RESEARCH REPORT

Assessing the Impact of Community-Level Initiatives

A Literature Review

February 2021

OPRE Report #2021-4

Assessing the Impact of Community-Level Initiatives

February 2021

OPRE Report #2021-4

William Congdon, Margaret Simms, and Carol De Vita

SUBMITTED TO

Kimberly Clum, project officer
Office of Planning, Research, and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Contract Number: HHSP233201500064I

SUBMITTED BY

Teresa Derrick-Mills, principal investigator
Urban Institute
500 L'Enfant Plaza
Washington, DC 20024

DISCLAIMER

The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research, and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.

This report and other reports sponsored by the Office of Planning, Research, and Evaluation are available at www.acf.hhs.gov/opre.

Contents

Acknowledgments

This report was funded by the Office of Planning, Research, and Evaluation (OPRE) in the US Department of Health and Human Services' Administration for Children and Families. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

We would especially like to thank our OPRE project officer, Kimberly Clum, for providing guidance and feedback throughout the project.

We would also like to thank Urban Institute staff who helped improve the quality of this report: Teresa Derrick-Mills, Ashley Hong, Kate Thomas, and our editor, Liza Hagerman. We provide a special note of thanks to Monica Rohacek, who originally conceptualized the report and did early work to identify the programs and evaluations that might be featured here.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

Photo from Shutterstock/Rawpixel.com.

Executive Summary

Addressing many of the most difficult social problems requires an approach that combines the efforts of multiple organizations, often operating in different sectors, coordinating across an entire community. Numerous programs overseen by the Administration for Children and Families (ACF) engage in these community-level initiatives, or community-level change efforts. These initiatives, however, are challenging to evaluate. To inform ACF on how to better capture these initiatives' impact, the Office of Planning, Research, and Evaluation (OPRE) at ACF commissioned this literature review to see what can be learned from recent efforts that have not only attempted to achieve community-level change, but have also attempted to produce credible evidence about the impact of these efforts. To draw out these lessons, this review begins by providing an overview of the methodological challenges community-level initiatives pose for assessing causal impacts. It then highlights some key lessons gleaned from recent evaluations of community-level change efforts.

Summary of Key Findings and Considerations

- Credibly assessing the causal impact of initiatives on outcomes of interest is a long-standing challenge in the evaluation literature. The most credible way to determine causality is through random-assignment studies. However, such studies are often infeasible or impractical for evaluating community-level initiatives.
- Impact evaluations for community-level initiatives involve special challenges. Evaluation designs must account for these challenges, including the difficulty of defining the "treatment," the importance of spillover effects within communities, the interactions between different initiative programs and services, and the difficulties of detecting effects in community-level outcomes.
- New methods are strengthening our ability to measure the effects of community-level initiatives. Recent advances in quasi-experimental methods suitable for community-level evaluation are promising. In addition, theory-based or mechanism evaluations—i.e., evaluations focused on uncovering the casual links between an initiative's activities and its outcomes—can be especially helpful for learning about an initiative's impacts. Moreover, mixing methods to include qualitative and quantitative approaches can strengthen causal claims.

- Recent evaluations of community-level initiatives such as the Community Development Block Grant, Promise Neighborhoods, and the Urban Health initiative, among others, offer particularly useful lessons for how to approach community-level evaluations.

Introduction, Background, and Methods

Assessing Effectiveness, Not Just Assessing Compliance

Over time expectations have grown about what it means to adequately assess government programs. While assessing whether funds were spent as intended and whether organizations are complying with all the rules that accompany their funding continues to be critically important, assessing the extent to which government programs are effective and building the tools and evidence to enable that assessment has become increasingly important as well. Since the 1990s, various legislative and executive mandates have emphasized the importance of monitoring the effectiveness of policies and programs and developing the systems necessary to make that possible. The more recent Government Performance and Results Act Modernization Act of 2010 (GPRAMA) and the Foundations for Evidence-Based Policy Making Act of 2018 have further emphasized the importance of basing policy and program decisions on solid evidence, using reliable data and information, and systematically incorporating research and evaluation into program and agency processes.

The federal government accomplishes important elements of its work through grantees (state and local governments, territory and tribal governments, and community-based organizations), and federal expectations for how grantees document and report on their work are also evolving. The federal government has been seeking more information from its grantees about the outcomes they achieve and the impact of the work. At the same time, grantees have been looking to support their own continuous quality improvement and to inform their strategic allocations of future resources (e.g. what strategies or initiatives are effective and which ones are not?) through the acquisition of actionable information.

These changing expectations and the different types of information they require mean that tracking the numbers of people served and the types of services they are receiving are no longer sufficient. In some types of programs, grantees are carrying out specific activities on behalf of the federal government. In those cases, the federal government directly specifies the outcomes the grantees need to achieve and how the grantee must track and measure the information. In many other federal programs, the grantees have substantial flexibility in the local programs they design to meet the needs of people in their communities or their communities as a whole. In these cases, the grantees need to determine the best ways to track and measure outcomes and impacts relating to the particular objectives and activities of their programs.

The increasing expectation that the federal government and its grantees measure effectiveness of programs has coincided with advances in the methods for studying program outcomes and impacts. Most of these advances, however, focus on studying impacts on individuals. In other words, they help us understand which programs are most effective in helping improve individual and family or household-level outcomes. A number of important programs, however, are focused on making changes at the community level, such as improving local economic conditions, or achieving public health goals. As discussed below, assessing the effectiveness of community change efforts can pose particular challenges, and sometimes require different or specialized methods for studying program outcomes and impacts.

Assessing Effectiveness of Community Change Efforts

Over the years, the federal government's expectations of grantees' reporting have grown, not only regarding their uses of federal funds, but also whether they are achieving the objectives that underlay the use of those funds. To demonstrate the effectiveness of their activities, grantees must first be able to identify expected outcomes, develop indicators that can effectively measure progress in achieving those outcomes, and collect data on those indicators. Grantees must then develop and implement a strategy for using those data and measures to assess the impact of their efforts on those outcomes. These are especially challenging activities, however, for grantees engaged in community-change efforts. Many—if not most—community-change efforts are multiyear endeavors, and although a grantee may be clear about the efforts' ultimate objectives, the grantee must also find ways to capture annual progress toward achieving those objectives, identifying the appropriate, incremental outcomes—and indicators of those outcomes—to do so. In addition, grantees often conduct community-change efforts as part of a collaboration with other organizations and must find ways to account for the impact of their own contribution to the overall effort.

Historically, many organizations that have been involved in community-level work and undertaken community-level activities have focused more on measuring the number of clients served by various initiatives rather than the impact of those services. In moving to measure the impact of those initiatives, community-level initiatives face a particular challenge: determining not only how the services affected the individual or family, but also the community as a whole. The purpose of this literature review is to identify lessons that ACF and its grantees may be able to apply to evaluations of community-change work from both the literature on measuring community-level change and evaluating specific community-level initiatives. In reviewing this literature, we capture lessons on various approaches to

defining the boundaries of a community as well as on how to define and measure desired outcomes. We also highlight the expertise that may be required to conduct evaluations of community-change efforts, what data collection methods and strategies are likely to yield useful information, and how these data might be aggregated or compiled to provide a useful view at the state and federal levels.

The discussion below proceeds as follows: Section 2 discusses some of the particular challenges associated with impact evaluation in the context of community-level initiatives. Section 3 provides examples of methods and findings from selected evaluations of other community-level initiatives. Section 4 concludes by drawing on lessons from both the highlighted examples as well as the broader evaluation literature to suggest key considerations for evaluating community impacts. Appendix A contains an overview of different evaluation methodologies used to measure the impacts, or effectiveness, of an initiative. Although these methods are not specifically designed to evaluate community-level initiatives, some of them show promise for evaluating community-level initiatives. Appendix B contains a matrix of the literature reviewed by key components for reference. Appendix C contains case summaries of eight evaluations of community-level initiatives.

Special Challenges in Evaluating Community-Level Initiatives

Community-level initiatives have at least two defining characteristics that distinguish them from other types of programs: first, by definition, they target communities or neighborhoods, as opposed to other units such as individuals or families. Second, they include not only particular programs that target the community level, such as community-based forms of housing assistance, but also, broader initiatives that include multiple programs and services that might vary across communities or over time. A succinct, functional definition that captures both of these elements is given by Prudence Brown, a long-time scholar of and consultant for place-based community change efforts: "geographically targeted initiatives that operate across systems" (1996, 162).

The distinctive features of community-level initiatives present complexities for evaluating effectiveness, making impact evaluation particularly challenging. Evaluating effectiveness, in this case, means measuring the extent to which the initiative causes or leads to changes in outcomes for community residents or institutions that would not have occurred without the program. The special evaluation challenges begin at the level of defining the initiative and outcomes. They also include conceptual and technical considerations related to measurement and detection.[1]

Definitional Challenges: Treatments and Outcomes

Evaluating the impact of a program or initiative requires clearly defining both the program or initiative of interest and the outcomes of interest. The first set of challenges for evaluating community-level initiatives stems from this requirement. For community-level initiatives, both the initiative itself and its outcomes may not be easy to define. For a start, community-level initiatives are often complex, involving a suite of programs and services. Further, the suite of programs and services may change over time and, if the initiative is being conducted in multiple localities, each locality may adapt the particular suite of programs and services to fit the specific needs and context of their community. In addition, community-level initiatives allow for a multiplicity of potential outcome measures and even levels at which outcomes might be specified.

[1] A detailed discussion of methodologies used and the accompanying challenges can be found in appendix A.

For the purposes of evaluation, the issue of defining the initiative under study raises challenges both for **conducting** evaluations of effectiveness **and** for **interpreting** their findings (Fiester 2011). Typically, community organizations tailor initiatives to the needs of their particular community, and those initiatives evolve over time as needs and programs change. The key evaluation question in this context becomes: what precisely is the thing we are trying to understand the impact of? Or, in evaluation terms, what is the treatment or intervention?

Different ways of answering this question lead to different tradeoffs: one is to define the specific initiative to be evaluated as the overall community effort in this area, without attempting to assess each component or the contribution of each partner, but instead evaluating whether outcomes are improved under the initiative as a whole compared with its absence (setting aside the practical challenges with doing so). This answers the question: how did the particular mix of programs and services delivered as part of the evaluated initiative affect outcomes in the studied community or communities over the evaluation period? This approach comes with limitations, however.

Knowing what difference a particular, complex (multifaceted) initiative actually made, at best, leads to only suggestive understandings of what initiative elements led to changes in which outcomes, and it carries some risk that practitioners and researchers may, in the absence of more detailed evidence, draw the wrong conclusions. First, when looking at an initiative in the aggregate, rather than as a clearly defined set of particular elements, it is hard to know whether it was the combination of elements that made the difference or whether only particular elements were driving the results. Therefore, we cannot know whether it is necessary to replicate the initiative in its entirety to achieve similar results in a different (but similar) community or if we only need to implement some of the program elements (and which ones).

Second, without a well-defined set of programming that is stable over time and consistently implemented, it can be difficult to know how well the results might translate to other populations, places, or time periods. This is what evaluators refer to as the "external validity" of the evaluation (Deaton and Cartwright 2016). Put another way, it is hard to know if the outcomes of the initiative were a result of factors unique to its implementation (in a particular time and place and with particular groups of people) or if the initiative would achieve those outcomes in other places. When an initiative is understood only as it is broadly defined, and implementing communities may in practice emphasize different elements or adapt the mixes of programs and services over time, external validity challenges are compounded. To assume that a similar initiative might lead to similar effects in different contexts requires, at the very least, a detailed understanding of how the program elements were implemented

and adapted for their original context and the ability to replicate or adapt implementation in new contexts.

Another way to manage the issue of "what makes a difference" is to focus on specific elements of an overall initiative. For example, an organization could decide to evaluate particular programs or services within the initiative that are relatively stable or well defined. This can allow an organization to try to understand or uncover the mechanisms that link these initiative activities to community impacts. It would also be possible to evaluate the effectiveness of programs or services that evolve over time using highly technical evaluation methods, but the use of these technical methods requires keeping track of detailed information on how the interventions change over time (Robins and Hernan 2009).[2] These approaches may generate more generalizable evidence about the effectiveness of individual elements of the community-level initiative but may not help an organization understand the initiative's overall impacts. An evaluation of, for example, a specific program that promotes early childhood education gives more traction on the question of whether that particular program might produce similar outcomes for comparable families in similar communities, but it only tells you about the effectiveness of that piece of the initiative, not the overall initiative. Especially where the programs and services that are part of the initiative are thought to interact in important ways—for example, an early childhood education program that is complemented by a nutrition assistance program—understanding the effect of either program in isolation may be less informative than understanding how the programs' interactions mattered or understanding the effects of the initiative as a whole.

A second set of issues that require clarity for the purposes of evaluating community-level initiatives relate to defining outcomes. Even when initiatives operate at the community level, outcomes can be measured at either the individual or community level (or both). For example, employment programs can be evaluated to assess how effectively they promote economic mobility for participants, but also for how effectively they improve overall economic conditions in the community they serve. The choice is in part a matter of what interests policymakers and program officials and how the goals of a particular program or initiative are defined. The key point from an evaluation perspective is that, because of some conceptual and measurement issues elaborated on below, the outcome level selected can affect what evaluation approach is required or how the results can be interpreted.

[2] More information on these technical methods—such as a difference-in-difference approach using synthetic control methods—can be found in Appendix A.

Conceptual Challenges: Program Spillovers and Interactions

A key set of conceptual challenges for evaluating community-level interventions arises because theory and evidence suggest community-level initiatives not only affect the people they directly target, but also that their effects are not necessarily singular and straightforward. The complex relationships among the inputs, outputs, and outcomes of community-level interventions are more aptly described as involving *nonlinearities* (Nichols 2013).[3] That is, community-level initiatives might lead to positive impacts on community-level outcomes in ways that are more than the sum of their individual-level impacts, because of spillovers across people within communities. And the outcomes may also be more than the sum of their parts, because of interactions between the programs and services delivered as part of the initiative.

Spillover Effect

First, community-level interventions might generate spillovers across individuals within a community. When community programs are effective at ameliorating the causes or consequences of poverty for the people they serve directly, community members who are not served might also benefit and demonstrate improvements. For example, programs that improve participants' health might also protect the health of others in the community. That is, a program that provides participants with, say, free flu shots is likely to protect and improve health outcomes not only for people receiving the service, but also for their neighbors and other community members not served by the program, by making it less likely they will be exposed to the influenza virus.

For the purposes of evaluating community-level initiatives, this can pose a challenge by making it difficult to assess the full social impact of community-level initiatives with evaluations that compare outcomes across individuals or households (Manski 1993). If people directly served by a program are compared with those in the same community who are not served, the comparison will fail to generate valid impact estimates because even those not directly served by the program may be affected by the program indirectly. If, on the other hand, people served by the program are compared with those in a community without the program, the overall impact will understate the true program benefits by failing to capture spillovers. Where there is reason to believe that this type of spillover is substantial, it

[3] For a more detailed definition and discussion of evaluation techniques and measurement issues, please see appendix A.

becomes important to either evaluate using community-level outcomes or to design the evaluation to measure this type of effect directly (Crépon et al. 2013).

Interaction Effects

A second set of potential nonlinearities is driven by the fact that community-level initiatives are often predicated on theories of change which suggest that there are potentially important interactions between discrete programs or services in generating individual- or community-level impacts. That is, the delivery of a particular set of programs in combination—for example, job training, community support, and work incentives—might be more effective than the sum of their parts (e.g., Bloom et al. 2005). If there are important interactions between programs and services in a community-level initiative, narrow tests of specific programs within the initiative may not capture overall initiative impacts and it may be important to evaluate an initiative as a total package. Alternatively, evaluation approaches that incorporate models of how program elements might logically interact with each other, or are designed specifically to measure some of these interaction effects, might help build evidence on program effectiveness, even in the presence of such interactions.

Technical Challenges: Detection, Measurement, and Identification

A final set of challenges to understanding the impacts of community-level initiatives is related to more practical concerns having to do with the ability of an evaluation to detect effects and with issues related to measuring outcomes.

