presumed to hold the most promise merit some level of critical attention in which issues of worthiness and replicability receive systematic review.

Finally, for anyone committed to the project of developing social programs that work, we hope this discussion has helped

to make replication more approachable and interesting as an emerging innovation in social investment strategy. The need for innovation today is not confined to the search for better social programs. Innovation is also needed in the ways by which society's limited resources for social investment can be leveraged for greater impact. Planned replication of proven initiatives represents a critical step in this latter direction. If readers of this paper have come away with a more pragmatic understanding of how to decide when planned replication is warranted, our purpose will have been served.

Kathryn Furano is assistant director of the West Philadelphia Improvement Corps. Linda Z. Jucovy is an independent writer and researcher. David P. Racine is the president of Replication and Program Strategies, Inc. Thomas J. Smith is vice president and director of special projects with Public/Private Ventures.

ADDITIONAL READING

Charles Stewart Mott Foundation. 1990. ***Replication: Sowing Seeds of Hope.*** Flint, Michigan: Charles Stewart Mott Foundation.

Conservation Company and Public/Private Ventures. 1993. ***Building From Strength: Replication as a Strategy for Expanding Social Programs that Work.*** Philadelphia: Replication and Program Services, Inc.

DeLone, Richard H. 1990. ***Replication: A Strategy to Improve the Delivery of Education and Job Training Programs.*** Philadelphia: Public/Private Ventures.

Public/Private Ventures. 1994. **"Replicating Exemplary Programs: Lessons from Experience."** Philadelphia: Public/Private Ventures.