Beginning with the issue of detection, in impact evaluation, like any scientific measurement, it is easier to detect effects when they are large compared with everything else happening in the background. A fundamental challenge for evaluating the community-level impacts of initiatives is that they often represent very small investments compared with the size of the local economies and communities whose outcomes they seek to improve. As a result, even if such initiatives have true, positive effects on outcomes such as poverty rates or economic mobility, those effects may be too small to detect with available methods or to detect at a reasonable cost, given the other sources of variation in those outcomes (Bloom 2006). For example, if local poverty rates tend to vary year-to-year simply because of normal fluctuations in economic conditions, it might be hard to distinguish an initiative's

effects in seeking to reduce poverty from changes in the poverty rate that are simply a result of those fluctuations.

Where feasible, a common way to improve an evaluation's ability to detect small effects (otherwise known as the "power" of an evaluation) is to increase the total sample of observations among both those who are receiving the intervention and the comparison group. Small differences in outcomes between, say, one or a few individuals who participate in or are exposed to an initiative and one or a few others who do not may very well be due to chance. We, therefore, cannot distinguish between any effects of the initiative and the results of chance alone. But as the number of people in each group grows, the likelihood of differences due to chance grows smaller, and even small impacts of the initiative can be detected.

However, increasing the sample in this way is often impractical, expensive, or limited as an option in the context of community-level evaluations because the unit of analysis is typically whole cities or communities, and the number of cities or communities that can be included as sites in an evaluation is typically small. One solution to this challenge is to take advantage of variation in program access or delivery for individuals and analyze individual-level outcomes. In other words, examining the changes experienced by the hundreds or thousands of individuals within a city who have differing exposure or access to a program, rather than the changes experienced across 10 to 20 communities, might increase that analysis's ability to detect smaller but important changes occurring. Even for programs or services that vary principally across communities rather than individuals, using individual-level data rather than community-level data may sometimes help improve the evaluation's power to detect changes, although this improvement will tend to be more modest.

Another way to improve the chances of detecting even small effects is to reduce the amount of expected variation in the outcome for the analysis sample in the absence of the intervention. If, for example, test scores for first graders naturally exhibit a wider range of variation than test scores for, say, third graders, it should generally be easier to detect the effects of an educational intervention for elementary school students on third-grade scores than on first-grade scores. By reducing potential sources for noise (or variability in what you are studying), it can be easier to detect a signal. As a consequence, another way to increase an evaluation's ability to confidently distinguish a program's effects from other factors that might contribute to differences between the treatment and comparison groups is to construct the sample to reduce the potential role of those other factors. For example, communities in different states might be subject to different types of economic shocks or policy environments that affect outcomes. In this case conducting an evaluation that compares treatment and control communities within a single state could reduce the likelihood that different economic

conditions cause variation in outcomes. An approach like this would improve the evaluation's ability to detect program impacts. In addition, this is where the choice of evaluation methods can matter for community-level evaluation; for example, methods such as synthetic control case studies and matching estimates—which compare outcomes among treated communities or individuals with comparison communities or individuals that are otherwise similar based on their observable characteristics—focus on reducing variation in the analysis sample as part of their approach.[4]

A second set of issues has to do with the unit of analysis: impacts are ultimately assessed using data that attempt to measure concepts corresponding to outcomes of interest. In the context of community-level evaluation, the choice of the unit of analysis can be significant. Of special concern for those assessing the impact of community-level initiatives is that the selection or migration of individuals, families, and households in or out of treatment communities can potentially affect results. For example, some adults receiving workforce development services might leave the communities in which they receive the service to find or accept employment. As a result, community-level initiatives might change community outcomes both through effects on people within communities and by changing the composition of communities.

If, for example, one effect of a community-level initiative is that people who receive services become more mobile and succeed at attaining improved economic outcomes by moving away from the treated community, this might depress outcomes for the original community, even as it improves outcomes for those people and potentially society as a whole. The direction of this effect is ambiguous; for example, federal housing investments may have improved neighborhood outcomes in part at the expense of displaced residents (Popkin et al. 2004). Both the program's direct effects on people served and any compositional effects—that is, effects because of changes in who lives in the community as a result of a program or initiative—may interest policymakers, but understanding their respective contributions to collective outcomes requires careful measurement.[5]

A final set of issues relates to measurement. A potentially important issue with community-level initiatives is the question of how long it may take—even in theory—for detectable effects on community-level measures to emerge. For example, some of the most substantial impacts from one

[4] A more detailed discussion and definition of alternative evaluation techniques, including synthetic control case studies and matching estimation methods, is included in appendix A.

[5] For example, identifying both direct and compositional effects might require the ability to measure both the outcomes for and location of treated (and possibly comparison) individuals over time. This would include the ability to measure outcomes of those who move away from the original treatment site to other communities by tracking individuals over time or arranging to identify them in administrative data covering larger geographies, such as state- or national-level records.

housing mobility demonstration experiment appeared only after many years, when adult outcomes could be observed for individuals who moved as children (Chetty, Hendren, and Katz 2015). At a very practical level, this concern speaks in part to the value of administrative data—such as operational data collected by the program under study or data collected for administrative purposes by other programs, such as tax or wage records—in program evaluation, which can sometimes make tracking outcomes over longer periods of time more feasible.

The matrix in appendix B further highlights some of these key findings from selected studies in the impact evaluation literature, along with their potential implications for other efforts to evaluate community-change.

Evaluations of Community-Level Initiatives: Selected Cases and Findings

There are numerous evaluations of community-level initiatives, though they differ in their evaluation approaches. It is instructive to examine a sample of them to see how principles and strategies discussed above can be implemented in different program contexts.

Appendix C contains eight case studies of community-level initiatives: the Community Development Block Grant Program; Making Connections; Northside Achievement Zone; Plain Talk; Promise Neighborhoods; Strive Together; Vibrant Communities; and the Urban Health Initiative. They represent a mix of programs funded either by the federal government or private foundations. In most cases, the initiatives have moved toward quantifying their impact, but a few focus more on implementation than impact. These have been included because they raise important issues about how to shape the project and data collection to generate the information needed for impact analysis or because they raise important issues about the program structure needed for successful program operation.

While the details of each evaluation are summarized in appendix B, we examine some commonalities and differences in the approaches taken by the different programs.

Defining the Problem/Identifying What to Change

Most of the programs examined defined the proposed changes in narrow ways—for example, reducing teen pregnancy, increasing educational performance, and improving health outcomes for children. But others had broader goals or multiple strategies, including altering systems and policies to shape life prospects (Vibrant Communities) or reducing disparities in well-being among neighborhoods (Community Development Block Grant, or CDBG). The strategies for achieving the objectives were also divergent. Some proposed very broad strategies such as neighborhood improvement and parent's economic success (Making Connections), while others were more narrowly focused on a set of programs around a single strategy, such as improving academic achievement (Northside Achievement Zone).

Measuring Change

The indicators chosen to measure change were either one or two (educational outcomes in the Northside Achievement Zone) or many (employment rates, household income, crime rates, etc. in Making Connections). Programs with a long-term goal, such as Strive Together, firmly grounded their change measures in a theory-of-change model. For them, educational outcomes at each stage determined adult success. Others focused their measures on a few or several outcomes that were a few steps away from the ultimate goal or objective, taking a mechanism or theory-based approach in their evaluation. For example, Plain Talk had the long-term goal of reducing teen pregnancy, which was facilitated by providing trusted advisors (such as parents) with more information that they could share with their children about reproduction, birth control, and available services. This was based on the assumption that more information from a trusted source would lead to changes in teens' behavior. In this context, outcome measures were focused on whether adults' knowledge increased and whether teenagers were more likely to seek information from those adults, rather than on the pregnancy rate.

Strategies for Measuring Change

The diversity in expected outcomes also generated variation in how the initiatives chose to collect and summarize data on their progress. Most relied on a combination of quantitative and qualitative measures. The qualitative data were collected through surveys, interviews and site visits/observations. Quantitative data from the sites (e.g., administrative and performance measurement data) were compared in most cases with publicly available data on comparison groups. The educationally focused initiatives (Strive Together, Northside Achievement Zone) used data from other communities, school districts, or state averages as comparisons for assessing progress over time. The CDBG measures were interesting in several ways and might be particularly relevant for other ACF-funded community-change level efforts because the evaluation discussed setting a minimum level of funding necessary to expect impact and because the evaluators developed a comparison group by averaging outcomes in other similar communities.

Key Considerations for Evaluating Community Impacts

Some key practical considerations to consider in implementing data collection and evaluation of community-level initiatives include the following:

Consider the Tradeoffs Associated with Evaluating Particular Community-level Initiatives and Building Knowledge that Can Apply More Broadly

It is possible to learn whether particular initiatives are having impacts on community-level outcomes of interest. However, it may be challenging to do so in a way that generates generalizable knowledge about how and why initiatives work (or do not), whether they would work in other communities or under other conditions, or what elements of an initiative should be scaled up or back to improve outcomes. The need to account for spillovers across individuals and programs, together with the variation in initiatives and the fact that they may evolve over time, means that impact evaluations of the overall initiative may have limited external validity. In other words, while evaluations of an overall initiative can provide rigorous evidence as to whether it is working as implemented in its current location, this may still provide only limited information as to whether a similar initiative would generate similar outcomes in other communities or with other populations. Narrower tests of the impact of specific components of the initiative—such as particular programs or services associated with it—might generate more generalizable knowledge; for example, about the impacts of those components or important mechanisms that link those components to particular outcomes. It is important to remember, though, that these narrower tests pose their own tradeoff which is that they may give an incomplete picture of the overall impact of an initiative that combines multiple programs.

Consider Building Capacity for the Development of Synthetic Case Control Methods[6]

While in principle randomized evaluations—which compare outcomes between groups of individuals or communities randomly assigned to either receive an intervention or not—can be used to evaluate community-level programs (Prinz et al. 2016), in practice the conditions that would make it possible to conduct a randomized evaluation may not be appropriate for community-level interventions or could only be made use of with considerable expense. Quasi-experimental approaches—which compare outcomes between individuals or communities receiving an intervention with otherwise similar individuals or communities that do not—may bring the right balance of feasibility and validity for evaluations of community-level outcomes. The method with perhaps the greatest promise, and a growing body of examples, may be the synthetic control approach, which compares outcomes for a treatment site with a weighted average of comparison sites.[7] It typically uses information on the nature and timing of the initiatives, existing data sources for variables predictive of the designated outcomes, historical data on these outcomes across treatment and comparison sites, and a pool of potential comparison sites where the initiative has not been implemented. In principle, this approach could be used to conduct impact evaluations of community-level initiatives. If valid site-specific estimates of community-level program impacts can be generated, it may be possible to aggregate results across sites to arrive at estimates of the overall program impacts. Doing so would require that sites have access to data that could be used to measure the same outcomes at each site. Evaluators would also have to address variations in how initiatives might be implemented across sites or over time (either by selecting particular treatment elements, defining the treatment very generally, employing estimation methods that can account for treatments that change over time, or some combination) and may need to account for differences in population characteristics across sites (Nichols 2013).

Leverage Theory and Existing Evidence in Generating New Evidence

Theory-based evaluation approaches or mechanism experiments may potentially play an important role in generating evidence about community-level initiatives' effectiveness. These evaluation approaches depend, however, on an initiative having a clearly articulated theory of change. Evaluations using these

[6] See appendix A for complete descriptions of the various evaluation techniques.

[7] A more detailed discussion and definition of synthetic control methods is included in appendix A.

approaches might focus on detecting the role of the initiative in influencing the mechanisms by which the programs or services offered by the initiative are believed or understood to lead to the desired outcomes. For example, a community-level initiative to improve health outcomes by increasing access to fruits and vegetables might evaluate the effects of the program on the prices for fruits and vegetables in local stores, relying on other sources of evidence or knowledge that connect prices to consumption and, ultimately, population health outcomes. Theory-based evaluations combine findings from evaluations that focus on particular links in a causal chain with other sources of information—including theoretical models or bodies of empirical evidence—to support broader conclusions about programs' overall effectiveness. These evaluation approaches work best when leveraging existing evidence by building on an area that already has a developed evidence base or when targeting important gaps in the evidence on particular program elements or links in a causal chain. These evaluation approaches are less effective for newer areas where little evidence already exists. In addition, developing effective mechanisms or theory-based evaluations requires that program officials and evaluators possess or develop knowledge on existing evidence and theory before designing and conducting such evaluations.

Build a Body of Evidence Related to Outcomes That Includes Impact Evaluation

While only some empirical methods are adequate for drawing valid conclusions about causal impact, a wide range of methods can contribute to building a body of evidence that strengthens, in aggregate, overall understanding of which programs work, how, and for whom. Qualitative methods and participatory research can inform and complement quantitative impact evaluation methods. Mixing methods—that is, using both qualitative and quantitative approaches as part of the same study—can enhance the overall strength of the evidence on program impacts.

Take Advantage of Complementarities between Impact and Performance Measurement

Performance measures and the data collection efforts that support them can be designed with an eye toward their potential use for impact evaluations. Similarly, the results of impact evaluations should be used to inform the development of performance measures. In particular, program and other administrative data have the potential to be a powerful and cost-effective tool for both performance

measurement and program evaluation, especially when it can be joined to other datasets that include outcome measures.

Appendix A. Evaluation Methods for Measuring Community-Level Initiatives' Effectiveness

Effectiveness, Impact Evaluation, and Causal Inference

Program evaluation, at the community level or otherwise, can take many forms—from understanding processes and implementation to assessing program impacts on outcomes of interest. To understand if and how programs are externally effective—that is, whether they succeed in achieving their intended goals—the key form of evaluation is impact evaluation. Impact evaluation seeks to answer the question: did the program cause a change in the outcomes that policymakers, program officials, or society care about? In the context of community-level strategies, this might be an assessment of whether a program leads to improved outcomes for individual participants (e.g., better earnings and employment outcomes), improved outcomes at the community level (e.g., lower local poverty rates), or both.

The central challenge for impact evaluation is establishing causation, or what is sometimes referred to as the attribution problem (Chigas, Church, and Corlazzoli 2014). What, precisely, does it mean to say that a program caused a change in an outcome, isolated from all the other factors that might contribute to changes in an outcome? The most common approach in the social sciences is to define causality by way of reference to a counterfactual: if in the presence of a program a community sees, for example, higher rates of economic mobility than it would have in the absence of the program, the program is said to have caused the improvement (Holland 1986). Put this way, the challenge for impact evaluation emerges clearly: once a program is in place, we can only observe the world as it is now; we can no longer see what would have been the case without the program—the counterfactual.

Impact Evaluation and Performance Outcomes Measurement

While performance measurement, the monitoring of program performance using established measures, is distinct from impact evaluation in important ways, it shares some common elements and goals,

supporting the general objective of designing, managing, improving, and investing in programs and initiatives in an evidence-based way (Hatry 2013).

The key difference between the two is that although both use data to assess program performance, performance measurement is not designed to make causal statements about program impacts. That is, even where performance measurement is focused on program outcomes—for example, measuring employment outcomes of program participants—it is concerned principally with measuring those outcomes, or comparing them against established goals or targets. It does not compare those outcomes to a counterfactual, as does impact evaluation.

The two concepts can be complementary in important ways (Davies 1999). On the one hand, performance measurement can potentially support impact evaluation. For example, the data collected as part of performance measurement systems can sometimes be used for the purposes of impact evaluations. Performance measurement can also help inform decisions about when to conduct an impact evaluation as well as the interpretation of findings from conducting an impact evaluation. For example, if performance measures show very few participants complete a workforce training program, this can both suggest that such a program is unlikely to generate impacts and help evaluators understand why. Likewise, impact evaluations can strengthen the design of performance measurement systems by generating evidence about which program measures are among the most important to track and manage, and they can help set meaningful performance goals.

Randomized Evaluations

One solution to the challenge of establishing causality is identifying, or generating, a group that looks as much as possible like the treatment group would have in the absence of the program—a comparison group—and comparing their outcomes with those of the treatment group. The purest way to generate this comparison group is through random assignment, whereby people are randomly assigned to receive the treatment or "business as usual" (also referred to as the "control" condition). This is the logic that leads to randomized experiments as a common form of impact evaluation: when people are randomly assigned to receive a program (or not), the group that does not receive the program can be expected to look like the counterfactual. In other words, their status should be that of those treated if they had not received the service. As a result, differences in outcomes between those individuals or communities randomly assigned to receive a program or not represent the causal impact of the program (Rubin 1974).

This high degree of internal validity—the confidence with which their results can be interpreted as causal—is a central attraction of randomized impact evaluations. A drawback to random assignment evaluations is that they can sometimes be impractical, infeasible, or expensive. Conducting a randomized evaluation depends, at least, on the ability to withhold the program from some group for at least some period; adequate data; and a sufficient number of units to randomize to detect effects of a meaningful size (Duflo, Glennerster, Kremer 2008).

Quasi-Experimental Methods

Because random assignment is feasible and worthwhile under only a particular set of circumstances, other approaches are often necessary or preferable. One set of alternative methods, which seeks to approximate some features of random assignment studies, are quasi-experimental methods (Athey and Imbens 2017). These include difference-in-difference estimation, matching methods, and regression discontinuity designs. Like random assignment, these methods seek to identify impacts by comparing outcomes for groups that receive the program with a comparison group that does not; their common challenge is finding or constructing comparison groups with a defensible claim to approximating the counterfactual.

DIFFERENCE-IN-DIFFERENCE APPROACH

A difference-in-difference approach compares outcome trends in sites with a program before and after its implementation with trends in sites without the program (Card and Kreuger 1994). In other words, do clients in the program advance more than those in a similar site with similar characteristics? Difference-in-difference methods can relax somewhat the practical and methodological requirements of randomized studies. They can be employed in instances where program officials or evaluators are not able to assign people or sites to receive a program. That is, they can be used to assess impact by comparing, say, cities with and without a program in cases where the determination of which cities would receive the program, and when, was made based on factors unrelated to conducting an evaluation. They are suitable, also, when only aggregate data, such as city- or county-level poverty rates, rather than data on individuals, such as individual or family income, are available. However, they are more open to challenges regarding their internal validity. In particular, interpreting the results of a difference-in-difference approach as causal relies on the assumption that the outcome path for program sites would have paralleled that of the comparison sites in the absence of the program.

DIFFERENCE-IN-DIFFERENCE APPROACH: SYNTHETIC CONTROL METHODS

A recent variant of the difference-in-differences approach, synthetic control methods compares outcome trends in a program site with the trend for a weighted average of comparison sites, rather than for particular comparison sites (Abadie, Diamond, and Hainmueller 2010). That is, rather than comparing outcome trends in, say, one city with the program against trends in other similar cities without the program—as in a traditional difference-in-difference estimation—the synthetic control approach compares outcomes in the city with the program against outcomes for a "synthetic" comparison city that is statistically constructed using data from potentially many cities without the program.

When it is possible to construct a synthetic control in this way, doing so may offer some advantages over traditional difference-in-difference estimates. Employing a data-driven method for generating a comparison that should more closely approximate the counterfactual may improve claims to internal validity. In addition, this approach provides greater transparency and a more mechanical solution to the selection of comparison sites, by assigning their weights systematically, rather than relying on the evaluator's discretion. The main drawback of this approach, in addition to the challenges that come with more methodological complexity, is that it may not always be possible to construct a synthetic control that closely resembles the treatment site. In particular, generating the synthetic control requires having data on both the treatment and comparison sites over time periods—usually some number of years—before the program's implementation to determine the weights used to construct the synthetic control. And even with sufficient data, the method will only generate a synthetic control suitable to use for evaluating outcomes when some weighted average of available comparison sites does in fact closely resemble the program site.

MATCHING METHODS

Matching methods, like the synthetic control approach, seek to empirically identify which individuals or sites might serve as a good comparison group. Approaches such as propensity score matching compare outcomes among program recipients or sites with outcomes for individuals or sites who look similar on observable characteristics (Dehejia and Wahba 2002). This approach also requires good data on outcomes and observable characteristics of both the treatment group and untreated sites or clients that serve as a comparison group.

REGRESSION DISCONTINUITY DESIGNS

For programs where eligibility is determined based on meeting a threshold value of some criterion—for example, having an income level below some specified dollar amount—another way to create a

comparison group is by comparing individuals served by the program or initiative to those who just missed the cutoff. This approach, called regression discontinuity, exploits discontinuities in program availability or implementation to identify impacts (Imbens and Lemieux 2008). For example, if admission to an education or training program is available only to those with test scores above a certain threshold or those above a certain age, the program's effect can be estimated by comparing outcomes for those just above the cutoff with those just below it. The logic of this approach is that individuals who just miss the program cutoff—whose tests scores are only a few points too low or who are only a few months too young to participate—are likely to be otherwise similar to those who just make the cutoff.

This approach has important limitations: first, impact estimates may represent effects of the program only in the region of the discontinuity; individuals or sites just below a cutoff, for example, are likely to serve as a good counterfactual for those only a little above such a cutoff but not for those well above it. As a result, impact estimates generated using these methods might deviate from or only approximate the average treatment effect of the underlying program or service on the overall service population. Put another way, a program's effects for other groups it serves, or might potentially serve, and who have different needs or characteristics from individuals near the cutoff, might be different from those estimated using a regression continuity design. In addition, these approaches are only applicable, by definition, in special instances where programs are designed and implemented in a way that creates the necessary discontinuity, and they are best suited to instances where the discontinuity depends on a continuous variable, such as age or test scores. While the above example of a training program that conditions eligibility on a test score provides a sharp discontinuity that can be the basis for estimating treatment effects using this method, an otherwise similar training program that conditioned eligibility, for example, on current employment status would not be well suited for this approach.

Mechanism and Theory-Based Approaches

A different set of approaches, which in some instances can substitute for full program evaluations, includes those that specify and evaluate intermediate links in the causal chain that connect programs to outcomes. These approaches are variously referred to as mechanism experiments (Ludwig, Kling, and Mullainathan 2011), mediation analysis (Morgan-Lopez and Burr 2017), theory-based evaluation (Leeuw 2012), or contribution analysis (Mayne 2001). These methods, collectively, take the approach of generating evidence about program effectiveness by clearly specifying the theory—and, in particular, identifying the mechanisms (also referred to as mediators)—by which programs generate, or are believed to generate, changes in outcomes. Evaluating program impacts on those mechanisms, or the

role of those mechanisms in generating outcomes, then provides a way to generate evidence on program effectiveness without evaluating program effects on outcomes directly. The conceptual relationship between programs or policies (P), mechanisms (M), and outcomes (Y) in this framework is illustrated schematically in figure A.1 below.

FIGURE A.1
Policies, Mechanisms, and Outcomes

Figure adapted from Ludwig, King, and Mullainathan (2011).

For example, consider a hypothetical community-level initiative to improve community health outcomes by providing incentives to local grocers to stock fruits and vegetables at reduced prices. In this case, the program (P) is the incentive to grocers, the outcome (Y) is a measure of community-level heath, and the mechanism (M) linking them is that reduced fruit and vegetable prices will lead to greater consumption. Research already shows the link from M to Y: reducing the price of fruits and vegetables can increase consumption (Bartlett et al. 2014), and greater consumption is associated with improvements in population health (He et al. 2007). So, it may be sufficient for gauging whether the program is advancing its ultimate goals to evaluate the link from P to M: does the program lead to reductions in consumer prices?

Mechanism experiments can open up evaluation opportunities that inform program-impact assessment without the costs and requirements associated with a full program evaluation. But to be valuable for understanding program impact, they require clearly articulated theories and mechanisms, as well as evidence about other links in the chain between programs or services and outcomes of interest. In the example above, for instance, the already existing evidence that reduced fruit and vegetable prices led to increased consumption, and that increased consumption is associated with improvements in population health, is essential. It is only in the context of the evidence on those links in the causal chain that evidence of consumer price reductions because of the program can suggest that the program is achieving its ultimate objectives.

Other Approaches and Mixing Methods

Other approaches can also help contribute to understanding program impacts. Qualitative research, such as case studies, interviews, and ethnography, can inform and complement quantitative impact evaluation methods. Participatory research methods that emphasize the inclusion of voices and insights from within communities can also play an important role in building reliable evidence on program impacts (Cornwall and Jewkes 1995). Both approaches can potentially be used to generate hypotheses that existing theory may fail to reflect, provide evidence on the role of particular mechanisms in generating program outcomes, or otherwise contribute to a body of evidence that strengthens overall assessments of the causal role of programs in affecting outcomes.

Note, these varying approaches to generating information about program impacts are not mutually exclusive, and in fact they are often complementary. In general, mixing methods—including, for example, qualitative components of randomized program evaluations—can help enrich and strengthen causal claims (Schorr and Farrow 2011). Moreover, the most appropriate weights to place on different research methods can depend on the stage of initiative development and implementation and whether impact evaluation is being conducted in the context of developmental, formative, or summative evaluation (Preskill, Parkhurst, and Juster 2014).

Appendix B. Matrix of Key Methodological Articles

To highlight and expand upon particularly relevant sources from the impact evaluation literature, this appendix includes a matrix with additional information from key sources referenced in the memo, including full references, article abstracts, and important themes. The articles and chapters in the matrix are reviewed sources that represent either (1) overview discussions focused on issues, challenges, and methods for community-level evaluation, (2) up-to-date reviews of key segments of the evaluation literature (e.g., a recent review of quasi-experimental methods), or (3) recent or illustrative examples of particular methods of interest (e.g., an example of synthetic-control-methods application). For each article or chapter, we include notes on the methods discussed or employed as well as the way in which they potentially address challenges associated with community-level evaluation

TABLE B.1

Effectiveness Evaluation Literature Review Matrix

Citation	Abstract (condensed)	Method(s) of focus	Topics covered	Challenges and overcoming them
Conceptual issues and community-level evaluation				
Fiester, Leila. 2011. "Measuring Change While Changing Measures: Learning In, and from, the Evaluation of Making Connections." Baltimore, MD: The Annie E. Casey Foundation.	"As detailed in this report, Casey's evaluation efforts were challenged by shifts in Making Connections' design, management and emphasis over the course of the 10-year initiative. At the same time, these challenges produced a flexible, "learning-while-doing" approach to assessing site progress toward achieving better results for neighborhood children and families—a needed departure from more rigid and less meaningful methods of measuring community change initiatives. As the report concludes, Making Connections' evaluation broke new ground—especially in the sharp focus it gave to measuring local capacity to achieve and sustain results and in developing new,	• Place-based program impact evaluation in the presence of treatments that evolve over time	• Measuring population-level change • Framing a program evaluation with multiple purposes • Multiple data needs and uses at various scales	Community-level interventions often evolve over time, which poses a challenge to evaluation. To address challenges posed by evolving treatment programs, authors propose strategies such as identification of outcomes; data-collection strategies; development of local site-specific evaluative capacity; and clarifying evaluative frameworks.

Citation	Abstract (condensed)	Method(s) of focus	Topics covered	Challenges and overcoming them
	performance-based ways to manage and improve the initiative's implementation."			
Ludwig, Jens, Jeffrey R. Kling, and Sendhil Mullainathan. 2011. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives* 25 (3): 17–38.	"Randomized controlled trials are increasingly used to evaluate policies. How can we make these experiments as useful as possible for policy purposes? We argue greater use should be made of experiments that identify behavioral mechanisms that are central to clearly specified policy questions, what we call mechanism experiments. These types of experiments can be of great policy value even if the intervention that is tested (or its setting) does not correspond exactly to any realistic policy option."	• Use of mechanism experiments where a policy evaluation would be inappropriate, with examples and explanations • Mechanism experiments differ from policy evaluations in their focus (testing mechanisms rather than policy impacts) but not necessarily their methods; mechanism experiments can be randomized evaluations	• Applicability and strengths of mechanism experiments to inform policy, especially instead of policy evaluations	Mechanism experiments can address challenges associated with obtaining rigorous evidence on program impacts, where program evaluations may be infeasible by combining evaluations of program impacts on mechanisms with other sources of information such as behavioral models, parameters, and prior beliefs
Preskill, Hallie, Marcie Parkhurst, and Jennifer S. Juster. 2014. "Guide to Evaluating Collective Impact. Part 1: Learning and Evaluation in the Collective Impact Context." Washington, DC: FSG.	"This section describes the importance of continuous learning and presents an evaluation framework to guide the design of different performance measurement, evaluation, and learning activities. The purpose of the framework is to help readers conceptualize an effective approach to performance measurement and evaluation, given their initiative's stage of development and maturity."	• Continuous learning, performance measurement, and program evaluation in the context of collective impact • Evaluation of both individual and policy-level outcomes	• Developing a framework for performance measurement and evaluation of collective impact efforts • Strategies to plan for evaluation	Address challenges posed by having multiple actors as part of a collective impact initiative by using a shared measurement system.
Preskill, Hallie, Marcie Parkhurst, and Jennifer S. Juster. 2014. "Guide to Evaluating Collective Impact. Part 2: Assessing Progress	"This section offers guidance on how to plan for and implement a variety of performance measurement and evaluation activities aimed at assessing an initiative's progress, effectiveness, and impact. It includes sample performance indicators, evaluation questions, and outcomes for collective	• Developmental, formative, and summative evaluation matched to the collective impact initiative's stage of development	• Collective impact initiatives: sample outcomes, indicators, research questions	Address challenges associated with outcome measurement in the collective impact context by assessing impact on indicators, including measures of professional

Citation	Abstract (condensed)	Method(s) of focus	Topics covered	Challenges and overcoming them
and Impact." Washington, DC: FSG.	impact initiatives in different stages of development, as well as advice on how to gather, make sense of, and use data to inform strategic decision making, how to communicate evaluation findings, how to choose and work with evaluators (when desired), and how to budget for evaluation."	• Planning for and implementing a variety of performance measurement and evaluation activities, including sample performance indicators, evaluation questions, and outcomes	• Case studies: initiatives applying evaluation	practice, individual behavior, cultural norms, funding flows, and policy outcomes.
Preskill, Hallie, Srikanth Gopal, Katelyn Mack, and Joelle Cook. 2014. "Evaluating Complexity Propositions for Improving Practice." Washington, DC: FSG.	"While evaluation has traditionally focused on assessing programmatic impact according to pre-determined indicators, a new approach is needed for evaluating complex initiatives, as well as initiatives operating in complex environments where progress is not linear, predictable, or controllable. Nine propositions can help evaluators navigate the unique characteristics of complex systems, improve their evaluation practice, and better serve the needs of the social sector."	• Flexible evaluation plans and budgets • Systems mapping • Social network analysis • Interviews • Rapid feedback debriefs • Summaries/learning memos • Critical incident reviews • After-action reviews • Timeline of key event • Review of information related to context • Reflective practice • Design labs • Focus groups • Most significant change • Appreciative inquiry • In-depth case studies • Ripple effect mapping • Observations • Digital storytelling • Snapshot survey • Bellwether interviews	• Characteristics of complex systems • Nine propositions (overarching strategies) for evaluation of complex systems • Helpful tools/methods for each proposition	1. Design and implement evaluations to be adaptive, flexible, and iterative 2. Seek to understand and describe the whole system, including components and connections 3. Support the learning capacity of the system by strengthening feedback loops and improving access to information 4. Pay particular attention to context and be responsive to changes as they occur 5. Look for effective principles of practice in action, rather than assessing adherence to a predetermined set of activities 6. Identify points of energy and influence, as well as ways in which momentum and power flow within the system 7. Focus on the nature of relationships and interdependencies within

Citation	Abstract (condensed)	Method(s) of focus	Topics covered	Challenges and overcoming them
		• Web analytics • Contribution analysis • Causal diagrams • Surveys • Time series designs		the system 8. Explain the nonlinear and multidirectional relationships between the initiative and its intended and unintended outcomes 9. Watch for patterns, both one-off and repeating, at different levels of the system
Schorr, Lisbeth, B., and Frank Farrow. 2011. "Expanding the Evidence Universe: Doing Better by Knowing More." Washington, DC: Center for the Study of Social Policy.	"Thanks to the last two decades of research and experience, we now know so much more than ever before about what it takes to improve outcomes for disadvantaged children and families….But our expanded knowledge has not led to better outcomes at a magnitude that matches the need. Among the reasons: we have too often failed to marshal the full extent of available evidence and to generate new, real-time knowledge from experience. The paper explores how we might make use of all the evidence we now have from multiple sources, including research, theory, practice, as well as the findings from program evaluations, and how we might aggressively gather new evidence about the nuanced and powerful strategies for change that are currently emerging."	Use multiple evaluation methods that align with the multiple purposes of evaluation, the nature of the intervention, and the stages of implementation including the following: • Qualitative studies and process evaluations, documenting the way in which ideas, resources, and stakeholders are brought together to launch and implement an initiative; • Cohort studies, to look more intensively at the results for specific target groups of children or families within the broader population whose needs are addressed by a complex intervention; • Neighborhood surveys of impact, to begin to	• Expanding the definition of "credible" evidence • Systematically generating and using evidence to inform complex initiatives. • Attributes of alternative methods that can help determine success and understand the workings of complex social programs.	Proposed approach to evaluating complex efforts: • Begin with a results framework • Use strong theory to connect activities to results • Expect to compare results, but don't expect to find a perfect comparison group to "prove" causality • Use multiple evaluation methods that align with the multiple evaluation purposes, the nature of the intervention, and the implementation stages.

Citation	Abstract (condensed)	Method(s) of focus	Topics covered	Challenges and overcoming them
		get data on population-level results; and • Case studies of specific interventions, to understand in-depth the way in which they do or do not affect change.		
Technical approaches and community-level evaluation				
Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2012. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 104 (490): 493–505.	"Building on an idea in Abadie and Gardeazabal (2003), this article investigates the application of synthetic control methods to comparative case studies. We discuss the advantages of these methods and apply them to study the effects of Proposition 99, a large-scale tobacco control program that California implemented in 1988….Using new inferential methods proposed in this article, we demonstrate the significance of our estimates. Given that many policy interventions and events of interest in social sciences take place at an aggregate level (countries, regions, cities, etc.) and affect a small number of aggregate units, the potential applicability of synthetic control methods to comparative case studies is very large, especially in situations where traditional regression methods are not appropriate."	• Synthetic Control Methods (case study), which compare outcome trends in a program site with the trend for a weighted average of comparison sites, rather than for particular comparison sites	• Data-driven procedures to select synthetic comparison units in comparative case studies • Using a placebo study to produce quantitative inference in case studies	Synthetic control methods have an advantage for community-level interventions in their ability to generate impact estimates with only a single treatment site, using aggregate data. In addition, synthetic control models overcome typical challenges of comparative case or difference-in-difference studies, including the following: • Reduced ambiguity about how comparison units are chosen • Inferential techniques that better measure uncertainty of estimates
Athey, Susan, and Guido W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy	"In this paper, we discuss recent developments in econometrics that we view as important for empirical researchers working on policy evaluation questions….First, we discuss new research on	• Regression discontinuity designs • Synthetic control methods	• New methods for drawing causal inference from observational data	This paper surveys a range of quasi-experimental approaches to generating rigorous measures of program impacts. Because

Citation	Abstract (condensed)	Method(s) of focus	Topics covered	Challenges and overcoming them
Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32.	identification strategies in program evaluation, with particular focus on synthetic control methods, regression discontinuity, external validity, and the causal interpretation of regression methods. Second, we discuss various forms of supplementary analyses, including placebo analyses as well as sensitivity and robustness analyses, intended to make the identification strategies more credible. Third, we discuss some implications of recent advances in machine learning methods for causal effects…"	• Difference-in-difference study • Estimating average treatment effects in settings with multivalued treatments • Causal effects in networks and social interactions • External validity • Leveraging experiments • Placebo analyses • Robustness, identification and sensitivity • Machine-learning methods for average causal effects • Machine learning for heterogeneous causal effects	• Strategies for approaching supplementary analyses, including machine-learning models	randomized evaluations are often infeasible for the purposes of evaluating community-level interventions, these methods are potentially more promising. Authors propose solutions for overcoming barriers to discussing causality in the context of estimating the impact of policies using nonexperimental methods.
Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *The Quarterly Journal of Economics* 128 (2): 531–80.	"This paper reports the results from a randomized experiment designed to evaluate the direct and indirect (displacement) impacts of job placement assistance on the labor market outcomes of young, educated job seekers in France….After eight months, eligible, unemployed youths who were assigned to the program were significantly more likely to have found a stable job than those who were not. But these gains are transitory, and they appear to have come partly at the expense of eligible workers who did not benefit from the program."	Randomized experiment designed to measure spillover effects (case study)	• Experimental design and randomization • Estimating externalities: - Unconstrained reduced-form models - Pooled reduced-form model • Instrumental variable estimates of program impact	A particular challenge associated with evaluating community-level interventions is identifying program impacts in the presence of spillover effects across individuals, where even individuals who do not receive programs or services may be affected by the presence of a program in a community. This paper illustrates an empirical approach to addressing this challenge by identifying and

Citation	Abstract (condensed)	Method(s) of focus	Topics covered	Challenges and overcoming them
				measuring spillovers by varying the fraction treated across areas.
Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, Vol 3, edited by T. Schultz and John Strauss. Amsterdam and New York: North Holland.	"This paper is a practical guide (a toolkit) for researchers, students and practitioners wishing to introduce randomization as part of a research design in the field. It first covers the rationale for the use of randomization….Second, it discusses various ways in which randomization can be practically introduced in field settings. Third, it discusses design issues such as sample size requirements, stratification, level of randomization and data collection methods ….Finally, it discusses some of the issues involved in drawing general conclusions from randomized evaluations, including the necessary use of theory as a guide when designing evaluations and interpreting results."	Randomized evaluations as a research tool to test theories.	For randomized evaluations: • Justification and implementation • Experimental design considerations • Data analysis • Considerations for implementation	Random-assignment studies are often used as a benchmark for internal validity. This paper provides a practical guide to the issues associated with designing and implementing a randomized evaluation and proposes various strategies to overcome challenges of randomized evaluation in various situations: • Combine economic modeling with variation from randomized evaluations to evaluate a larger set of parameters • Design experiments to explicitly test theories of economic behavior
Nichols, Austin. 2013. "Evaluation of Community-Wide Interventions." Washington, DC: Urban Institute.	"Random-assignment experiments are the gold standard for assessing the impact of a policy or program for good reason, but they are not always a good option. For place-based initiatives, spillover effects and other factors make random assignment studies inappropriate. There are other methods, some of which use propensity score reweighting, that can credibly estimate impacts for these programs."	Experimental and quasi-experimental methods for community-level interventions	Approaches to challenges associated with place-based evaluation because of spillovers and dynamic adjustment of treatment	Challenges: spillover effects, heterogeneous and evolving treatments Suggested solutions: synthetic-control models; g-estimation; propensity score reweighting

Appendix C. Case Studies

1. The Impact of CDBG Spending on Urban Neighborhoods
    - » An evaluation of the US Department of Housing and Urban Development's Community Development Block Grant (2002)
2. Making Connections
    - » Annie E. Casey Foundation initiative spanning 10 years and 22 US communities (2011)
3. Northside Achievement Zone
    - » Neighborhood initiative in Minneapolis funded, in part, by Promise Neighborhoods (2010, 2015, 2017)
4. Plain Talk Initiative
    - » Annie E. Casey Foundation initiative spanning 5 years and 5 US neighborhoods (2001, 2006)
5. Promise Neighborhoods
    - » An evaluation of the US Department of Education's Promise Neighborhoods (2015)
6. Strive Together Initiative in Bexar County Texas
    - » Strive Together is a national nonprofit cradle-to-career network that coordinates resources across 70+ communities across the US (2017)
7. Urban Health Initiative
    - » Robert Wood Johnson Foundation initiative spanning 10 years in 5 US cities (2002)
8. Vibrant Communities
    - » Canadian foundation funded for 10 years in 13 Canadian cities (2010, 2012)

The Impact of CDBG Spending on Urban Neighborhoods

Purpose/focus of study
- Develop a methodology for evaluating the impact of CDBG funding and identify performance measures that align with program objectives and could be used to measure a range of important neighborhood quality variables
- Identify an appropriate comparison group

Measures of change
- Improvements in neighborhood quality
  - Social disadvantage (female headship, teen pregnancy, welfare dependence)
  - Housing type and quality
  - Crime levels

Data sources
- Administrative and Census data for neighborhood quality measures
- Median loan amount (available in HMDA files—Home Mortgage Disclosure Act data)
- Dun and Bradstreet counts of businesses in community

Sample size or number of programs
- Communities within 17 cities receiving substantial CDBG funding in the 1994–96 period (the inclusion criteria of substantial funding was based on CDBG spending per poor resident and defined as a threshold level of mean per tract expenditure per poor resident determined as part of the analysis)

Comparison group or benchmark
- Neighborhoods with similar characteristics but without substantial CDBG funding

Time period or length of study
- 1994–99

Project Overview

The Community Development Block Grant (CDBG) is often considered HUD's flagship urban improvement program. Established in 1974, the CDBG program allocates federal funds to states, cities, and urban counties according to a formula based on population, poverty levels, housing stock age, and other need factors. Unlike earlier federal urban redevelopment models that were tied to specific purposes, the CDBG program gives local government wide discretion over the use of the funds. HUD has very little influence over local choices pertaining to goals and strategies, except that the funds are intended to reduce disparities in well-being among neighborhoods. Many localities use CDBG funds for activities that aim to improve household income, employment, business activity, homeownership, and housing investment.

This evaluation was conducted by a team of researchers at the Urban Institute, led by Chris Walker, for the US Department of Housing and Urban Development (HUD). The study focused on neighborhoods that received CDBG funding between 1994 and 1996. The study's purpose was to identify valid, reliable, and generally accepted performance indicators that would measure how well CDBG investments correlated with program outcomes—that is, the impact CDBG funds had on improving urban neighborhoods, as measured by generally available data related to outcomes.

Evaluation Overview

The overall goal was to develop and test several performance measures that might form the basis for a future performance measurement system. The evaluation had four specific goals:

1. Identify a small number of readily available, generally accepted, and easily replicable indicators of neighborhood quality of life that are suitable for assessing CDBG impacts.
2. Develop a definition of "substantial" CDBG investments in a neighborhood.
3. Recommend benchmarks against which to assess the performance of neighborhoods that have received substantial levels of CDBG investments.
4. Compare the study's results with local informants' understanding of the impact of CDBG on their neighborhoods in the late 1990s.

Among cities that received substantial CDBG funding in 1994 and 1996, seventeen cities were selected for the study sample. Selection was based on multiple criteria, including

- the richness and availability of data that could be assembled at the neighborhood level;
- geographic representation, using the four census regions;
- metropolitan area job growth between 1994 and 1997, to ensure inclusion of communities with varied job-growth levels over that period. The data were derived from HUD's *State of the Cities 2000* report and classified into four categories (e.g., no-growth, low-growth, moderate-growth, and high-growth); and
- the 2000 CDBG allocation for each city, including some larger cities with variation in CDBG investment across census tracts within cities.

In addition to developing databases and performance measures, the research team also conducted interviews with local informants in four of the 17 cities sampled. The purpose of the interviews was to assess to what degree local perceptions corresponded to the researchers' categorizations of neighborhood performance.

Evaluation Data Sources and Analysis

The first step in the study was to identify data that were universally available, reliable, frequently collected, and generally accepted as valid measures of neighborhood quality. Finding potential indicators was difficult because many measures were uneven in quality and not consistently collected

across cities. The researchers used factor analysis to identify six factors that represent neighborhood quality of life. These factors were assessed to be stable across cities and time:

1. Social disadvantage (e.g., female headship rates, teen birth rates, welfare usage, etc.)
2. Housing type and tenure (e.g., percentage of single-family homes and homeownership rates)
3. Prestige (e.g., median home values and percentage of residents with college degrees)
4. Business and Employment (e.g., number of businesses and jobs)
5. Crime (e.g., property and violent crime rates)
6. Housing vacancy rates (e.g., residential vacancy rates.)

Next, the study needed to define a "substantial" CDBG investment. Defining "substantial" investment also proved challenging because the amount of funds needed to show significant improvement might vary for different neighborhoods and by the city's socioeconomic conditions. For example, a declining city or neighborhood might require larger amounts of CDBG expenditures to produce an observable effect on neighborhood quality than would a growing city or already-improving neighborhood. In other words, "substantial" must be operationalized contingent on the conditions of the area. HUD's Integrated Disbursement and Information System (IDIS) and other HUD administrative data were used to construct a database of CDBG neighborhood expenditures.

To evaluate the suitability of various potential candidates for performance measures, the researchers used a series of simple regression models using only two variables at a time—CDBG spending and the given performance measure. The independent variable, CDBG spending, was averaged over three years to smooth out year-to-year fluctuations by census tract and to observe a lag between CDBG investment and the outcome of interest (generally measured in 1999).

In general, the study found that more CDBG spending resulted in improvements to neighborhood quality in all 17 study cities. Two indicators (residential mortgage lending activity and business and employment activity) proved to be good proxies for some, but not all, dimensions of neighborhood quality. The data underlying these indicators (namely, median loan amount from Home Mortgage Disclosure Act data and the number of businesses from Dun and Bradstreet) are readily available to all CDBG grantees and relatively inexpensive to obtain.

Summary of Evaluation Findings

The researchers concluded that this general research design was a good first step in developing a mechanism to assess the relationship between CDBG funding and neighborhood outcomes, particularly given the substantial gaps in information about effects of the CDBG program. However, they caution that the study was not broad enough to prove conclusively that CDBG funding is positively correlated with specific measurable outcomes.

To strengthen this approach, the researchers offer the following suggestions:

- A future study should be based on a nationally representative sample of jurisdictions or include all CDBG grantees. This would allow stronger correlations between variables.
- The inclusion of other public investments (such as investments in infrastructure or crime or schools or job programs) and more years of investment would capture a broader and more realistic view of community development investments.
- Improvements in administrative recordkeeping would allow for better tracking of CDBG expenditures. At the time of this study, it was recommended that more could be done regarding data cleaning, updating data-collection protocols, and ensuring more complete geographic coverage.

The research suggests that a neighborhood classification system cannot measure every nuance of neighborhood health and vitality. Neighborhoods are complex structures, and most models are only able to capture a few dimensions of the whole.

Thinking about the design and use of a performance measurement system as a tool for communities interested in assessing their own community development performance could be an appropriate construct for these studies. Local administrators in communities included in this study expressed considerable interest in the research goals but were concerned that such an approach would be used to set federal standards that might be used to sanction "poor" performance based on a few statistical standards that do not capture all the nuance of their neighborhood. Nevertheless, local administrators were more receptive to setting benchmarks by which they could assess their own progress in improving their neighborhoods.

Source

Walker, Chris, Chris Hayes, George Galster, Patrick Boxall, and Jennifer Johnson. 2002. *The Impact of CDBG Spending on Urban Neighborhoods.* Final report submitted to US Department of Housing and Urban Development. Washington, DC: Urban Institute.

Making Connections

Purpose/focus of study
- Document the process of program development and modification over the initiative's lifespan
- Assess the impact of the program on families in the designated communities

Measures of change
- Increased family earnings
- Increased family assets
- Children healthy and ready to succeed in school
- Increased civic participation
- Strong informal supports and networks
- Access to quality services and supports

Data sources
- Surveys and interviews (telephone and in-person)
- Site visits in three waves

Sample size or number of programs
- 10 neighborhoods in cities around the country—Denver, Des Moines, Iowa, Hartford, Connecticut, Indianapolis, Louisville, Milwaukee, Oakland, California, Providence, San Antonio, and Seattle/White Center
- 700 to 850 families per site

Comparison group or benchmark
- Benchmark or baseline data on conditions for children and families and overall conditions in the same neighborhoods

Time period or length of study
- 8 years (2002–10)

Project Overview

Making Connections was one of the first large-scale efforts to implement and assess change at the community level. Sponsored by the Annie E. Casey Foundation, the initiative, which began in 1999, spanned ten years and cost roughly $60 million. The Foundation initially selected 22 communities, with 10 of the 22 sites eventually funded for full implementation. The initiative aimed to improve outcomes for vulnerable children by transforming their neighborhoods and helping their parents achieve economic stability. The theory of change was that these goals would be achieved by connecting children and families with a range of services and supports and creating stronger social networks. Making Connections also sought to empower neighborhood residents and build the neighborhood's leadership capacity. Over the course of the initiative, changes were made to the initiative's methods and approach, data needs, and evaluation tools to adjust to ever-changing circumstances.

Evaluation Overview

The evaluation design evolved over the course of the initiative; however, it rested on a set of core assumptions, namely: families matter, place matters, connections matter, resident engagement matters, data matter, and results matter.

Five high-level research questions guided the work:

1. **What changes have occurred in the Making Connections neighborhoods?** For example, did employment rates in the neighborhood change? What about household income, access to affordable housing, and crime or incarceration rates?
2. **What changes have occurred for the families?** Were there changes in earnings, income, debt reduction, homeownership, school readiness, and children's health?
3. **What changes have occurred in community capacity or systems of support or opportunity?** For example, did connections to support services or networks to secure employment or access health care improve? Did use of data promote accountability?
4. **What strategies did the Making Connections sites pursue?** What were the contributions of these interventions toward achieving change?
5. **To what extent is it possible to observe increased capacities or improved trends and/or positive outcomes in Making Connections sites?** Is there evidence of public will to continue to advocate for change? Is there evidence of strong alliances or sufficient funding to support ongoing activities?

The Foundation evaluation team initially developed nearly 300 target outcomes and indicators that might demonstrate success. These were in the areas of family strengthening, alliance building, advocacy, and collective action; connections to informal social networks, formal helping systems and economic opportunities; building neighborhood assets, family functioning, and child and family well-being. Making Connection sites used these target outcomes to select (or develop) their own outcome indicators that reflected their local goals and aspirations.

The report identifies three challenges to evaluating community change initiatives.

1. how to measure population-level change for an initiative that seeks community-level results and operates on a small-scale programmatic level;
2. how to frame an evaluation that has multiple and sometimes competing purposes, such as measuring outcomes, building local capacity, and empowering neighborhood residents;

3. how multiple data needs and uses for different purposes (such as management and implementation issues) can drive evaluation options and choices.

Evaluation Data Sources

The evaluation of Making Connections required multiple data sources to address multiple needs and administrative functions. The evaluation's key elements included a cross-site survey, theories of change and impact, definitions and measurements of key concepts, and core capacities assessment tools to assess and manage progress in sites. The evaluation design was not experimental, and no effort to create or identify a control group was attempted. Instead, data collected throughout the evaluation was compared with baseline data, collected on child, family, and neighborhood conditions at the start of the full-implementation period of Making Connections. Baseline data were collected via a field survey conducted from 2002 to 2004 in each of the ten Making Connections neighborhoods and in each county that contained the Making Connections neighborhood.

CROSS-SITE SURVEY

The survey was intended to measure change within and across sites. It was administered three times throughout the initiative. In wave I, a representative sample of 700 to 800 households in the 10 Making Connections sites was drawn. The survey maintained a longitudinal sample over the three waves of data collection. In subsequent waves, the sample increased to 800–850 households to track children and families that moved out of the neighborhoods and add new families living in the neighborhood. The family was the unit of analysis, and specific questions were asked about a randomly selected "focal child" in each household. The baseline survey was conducted by telephone; subsequent surveys were conducted in-person by local interviewers recruited by the local team and national survey firm. Surveys took approximately 45 minutes and covered standard demographic characteristics; household composition and living conditions; children's health, education, and well-being; family economic hardships, employment, income, and assets; informal connections to the neighborhood and beyond; civic responsibility and activism; financial and human services used; civic and commercial amenities in the neighborhood, etc. In addition to the standard questions, each site could add up to 15 site-specific questions developed by the local team.

THEORIES OF CHANGE AND IMPACT

To better understand how to effect change, each site developed theory-of-change and theory-of-impact models. These models were specific to the particular issue(s) the site was working on and therefore

varied from site to site. Theory-of-change models were intended to help sites visualize the steps needed to create change; theory of impact helped sites understand how to get to scale (such as what capacity or alliances had to be developed, a strategy to expand the work, etc.). The purpose for creating these models was to help build capacity for local empowerment, data collection, and conducting evaluations that might help sustain and expand the efforts in the future.

DEFINITIONS AND MEASUREMENT OF KEY CONCEPTS

The national and local teams worked together to establish common definitions of key concepts that were integral to the initiative. For example, what is a family; what is meant by going to scale; what is accountability, success, or sustainability? Common definitions enabled all sites to measure concepts in fairly uniform ways or at least understand where critical concepts may differ from site-to-site. Where appropriate, this approach also enabled some cross-site comparisons.

CORE CAPACITIES ASSESSMENT TOOLS

Because Making Connections placed strong emphasis on building local capacity, it was essential for the evaluation team to devise ways to measure capacity and collect and use data that reflected the adequacy of interventions and areas of further improvement. This work identified a continuum of five developmental stages that communities pass through as they strengthen their collective capacity. At the lowest end of the scale is "maintaining business as usual;" at the upper end is "effective practices taking hold; transformation of business as usual." Evaluators also developed a 15-point rating system to assign numerical values to each site's capacity assessment. Each developmental stage has indicators defining success, which is assigned three values. A lower value indicates that the capacity-building work is in an early phase of planning or implementation; a medium value indicates that implementation is underway and the strategy is beginning to show progress; and a high value indicates that the community is ready to move to the next level or stage.

Summary of Evaluation Findings

The evaluation found that in some sites and for some populations Making Connections achieved many of its goals. However, it was not a panacea for reducing poverty and building capacity in all neighborhoods or for all population groups. The conclusion of some Foundation officials and program operators was that the initiative provided a strong platform for local problem-solving by establishing a set of tools, capacities, networks and norms. Evaluators caution that it is not advisable to make cross-site comparisons because all communities are different.

One key takeaway point from this complex initiative was that multiyear, multisite evaluations require flexibility and good communication between all participants. For example, management and evaluation teams worked together for many years and never fully resolved their differences regarding data collection and reporting. Site teams wanted data that were current and actionable; evaluation teams wanted longer-term, comparative data. As a result, the initiative required a wide range of tools to serve different needs and perspectives, and most importantly, strong communications between the initiative managers, site teams, and evaluation staff.

Other key findings include the following:

- To be successful, the effort needs a range of core community capacities in place, including vision, resident engagement, ability to use data, and community and financial support.
- Longitudinal survey data offer insights into the initiative's key issues but are not adequate for day-to-day management needs.
- A performance tracking system that includes outcome measures helps improve management of a complex, multifaceted change initiative.
- Large-scale community change cannot be supported and sustained without the buy-in of local partners and leaders who are in charge of the transformation.

Source

Fiester, Leila. 2011. *Measuring Change While Changing Measures: Learning in, and from, the Evaluation of Making Connections*. Baltimore: Annie E. Casey Foundation.

Northside Achievement Zone Outcomes

Purpose/focus of study
- Conduct a return on investment (ROI) study of program
- Measure improvement in academic performance and quality of life

Measures of change
- Return on investment study: that benefits to individual and society exceeded program costs
- Academic performance measures for enrolled students:
  - Kindergarten proficiency
  - Math proficiency
  - Reading proficiency
  - High school graduation

Data sources
- Administrative data for children in the neighborhood
- Household surveys
- Peer-reviewed social and economic literature for outcomes at different levels of education (for ROI)

Sample size or number of programs
- 1,700 families and 3,400 children as of 2013

Comparison group or benchmark
- For ROI
  - Projected outcomes for typical three-year in same community not enrolled at least five years in the program
- For student outcomes
  - Scores for students living in the same areas but not enrolled in program
- For quality of life
  - Benchmark was quality of life at program entry

Time period or length of study
- Five-year grant
- Findings in report are for the first three years

Project Overview

The Northside Achievement Zone (NAZ) began in 2008, when a group of nonprofits and schools in the Northside neighborhood of Minneapolis, one of the most impoverished parts of the city, came together to collectively provide a continuum of supports to parents and children in the neighborhood. Their goals were to improve academic performance, build a culture of achievement, and end generational poverty. In 2012, NAZ received a five-year federal Promise Neighborhood grant. By the end of the grant period, NAZ had partnerships with over 40 organizations and served nearly 1,700 families and 3,400 children.

To improve academic achievement, NAZ promotes integrated early childhood learning and development, provides both in-school and out-of-school support for K–12 students, and offers career and college planning support. Recognizing that families and neighborhoods play a vital role in children's success, NAZ helps families maintain or build a stable housing environment, as well as career and financial stability. Parents learn skills that are intended to help them assist with their children's educational success, and where appropriate, to further their own educational achievements.

Wilder Research of St. Paul, Minnesota, conducts yearly evaluations of NAZ to assess the program's accomplishments as a Promise Neighborhood for the children, families, and community. In 2017, Wilder issued a five-year assessment of the NAZ Promised Neighborhood, which is the basis of this summary. However, the project is ongoing and evaluation updates continue to be done, which can be found at http://northsideachievement.org/how-were-doing/results/results-naz/.

Evaluation Overview

The evaluation work has had three components—community surveys, year-end reports on student progress and quality-of-life outcomes, and a prospective return on investment assessment—that provided an integrative assessment of the progress and outcomes of the initiative. The first community survey was taken in 2010, with a follow-up survey conducted in 2013. Additional surveys are planned for every two to three years. The first two surveys had three goals: (1) to collect in-depth data about the well-being of children in the Zone, (2) to obtain parents' perception of the Zone, and (3) to monitor progress on the key NAZ outcomes.

Corroborating the survey information, end-of-year progress reports assess the academic achievements of children participating in NAZ. Age-appropriate measures are selected and correspond to NAZ activities. For example, the study tracks kindergarten readiness, reading proficiency (grades 3 through 5), math proficiency (grades 6 through 8), and on-time high school graduation rates. The research team also produces outcome measures that show the impact of NAZ supports in terms of assisting families and improving quality of life. Some of the measures tracked are family engagement in NAZ programs, housing support, and achieving career and finance goals.

Finally, the research team prepared a model to assess the potential costs and benefits of NAZ for a "typical" 3-year-old child who lives in the zone and participates in NAZ.

Evaluation Data Sources

Each component of the evaluation used different data sources.

The community survey was designed and constructed by the research team and NAZ staff. Data collection was conducted by the Urban Research and Outreach-Engagement Center (UROC) at the University of Minnesota. The surveys were conducted in-person by random door-to-door canvassing of the neighborhood. To be eligible, there had to be a child age 18 or younger living in the home. Only adults were interviewed. Respondents were asked about their impressions of the neighborhood and their schools, participation in various programs and activities, access to health care and transportation, and educational aspirations for their children. The response rates for the surveys were 47 percent in

2010 and 69 percent in 2013. While the participants in the program activities were not part of a randomized controlled sample, the methodology for selecting survey respondents (random households in a low-income neighborhood, door-to-door interviews) produced a generally representative sample of households in the zone. Given the high response rates, the findings can be interpreted with confidence, but also some caution.

The year-end reports, prepared by the research team, typically use standardized student test-score data and compare the test scores of NAZ-enrolled students (defined as enrolled for at least one year in NAZ) with the test scores of zone-wide (i.e., non-NAZ) students. The standardized student test-score data include measures of kindergarten readiness (as measured by the Beginning of Kindergarten—BKA—test administered by Minneapolis Public Schools) and reading and math proficiency measures from the Minnesota Comprehensive Assessment (MCA) tests. High school graduation rates are computed by the state as the share of the 9th grade cohort that graduates within four years. The Wilder research team broke out some of the results by comparing those of students who were attending a school that had fully implemented the NAZ service model or not and the type of NAZ support service the family received. Results are presented in aggregate form, not by individual child.

Quality of life outcomes rely on administrative data collected by NAZ staff. The data track the number of families and children that engage with NAZ through its various programs and activities and what types of help were received. The outcome measure describes the proportion who complete their goals or attain other outcomes related to the activities.

Prospective return on investment is a theoretical model estimating the costs and benefits that might accrue to society over the lifetime of a participating child. The model assumes that the child will participate in NAZ for at least five years, graduate from high school, and enroll in college. It also assumes various behavioral outcomes (such as less likely to smoke, commit crimes, use drugs, etc.) based on the findings of existing, peer-reviewed social and economic literature. It then goes on to calculate the economic value to society for improved behaviors. Because NAZ has been in existence for less than ten years, this is a theoretical model. However, the authors hope to update the model with empirical data as the program moves forward. The potential economic benefits are calculated based on four solution strategies: (1) early childhood education; (2) expanded learning, including achievement planning and mentoring; (3) parent education and engagement; and (4) career and financial support. Data for the model are derived from existing socioeconomic literature that calculates benefits for various behaviors and outcomes (for example, how much is saved from reduced drug use or reliance on public assistance). The authors state that they have tried to overestimate costs and underestimate benefits in constructing the model. Also, they have not considered factors related to movement in and

out of the neighborhood, which might suppress the actual earnings increase associated with higher education levels. They believe, however, the findings are conservative cost-benefit estimates for potential outcomes because the model and prior literature do not take into account the synergy that results from providing a multiple constellation of service support.

Summary of Evaluation Findings

COMMUNITY SURVEY

The surveys found that between 2010 and 2013, the visibility of NAZ increased. Residents had more knowledge about NAZ and generally positive impressions about the program. There had been a small increase in participation in early child care programs and activities, and a significant increase in participation in after-school programs. There was little change in the overall proportion of parents participating in parent-teacher conferences, but NAZ parents were more likely to do so than non-NAZ parents. Respondents had mixed feelings about the safety of their neighborhood: some thought it was safe; others expressed concerns. Surprisingly, most respondents were satisfied with their schools, even though statewide measures suggest these schools are low performing. The respondents' educational aspirations for their children were mixed: two-thirds said they wanted their child to go to college; one-third expressed lower goals. These results are being used by NAZ to assess their program activities and general program strategies, and to adjust as needed.

YEAR-END REPORTS ON STUDENT PROGRESS AND VARIOUS QUALITY-OF-LIFE OUTCOMES

In general, the evaluation reports indicated NAZ-enrolled students typically performed better than Zone-wide (non-NAZ enrolled) students; however, the extent of these differences varied from year-to-year and by measure. For example, measures of kindergarten readiness showed strong and positive effects of the NAZ program, while measures of math proficiency for older children showed only small and statistically insignificant effects. NAZ staff and academic partners are discussing possible causes for these variations and potential strategies to strengthen the NAZ experience.

Outcomes for families receiving program support were somewhat more difficult to measure, but results tended to be favorable. For example, roughly 40 percent of the 257 families who had a housing-related visit with a NAZ staff member had at least one documented episode of housing stabilization; 20 percent of the 59 adults with an active career goal started a new job. Not only do these data provide NAZ with feedback for making future modifications to the program, but they also provide evidence of effectiveness to potential funders who might invest in the program.

PROSPECTIVE RETURN ON INVESTMENTS

The model results estimate that the social return on investment in NAZ will be $6.12 for every dollar invested, with a net benefit to society of $167,467. The return on taxpayer investment is $2.74 for every dollar invested. Society's gains (estimated at roughly $200,000 for the average NAZ participant) come mostly from increased net earnings, based on increased educational attainment, career counseling, and increased productivity. There are also benefits from improved health outcomes, increased tax revenues, and other public savings such as lower crime rates, reduced need for special education, and less spent on public assistance and child welfare cases. The model estimates that total societal gains from NAZ would be more than $16.7 million in net benefits for every 100 participants served.

Sources

Diaz, Jose Y., Sarah Gehrig, Ellen Shelton, and Cael Warren. 2015. *Prospective Return on Investment of the Northside Achievement Zone.* St. Paul, MN: Wilder Research.

Idzelis Rothe, Monica, Ellen Shelton, Greg Owen. 2014. *Northside Achievement Zone: 2013 Community Survey Results: A Follow-Up to the 2010 Baseline Survey.* St. Paul, MN: Wilder Research.

Shelton, Ellen, Cael Warren, and Sarah Gehrig. 2017. *NAZ 2016 Annual Report, Including Accomplishments Over Five Years as a Promise Neighborhood.* St. Paul, MN: Wilder Research.

Note

Ongoing information about the project may be found at: http://northsideachievement.org/.

Plain Talk Initiative

Purpose/focus of study
- Document program development and assess implementation in communities
- Determine whether the program was successful in reducing teen pregnancy by developing a system of trusted adults

Measures of change

Immediate/intermediate objectives
- Changes in teen attitudes and behaviors around sex
- Increased adult-teen communication
- Increased knowledge of and access to reproductive health services

Long-term objective
- Reduction in teen pregnancy and childbirth

Data sources
- Interviews and surveys with teens and parents in the community and program participants
- Administrative data on the community

Sample size or number of programs
- Neighborhoods with low incomes in five cities—Atlanta, Hartford, New Orleans, San Diego, and Seattle
- For some aspects of the study, only three sites were included
- Total sample of teen/pre-teen girls was 698 in 1998 and 574 in 1994

Comparison group or benchmark
- Girls ages 12 to 18 in 1994 versus girls of the same age in 1998—in the same location

Time period or length of study
- 1994–98 for the teen outcomes study

Project Overview

From 1993 to 1998, the Annie E. Casey Foundation (AECF) launched and implemented Plain Talk, a program with the primary goal of lowering the teen pregnancy rate in five low-income urban neighborhoods. Plain Talk had three core strategies and components: (1) develop community consensus about the need to protect sexually active youth by encouraging early and consistent use of contraceptives; (2) give adults the information and skills they need to talk with adolescents about responsible sexual behavior and to share accurate information about pregnancy and STDs; and (3) work with local health agencies and service providers to ensure that adolescents have access to high-quality, age-appropriate reproductive health services, including contraception (Fiester 2006). AECF invited six sites to undergo a one-year planning process. After that first year of planning, five sites were invited to engage in full implementation of their planned program. The five sites were Mechanicsville in Atlanta, Georgia; Logan Heights in San Diego, California; White Center in Seattle, Washington; St. Thomas in New Orleans, Louisiana; and Stowe Village in Hartford, Connecticut. The original funding provided for 3 years of programming; the program was extended by another year to allow for four total years of program implementation. The sites were selected based on their high rates of teen pregnancy, high rates of poverty, and the existence of an implementing agency with a history of working well with

neighborhood residents. For implementation, each site had a lead agency, a project coordinator, and a community resident board to drive the work onsite. AECF wanted diverse sites (racially, socioeconomically, religiously, etc.) to examine the efficacy of the program across various populations. Further, sites had wide discretion in designing the intervention to fit with their own local needs and context. As a result, although the sites worked toward the same overarching goals, their approaches differed. For example, the San Diego site had a large Mexican-immigrant population (44 percent in 1993), and 44 percent of families lived below the poverty level, while the Indianapolis site was composed of African American and Caucasian families with a poverty rate of 27 percent (Fiester 2006, 10).

Evaluation Overview

AECF engaged Public/Private Ventures, a nonprofit and nonpartisan social policy organization, to evaluate Plain Talk's implementation and impact.[8]

1. Were sites able to create a community consensus about the need to protect sexually active youth?
2. How effective were the communication strategies in educating many adults and preparing them to talk to youth about responsible sexual behavior and where to access resources?
3. Did the sites improve access to high-quality, age-appropriate reproductive health services?
4. Did Plain Talk have an effect on adolescents' attitudes and behaviors (such as the communication behaviors and the reproductive health care behaviors that were key to the initiative's theory of change) and teen pregnancy rates?

The outcomes study drew on data and findings from the process study and, more specifically, addressed the following three research questions (Grossman et al. 2001, 8):

1. What were the characteristics of the youth interviewed (i.e., in the neighborhoods served by Plain Talk)?
2. How did communication patterns and youth's awareness of where to find birth control change over the evaluation period?

[8] During the first phase, the Casey Foundation provided financial support and oversight. After that phase ended in 1997, Public/Private Ventures conducted a study that documented program implementation and outcomes.

3. What were the associations between youth knowledge of (and attitudes toward) birth control and their use of contraception, incidence of pregnancy, and likelihood of being tested and treated for an STD?

Evaluators combined within-site and cross-site analysis, implementing multiple data collection strategies, such as surveys, ethnography, site visits, and review of administrative data.

The analysis did not include counterfactual sites, a choice driven by "the cost of surveying and by a belief that comparisons don't make sense for this type of intervention" (Fiester 2006, 17). Further, the evaluation was intentionally iterative, adapting as new knowledge emerged. For example, evaluators modified their research question probing increased communication between children and parents to one that probed the effectiveness of such communication. Uniquely, evaluators were part of the team implementing the program, rather than serving as completely independent observers. This was important because it allowed evaluators to "build local capacity for self-evaluation and to feed data back to implementers in real time" (Fiester 2006, 12). As the evaluation progressed, evaluators limited aspects of data collection, such as surveys, to three of the five sites to lower costs.

Evaluation Data Sources

To keep costs within the budget, evaluators limited data analysis to three sites (Atlanta, New Orleans, and San Diego), which were purposively chosen, as they were the first sites to get the program up and running and large enough to provide substantive data. The evaluation team relied on three types of data collection activities:

SURVEYS

Evaluators implemented two cross-sectional surveys that probed adolescents' knowledge, behaviors, and patterns of communication before and after Plain Talk. In each of the three sites, field workers randomly selected households to approach and interview; the survey was administered to all community members, not necessarily program participants, as the purpose of the program was to change overall community perspectives.

ETHNOGRAPHY

As qualitative data were a key component of this evaluation, ethnographers observed programming for 20 hours each week for over a year. The content of their observations included community outreach

and health education activities, interviewing adult participants and institutional partners, and collecting/reviewing written materials, among primary and secondary research activities.

ADMINISTRATIVE DATA REVIEW

Evaluators conducted quantitative analysis on hospital and health department data to contextualize the survey and ethnographic data. Administrative data they collected and analyzed included rates of pregnancy and birth by teenagers and the incidence of STDs among teens.

Summary of Evaluation Findings

The program's theory of change (below) had very specific steps theorized to change perspectives and reduce teen pregnancy and STD rates. Findings varied somewhat across the three sites but, in general, the outcomes evaluation concluded that there was a positive association between youth and knowledgeable adults communicating about sexuality, increased access to contraceptives, and teens' sexual knowledge and behavior. Increased knowledge within the community also made conversations easier and changed the dynamics from "no sex" toward safe sex or no sex. Other, more specific conclusions are as follows:

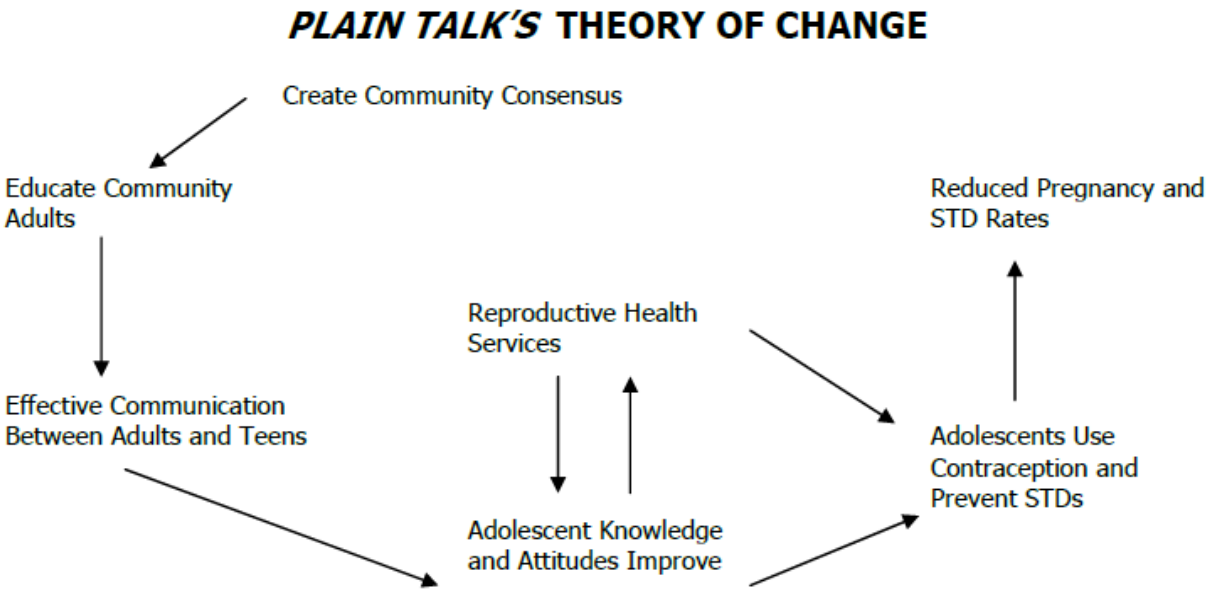
- The pregnancy rate declined in Plain Talk communities.
- Sites that used residents as trainers delivered more explicit sexual information and trained more residents than did sites that used professional trainers.
- Sites that focused equally on improving adults' knowledge about sex and their ability to communicate with teens about sexual issues had more success in changing youth's knowledge, attitudes and behaviors than sites that focused on one or the other.
- All sites saw improvements in the availability of reproductive health services.
- Including adults who were not parents, but who interact with youth in the program training, was valuable because parents were often reluctant to converse about sex with their children, even after training.

Overall, youth who talked to adults (as compared with those who did not), had more accurate knowledge about birth control, felt more comfortable about birth control, were more likely to have been tested for an STD, were less likely to have an STD were less likely to have created a pregnancy or become pregnant, and were less likely to have a child. These associations suggest that Plain Talk was working and that the improved communication led to improved outcomes for teen pregnancy and

sexual activity. The initiative's programming was so successful that it was replicated and disseminated as a "Plain Talk Starter Kit," which is a "practical guide to community-based programs seeking to reduce teen pregnancy, STDs and HIV/AIDS" (Grossman et al. 2001, 23). The Starter Kit is available here: Douglas, Elaine. "PlainTalk Starter Kit: A Practical Guide to Community-Based Programming to Reduce Teen Pregnancy, STDs and HIV/AIDS." Baltimore: Annie E. Casey Foundation.

FIGURE C.1

Plain Talk's Theory of Change



Source: Fiester (2006), 28.

Sources

Fiester, Leila. 2006. "'Looking for Shadows': Evaluating Community Change in the Plain Talk Initiative." Baltimore: Annie E. Casey Foundation.

Grossman, Jean Baldwin, Karen E. Walker, Lauren J. Kotloff, Sarah Pepper. 2001. "Adult Communication and Teen Sex: Changing a Community." Philadelphia: Public/Private Ventures.

Promise Neighborhoods Case Studies

Purpose/focus of study
- Determine how well the program was implemented
- Assess impact on children in program neighborhoods

Measures of change
- Program Implementation
  - Infrastructure development and service take-up rates
- Impact on early childhood education and well-being
  - Improved health and educational achievement

Data sources
- Program documents, interviews with program partners, site visits, and interviews with families for selected indicators of child well-being and program implementation, including healthy habits, stable housing, parents reading to them, and access to computer devices
- Administrative data for selected indicators of child well-being, including children's usual source of care, children ready to learn at kindergarten, children in early learning, children at or above grade level in math, reading, and language arts, attendance in junior high school, and high school graduation

Sample size or number of programs
- Five neighborhoods in program: three in the first cohort and two in the second—Berea, Kentucky, Buffalo, New York, Chula Vista, California, Los Angeles, and Minneapolis

Comparison group or benchmark
- Changes in indicators for population served over time

Time period or length of study
- 2010–14

Project Overview

Based on the Harlem Children's Zone model, Promise Neighborhoods seeks to offset the effects of growing up in poverty by building a comprehensive continuum of "cradle-to-career" supports so children are more likely to reach their potential. Launched in 2010 by the US Department of Education, three rounds of grants totaling nearly $100 million have been awarded to nonprofit organizations, institutions of higher education, and American Indian tribes. This includes 46 planning grants and 12 implementation grants to 48 lead agencies in 23 States and the District of Columbia. Implementation grants were awarded in two rounds (2011 and 2012), each for a five-year period. The Promise Neighborhoods Institute (PNI) at PolicyLink provides a national system of supports (such as technology tools, engagement with peers, and advice from experts) to Promise Neighborhoods and other communities interested in implementing similar place-based strategies.

To institute this cradle-to-career model, implementation grantees are expected to create a pipeline of supports that might include, for example:

- **early learning services,** such as training and professional development for existing center- and home-based providers, and support for new early learning programs;

- **K–12 educational supports,** including academic and enrichment activities provided before, during, and after regular school hours;
- **college and career preparation services,** such as mentoring, internships, college visits, application assistance, and career exploration; and
- **family and community supports,** such as parenting classes, adult education, health and nutrition services, housing assistance, community councils, and community gardens.

Grantees also are expected to construct theory-of-change or logic models to help guide their planning and strategies; foster partnerships to build and maintain a comprehensive, integrated, and sustainable system; develop longitudinal data systems to support decision making, quality improvement, and accountability; and report on 15 indicators annually that demonstrate progress toward achieving systems-level change.

To document the complexity of the Promise Neighborhoods and their implementation experiences, PNI contracted with Mathematica Policy Research to conduct five in-depth case studies of selected Promise Neighborhoods. The sample included three sites from the first round of implementation grantees (Berea, Kentucky; Buffalo, New York; and Northside Achievement Zone in Minneapolis, Minnesota) and two from the second round (Chula Vista and Los Angeles, California).

Evaluation Overview

The overall goal of the study was to document the complexity of implementing the Promise Neighborhoods cradle-to-career strategy on the ground and to identify the implications for developing a national evaluation that is true to the complex systems-change goals for the investments. The expected second phase of the current project will be to develop a national evaluation design informed by these case studies.

Three primary research questions guided the case studies:

1. How do Promise Neighborhoods build the infrastructure to support and sustain a pipeline of programs for children from birth through college and career?
2. How does the resulting system work on the ground? What are the take-up rates of high-quality services and schools?

3. Are Promise Neighborhoods meeting their partnership and service coordination goals? What barriers and facilitators do they face? What is needed to create a positive climate for successful partnerships and achievement of Promise Neighborhoods' goals?

Evaluation Data Sources

To develop comprehensive pictures of the grantees' Promise Neighborhoods and their implementation experiences, Mathematica collected and analyzed data from three sources:

1. **documents** collected from the sites and PNI, such as grant applications, organizational changes, sample partnership agreements, etc.;
2. **telephone interviews** with grantee directors; and
3. **site visits,** conducted in spring and summer 2014. Interviews and focus groups were conducted with the grantee management team, local evaluators, community partners, and participating families, and observations were made of key program activities. Enrollment and waiting lists were reviewed to determine take-up rates at various program sites.

The analytic approach was systematic, yet flexible, to assess and evaluate the large quantity of data collected.

Summary of Evaluation Findings

BUILDING INFRASTRUCTURE

To build infrastructure, the five case study sites took somewhat similar approaches. For example, the lead agency in each site expanded its own capacity to manage and provide structure for the complex efforts. This meant they needed to build their own areas of expertise and hire additional staff to fill new roles. Although lead agencies typically provided some direct services, they primarily partnered with schools, community-based organizations, government agencies, and other groups to cover the broad range of expertise needed to establish a cradle-to-career set of support services. To facilitate success, several key structures were developed. For example, sites created common data systems to share information and accountability for program success. Staffing structures were designed to facilitate ongoing communication throughout the pipeline. In some cases, staff from different programs or the lead agency were co-located at different program sites. In other programs, they spent time at the various locations. And financial support from a variety of sources, including local and national

foundations, private entities, and government agencies, was identified and pieced together to support ongoing and future Promise Neighborhood activities.

TAKE-UP RATES

The study found that take-up rates across sites and activities are driven by a combination of program capacity and participant interest. School-based activities reach the largest numbers of participants, however. Virtually all students who attended partner schools were touched in some way by Promise Neighborhoods' services. Smaller numbers participated in more intensive K–12 activities and in programs for younger children and adults.

MEASURING PERFORMANCE

Measuring performance was based on 15 indicators emanating from the Government Performance and Results Act (GPRA). These indicators, compiled from local and state program sources, included, for example, the number and percentage of young children participating in center- or home-based early learning programs; the number and percentage of students performing at or above grade level according to state mathematics and reading assessments; high school graduation rates, etc. Measuring performance was promising, but the results cannot be considered definitive at the early stage of the program in which this evaluation was completed. As expected, the first cohort of implementation grantees could report baseline data on most of the indicators, while the second cohort had fewer indicators to report. For those indicators with more than one year of data, the results were mixed. All sites reported upward trends in some measures and downward trends in others, but overall there were more upward than downward trends. The most consistent positive trends related to early childhood development. Although these early results are encouraging, they are based on only a few years of data and only five case studies. The researchers caution that the findings are not generalizable to Promise Neighborhoods.

CHALLENGES

A great deal was learned from the challenges encountered by these early implementation efforts. The five sites studied shared several experiences:

- **Lack of experience building a cradle-to-career continuum of solutions.** Although the organizations involved in the Promise Neighborhoods case studies all had experience serving their communities and working with partners, the size and complexity of this effort required new approaches. Building relationships among many different organizations can be a slow process. Developing and using joint data systems also takes time. PNI provided essential

technical support to help local sites set up their data systems and learn how to make data-driven decisions.

- **Varying levels of commitment and flexibility among stakeholders.** Although all the partners were drawn to Promise Neighborhoods' goals, each had its own organizational expertise, policies, and cultures. Blending these differing perspectives and approaches into a unified structure was difficult at times. The case study sites found that the policies and structures of school districts were often less flexible than those of other partners.
- **Staff and partner turnover.** Because program implementation can be a long-term process, it is not unusual for turnover to occur—either at the individual or organizational levels. These changes can result in slower implementation of the program and sometimes gaps in services.
- **Unrealistic expectations.** Funders and other stakeholders sometimes failed to realize that it will take more than two decades for children born in a new Promise Neighborhood to make their way through the full pipeline and complete college. An unrealistic desire for quick results to achieve program impacts was a considerable challenge.
- **A robust results framework with shared accountability.** Case study sites found it essential to receive training in how to use data to measure results, continuously improve, and share accountability. Training helped facilitate the effective use of data in these areas.
- **Strong interpersonal and institutional relationships.** Developing and maintaining a continuum of quality services requires strong relationships among a set of partners with a broad range of expertise. Sites found that colocation of staff and referral systems can facilitate the linkages needed to ensure seamless connections between service providers and families in transition. Building relationships with community members was essential. Promise Neighborhood staff noted the importance of being open to community suggestions.
- **Flexible, long-term, and sustainable capital.** Because of the long-term nature of Promise Neighborhoods' goals, the sites studied found it was essential to remain flexible as needs changed, new circumstances arose, and the staff responded to lessons learned. Because of the long-term nature of the program, the evaluators suggest that funders understand the need for flexibility and not make financial support contingent on using a particular program approach or partner. A long-term commitment to achieving the goals of Promise Neighborhoods will be critical for sustaining the program.

FACILITATING FACTORS

The case study findings indicated that two factors helped facilitate a positive climate and aided in achieving program goals: strong interpersonal and institutional relationships and a robust results framework with shared responsibility.

Source

Hulsey, Lara, et al. 2015. *Promise Neighborhoods Case Studies.* Final report submitted to Promise Neighborhoods Institute at PolicyLink. Princeton, NJ: Mathematica Policy Research.

Note

This additional resource provides an annotated list of Promise Neighborhood related resources. On pages 7–14, it provides 25+ sources and links related to collecting, analyzing, and using data in Promise Neighborhoods.

An Annotated List of Promise Neighborhoods Resources is located here: https://www.urban.org/research/publication/annotated-list-promise-neighborhoods-resources.

Strive Together Initiative in Bexar County, Texas

Purpose/focus of study
- Evaluate the initiative's effectiveness on enrolled children

Measures of change
- Increases in early grade reading scores
- Increases in middle grade math scores
- High school graduation rates
- Postsecondary enrollment and completion
- Employment success

Data sources
- Administrative and district and county level data

Sample size or number of programs
- 16 school districts within Bexar County Texas

Comparison group or benchmark
- Comparison group measure: progress against the county and statewide averages for each measure
- Benchmark measure: to reach an identified goal for each indicator by a specified date.

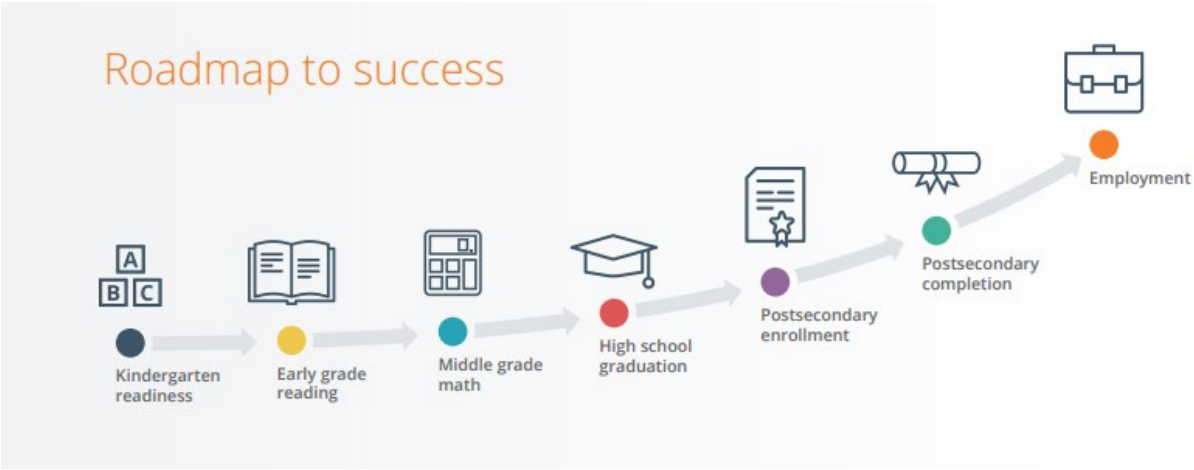
Time period or length of study
- 2011–12
- Program is ongoing

Project Overview

Started in 2010, Strive Together is a national, nonprofit network that coordinates resources in over 70 communities to support all children, regardless of their race or socioeconomic status. The Strive Together Cradle to Career Network aims to provide all children with equal promise and equal opportunity by supporting young people with an intentional roadmap to success (see figure C.2). The roadmap identifies seven key life milestones: kindergarten readiness, early grade reading, middle grade math, high school graduation, postsecondary enrollment, postsecondary completion, and employment. One of the 70+ communities is the P16Plus community of Bexar County in Texas. The P16Plus Council of Greater Bexar County has been affiliated with Strive Together since 2014 and implements Strive Together's collective impact framework to see the impact of their six programs in the county. The P16Plus initiatives range from "ReadyRosie," which supports kindergarten readiness, to "Diplomas Project" and "GenTX San Antonio," which engage students who are enrolling, attending, or graduating from postsecondary education. Sixty-four percent of Bexar County is economically disadvantaged, 12 percent are English language learners, 73 percent are Hispanic/Latino, and 10 percent are special needs students. P16Plus is largely guided by Strive Together's theory of action, which includes four key activities: engage the community, eliminate locally defined disparities, develop a culture of continuous improvement, and leverage existing assets.

FIGURE C.2

Strive Together Cradle-to-Career Roadmap to Success



Source: Strive Together (2017).

Evaluation Overview

In 2016, the P16Plus council and staff evaluated the efficacy of their programs using multiple forms of data analysis. They produced a Community Impact Report Card that assessed their use of Strive Together's collective impact framework. For each indicator, P16Plus identified a definition of the indicator and the data source to measure progress. For six of the seven milestones in Strive Together's roadmap to success, P16Plus identified indicators of success and how each of their (sometimes) overlapping programs address the indicators. For example, for early grade reading, P16Plus's goal is that "70% of third grade students will be reading at grade level by 2030." To measure against this goal, P16Plus created a core indicator for this goal ("passing rate of third grade students in reading on the State of Texas Assessments of Academic Readiness"). To measure progress toward this goal, they compare the district's passing rate against the state average, and compare each of their 16 district's rates for the third-grade reading level with that for the entire county, the county's economically disadvantaged, and different county and state population groups (race or ethnicity, economically disadvantaged, gender, and special education). See figure C.3 for more details.

FIGURE C.3

P16Plus Indicator Definitions and Sources

Indicator	Definition	Source
Kindergarten Readiness	Rate of students "very ready" in at least four of five domains (at or above 75th percentile)	Early Development Instrument, United Way of San Antonio and Bexar County
Third Grade Reading	Rate of students meeting Final Recommended Level 2 standard and 2016 Level 2 standard on third grade reading STAAR	Texas Education Agency STAAR Aggregate Data
Eighth Grade Math	Rate of students meeting Final Recommended Level 2 standard and 2016 Level 2 standard on eighth grade math STAAR	Texas Education Agency STAAR Aggregate Data
High School Graduation	Four-year federal graduation rate: percent of students in a cohort who received their high school diploma within four years, with limited exclusions	Texas Academic Performance Reports
Post-Secondary Enrollment	Rate of Bexar County high school graduates who enrolled in a Texas institution of higher education in the fall after their high school graduation	Texas Higher Education Coordinating Board, High School Graduates Enrolled in Texas Higher Education
Post-Secondary Attainment	Rate of Bexar County residents ages 25-34 who have an associate's degree or higher plus the rate of residents with a certificate estimated using the ratio of annual credentials awarded in Bexar County	American Community Survey 1-year Public Use Microdata Sample; Certificate estimate from Texas Higher Education Coordinating Board

Source: P16Plus (2017), 20.

Evaluation Data Resources

Six factors contribute to the program council and staff's ability to build capacity in education-related data and track student performance:

- Strong relationships with school districts and higher education institutions, which provide access to comparative data
- Access to and use of public data sources:
  - Individual program data
  - Early Development Instrument, United Way of San Antonio and Bexar County
  - US Census and American Community Survey data
  - Aggregate school district and Texas Education Agency data
  - Texas Higher Education Coordinating Board data

    - » US Department of Education data

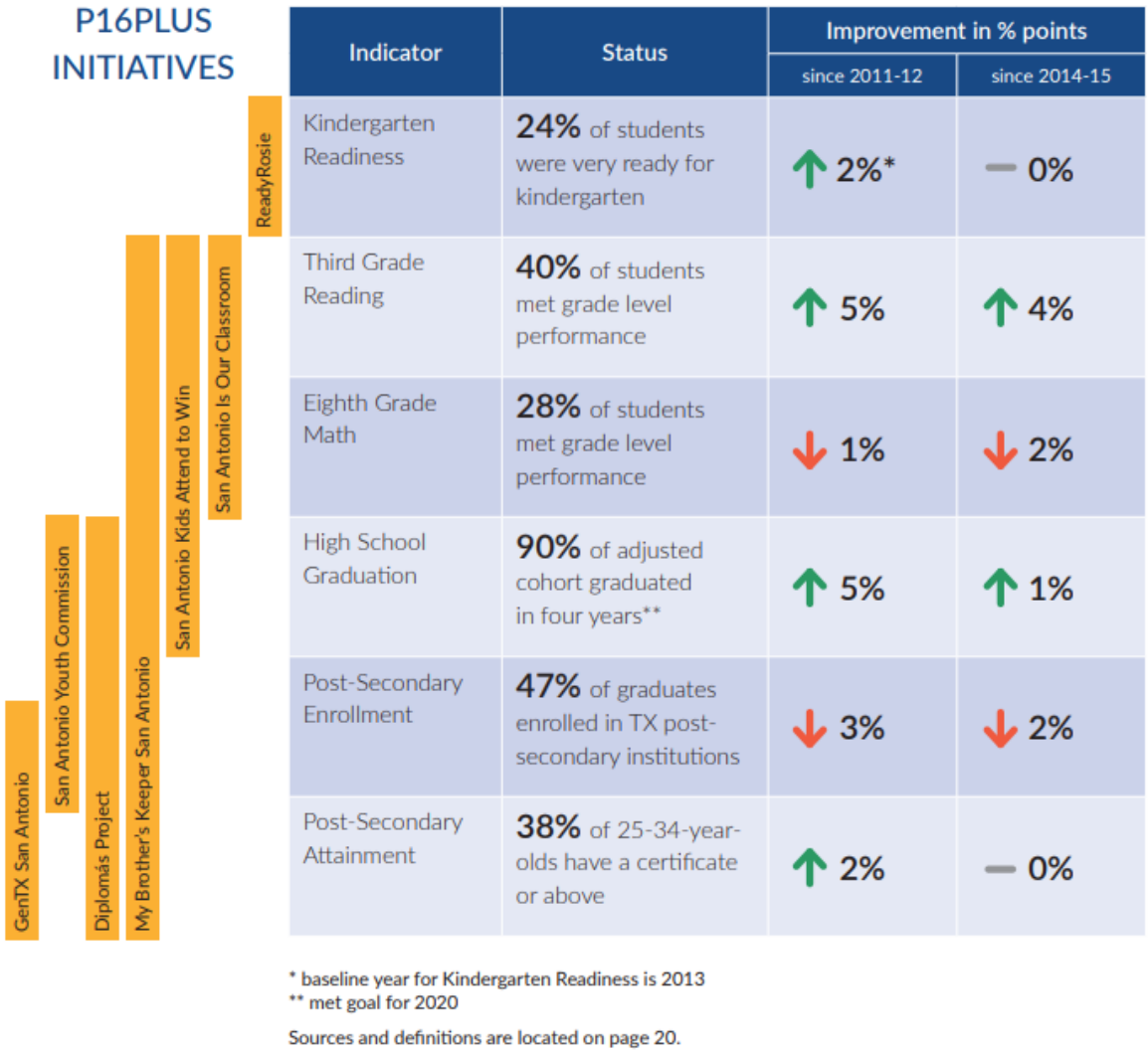
- Data Support Council, consisting of education data specialists from multiple school districts, higher education institutions, and community-based organizations, which guides the program's data initiatives
- Strive Together affiliation, which provided some outside evaluation of collective impact by the Strive Together network and provided a benchmark upon which to compare later performance
- Continuous Quality Improvement techniques that were learned and used, including data coaching and other tools and resources that enable partners to identify what is working and adjust their strategies to take account of the evidence
- P16Plus data team, which includes key individuals from the Bexar County initiatives

Summary of Evaluation Findings

Figure C.4 summarizes the six key goals and core indicators for the P16Plus programs in 2017.

FIGURE C.4

P16Plus Indicator Summary



Indicator	Status	Improvement in % points	
		since 2011-12	since 2014-15
Kindergarten Readiness	**24%** of students were very ready for kindergarten	↑ 2%*	— 0%
Third Grade Reading	**40%** of students met grade level performance	↑ 5%	↑ 4%
Eighth Grade Math	**28%** of students met grade level performance	↓ 1%	↓ 2%
High School Graduation	**90%** of adjusted cohort graduated in four years**	↑ 5%	↑ 1%
Post-Secondary Enrollment	**47%** of graduates enrolled in TX post-secondary institutions	↓ 3%	↓ 2%
Post-Secondary Attainment	**38%** of 25-34-year-olds have a certificate or above	↑ 2%	— 0%

* baseline year for Kindergarten Readiness is 2013
** met goal for 2020

Sources and definitions are located on page 20.

Source: P16 Plus (2017), 6.

Sources

P16 Plus. 2017. *Community Impact Report Card.* San Antonio, TX: Council of Greater Bexar County.

Strive Together. 2017. "Better Systems for Better Outcomes." Cincinnati, OH: Strive Together.

Strive Together. 2017. "8 Million Students, One Vision." Cincinnati, OH: Strive Together.

Note

Additional information about this ongoing initiative may be found at https://p16plus.org/.

Urban Health Initiative

Purpose/focus of study
- Determine if program achieved scale
- Determine if intervention was effective in changing outcomes for children and young people

Measures of change
- For scale
  - Number reached and served
  - Improved child health and safety in treatment communities
  - Changes in incidence of substance abuse, sexually transmitted disease, teen pregnancy, violence, crime incidence, and several educational achievement measures

Data sources
- Site visits and interviews with key informants, parents, and children
- Administrative data

Sample size or number of programs in treatment group
- Five cities—Baltimore, Detroit, Oakland, Philadelphia, and Richmond

Comparison group or benchmark
- Nine cities with matching demographics—Baton Rouge, Birmingham, Boston, Cleveland, Milwaukee, Newark, Pittsburgh, and St. Louis (compared as one group with another)

Time period or length of study
- 1996–05

Project Overview

The Urban Health Initiative (UHI) was a 10-year, $63 million program funded by the Robert Wood Johnson Foundation, operating between 1996 and 2005 with the goal of improving community-wide health outcomes for children in five cities: Baltimore, Detroit, Oakland, Philadelphia, and Richmond. The Foundation initially invited twenty cities to apply for funding and selected eight to undergo two years of data-driven planning. Five cities received an additional eight years of grant funding to implement their strategic plans. Grant funding was not meant to fund services but to support cities in creating systemic change. The Foundation's vision for the UHI was "to test whether, by bringing together a broad cross-section of the community to work together over a sustained period of time, it is possible to make measurable improvements in the health and safety of children" (Brown 2013, 4). "Most importantly, the cities were to implement these new strategies on a large scale—large enough to change child health statistics citywide" (Brown 2013, 1).

Lead agencies in each city undertook an inclusive planning process, identified the city's most pressing child safety problems, and mobilized local leaders, organizations, and residents to leverage resources and implement interventions to address those problems. Cities set local outcome targets which included, for example, increasing school readiness and third-grade reading skills and reducing preterm births, infant mortality, child abuse and neglect, child accidents and injuries, adolescent childbearing, youth use of alcohol, drugs and tobacco, and children's exposure to crime or violence. To

address their local goals, cities expanded or enhanced after-school, neighborhood safety, and parent home-visiting programs and supports for child care providers, and implemented other initiatives to support at-risk youth.

Evaluation Overview

The Foundation engaged a team of New York University researchers to evaluate UHI's implementation and impact. The evaluation addressed three primary research questions:

- Can a foundation-sponsored initiative be a catalyst for a cross-sector, collaborative process?
- Does such a process result in meaningful changes in policies and programs designed to serve children and youth in urban settings?
- To what extent do these changes in policies and programs improve the health and safety of children and youth?

The evaluation was focused on understanding impacts across the five sites, compared with evaluating each site individually. To strengthen their ability to draw causal conclusions about the impact of the UHI, evaluators developed a research plan that integrated two designs: theory of change and a quasi-experimental comparison group.

In the theory-of-change component, evaluators worked with the Foundation and local sites to articulate the program inputs and activities and their linkages to expected interim and final outcomes (figure C.5). Evaluators then collected data to determine whether the activities and interim outcomes unfolded as expected. For example, "Did the sites engage in a multisector planning process and did this process result in a shared vision of youth problems?"

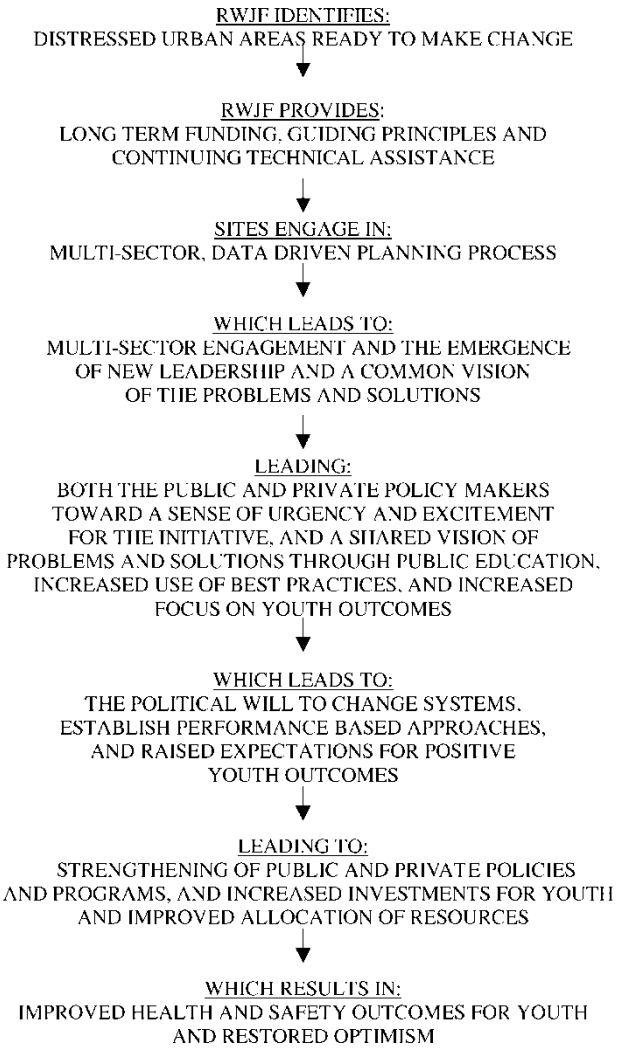
In the comparison group component, the evaluators first selected a group of nine comparison cities that were similar in socioeconomic and demographic characteristics. For the analysis, they contrasted activities, interim outcomes, and final outcomes in the UHI cities with the comparison cities to determine which activities and outcomes could reasonably be attributed to UHI and which would have occurred even in the absence of UHI. These comparisons proved to be an important element of the evaluation design. On some outcome measures, the indicators for the UHI cities declined over time, but those declines were smaller than in the comparison cities. This finding led the evaluators to conclude that the initiative had a positive impact.

Integrating theory-of-change and quasi-experimental comparison group modeling, evaluators attributed results to UHI only when two conditions were met. Results had to be

- predicted by program theory (see figure C.5 below); and
- occur in an area theorized to be affected by the program activities that were implemented.

FIGURE C.5

Abridged Theory of Change for the Robert Wood Johnson Foundation's Urban Health Initiative



Source: Republished with permission of SAGE Publications, Inc. Journals, from "Integrating a Comparison Group Design into a Theory of Change Evaluation: The Case of the Urban Health Initiative," *American Journal of Evaluation* 23, no. 4 (2002): 376; permission conveyed through Copyright Clearance Center, Inc.

Evaluation Data Sources

The evaluation team relied on five types of data:

- **Annual site visits** to each of the five UHI program city sites to collect information about progress and challenges. The team also visited the comparison cities to collect data on the approaches they were using to improve children's health and safety.
- **Interviews with 15 to 20 key informants every 12 to 18 months** in the five UHI cities to learn about the challenges and opportunities facing the community. The team also conducted interviews with civic leaders in the comparison cities to determine if they were implementing efforts similar to UHI.
- **Telephone survey of parents and children** conducted at three points in time with representative samples in the UHI cities, the comparison cities, and with a nationally representative sample. The surveys collected information about neighborhood and city conditions, performance of city institutions, youth problems, school and after-school activities, and other child-related challenges.
- **Secondary data on child health and safety outcomes** in both the UHI and comparison cities that could be tracked over time. These data included FBI crime reports, rates of sexually transmitted infectious diseases from the Centers for Disease Control and Prevention, educational data from the National Center for Education Statistics, vital statistics (birth and death records), and census data.
- **Secondary data on public expenditures on children and youth** in UHI program cities from three points in time.

Summary of Evaluation Findings

A final overview of the UHI summarizes the evaluators' overarching conclusions as follows:

- A foundation-sponsored initiative can encourage a broad and deep collaborative process. Although there are many projects around the country that are trying to address youth-related issues, many of these other efforts are narrow in scope compared with what was undertaken in the UHI.

- The process followed in the UHI model (i.e., collaboration, partnerships, data collection, etc.) resulted in some meaningful change in every site, even though the process was not always achieved as originally envisioned.
- Compared with similar cities, the five UHI cities demonstrated some measurable changes in policies and programs to affect the health and safety of children and youth, but these gains were often small. The results did not match the ambitious goals set out by UHI.

Specific findings included, for example:

- The UHI sites found ways to "work smarter for kids" and at least partially achieved many of their medium-term outcome goals. Comparison cities lacked the coordinated and consistent efforts that UHI cities undertook.
- There was no evidence that UHI sites were able to redirect public funds from remedial to preventive approaches.
- Interviews with and surveys of civic leaders and residents revealed some signs of a greater shared vision and more civic engagement around issues of health and safety of children.
- Parents perceived that something different was happening in their cities and that UHI was making modest improvements in their communities.
- Relative to parents in the comparison cities the survey data showed that parents in the program cities
  - » became more engaged in civic matters, as measured by respondents' self-report of voting in local elections;
  - » were more positive about their city's efforts to improve youth well-being;
  - » were less concerned about teen pregnancy and vandalism;
  - » viewed the availability and quality of after-school programs and other out-of-school services favorably;
  - » perceived positive changes in their neighborhoods and were more likely to rate their neighborhood as being a good or excellent place to raise kids and as having safe places for kids to play.
- UHI cities showed more positive trends (that is, had lower rates) on the following indicators relative to the comparison cities: births with late or no prenatal care; low birthweight births; second or later births to young mothers; mortality because of accidents, violence or suicide; youth arrests; and gonorrhea cases.

- UHI cities showed less positive trends (that is, had the same or higher rates) on the following indicators relative to the comparison cities: births to young mothers; infant mortality; school attendance; and high school graduation rates.

Outcome Tracking at the Local Level

In addition to the initiative-wide evaluation, each UHI city set performance targets and tracked progress toward those targets. Cities also reported on changes in funding, collaborations, and systems that resulted from their involvement in UHI. For example, the Baltimore initiative (called Safe and Sound) reported the health and safety of children and youth was measurably improved, with the only goal area showing no progress being the effort to improve young people's perceptions of adult and community support:

- In 2002, the city's infant mortality rate was the lowest on record—28 percent lower than in 1997.
- Child abuse and neglect decreased 32 percent compared with 1997, and teen births were down 25 percent.
- Reading scores of third graders improved by 64 percent over 1997.
- The high school graduation rate rose by 28 percent since 1997.
- In 2004, 40 percent of students entering kindergarten were assessed as fully ready to learn, compared with 27 percent the previous year.
- The city showed "modest improvements in the reduction of violent crime, homicides and arrests" (Brown 2013, 26).

UHI also changed the way Baltimore plans, evaluates, and funds services for children and youth.

- As a result of the RWJF initiative, the city now encourages public and private funders to use data, outcome measures, and the size and scale of the program when making funding decisions.
- Safe and Sound also worked with private-sector funders to negotiate agreements (called compacts) with state child-serving agencies to shift more funds to preventive services. This new policy and financing concept became known as the Maryland Opportunity Compact.

  By improving the effectiveness of citywide strategies and launching the Maryland Opportunity Compact, Safe and Sound, as of 2005, had leveraged more than $100 million to support efforts to help children and families and to sustain its own infrastructure.

Sources

Brown, Michael H. 2013. *Urban Health Initiative: Working to Ensure the Health and Safety of Children*. Princeton, NJ: Robert Wood Johnson Foundation.

Weitzman, Beth C., Diana Silver, and Keri-Nicole Dillman. 2002. "Integrating a Comparison Group Design into a Theory of Change Evaluation: The Case of the Urban Health Initiative." *American Journal of Evaluation* 23 (4): 371–85.

Vibrant Communities

Purpose/focus of study
- Determine the effectiveness of the network built as part of the project
- Determine the impact of the program on households in the communities treated

Measures of change
- Network operation
  - Use of intranet
  - Use of implementation resources
  - Community networking
- Reduction in poverty and increase in employment
  - Number of people with increases in assets and employment income
  - Policy changes that increase access to services and program participation

Data sources
- Site visits and interviews and staff surveys, website downloads for network operation, and policy changes
- Staff surveys, site visits, estimated number of beneficiaries of policy change, administrative/census data for reduction in poverty, and increased employment

Sample size or number of programs
- Thirteen Canadian communities and 170,000 households touched (27,000 comprehensively)

Comparison group or benchmark
- Number of people in the communities treated with increases in assets, employment income, cost savings (from subsidized transportation, for example), and improved housing

Time period or length of study
- 2002–12

Project Overview

Vibrant Communities (VC) was a foundation-funded pan-Canadian initiative conducted from 2002–2012 to implement place-based poverty reduction in 13 Canadian cities. In each city (called Trail Builder cities), the initiative implemented programmatic and systemic interventions, guided by multisectoral collaborations specific to each city. Programmatic interventions were designed to help people build human capital assets such as money management or job search skills; systemic interventions were intended to alter policies and systems that shape people's life prospects, such as providing more access to fresh food or developing a city bus pass system. Partners included local nonprofits, local governments, and individuals with personal experience living in poverty. Each city was encouraged to define and implement programming specific to that city's needs, as long as they followed the program's five core principles: (1) poverty reduction, (2) comprehensive thinking and action, (3) multisectoral collaboration, (4) community asset building, and (5) community learning and change.

The national sponsors provided a variety of mechanisms to link VCs to a "learning community," whereby cities and individuals could share ideas and learn from one another. The national sponsors helped establish an extensive website, regular e-newsletters, monthly convener calls, tele-learning

sessions with experts on various topics, and a coaching system to help individuals and groups create and develop their overall approaches.

Evaluation Overview

The three primary funders engaged an external consultant to work with VC staff to evaluate the initiative's implementation and impact. The evaluation team collaborated closely with internal and external stakeholders to prioritize research questions and conduct analysis and interpretation. The evaluation had two phases: first, evaluators summarized findings from existing statistics, case studies, and other reports from the initiative; and second, evaluators deepened the understanding of the role the national supports played in contributing to the VC outcomes. The case study described here focuses on the evaluation's second phase because much less is known about the types of infrastructure that can support the work of a broad-based and ambitious initiative.[9]

Evaluation Data Sources and Analytic Methods

The evaluation of the national support effort relied on four different types of data:

1. Two web-based surveys of local staff in Trail Builder cities and their partners, as well as people in non-Trail Builder cities who subscribed to the e-newsletter or participated in tele-learning sessions. The surveys asked, for example, about patterns of usage (i.e., how often did you go online; how much/what did you read); the perceived value of the supports; and how much the supports affected local poverty reduction efforts.

2. National staff monitored and tracked over time the number of downloads from the website, the number of subscribers to the e-newsletter, the number of participants at face-to-face or tele-learning events, etc.

3. In-depth studies of five communities were conducted to describe the linkages between the national supports and the local outcomes. During site visits or interviews, participants might be asked, for example, how much and in what ways the convening calls, coaching, or face-to-face events helped them.

[9] The phase I evaluation used existing statistical data, case studies, and interviews to investigate the VC work. The study concluded that because poverty takes many forms, successful outcomes require a comprehensive approach. The progress that VCs made to reduce poverty, create systems change, and build community capacity was uneven across the 13 VC sites.

4. An ad hoc committee of experts reviewed materials and provided additional perspective on the evaluation findings.

Summary of Evaluation Findings

For an action-oriented initiative as broad and complex as VC, the evaluators concluded that providing supports and infrastructures to facilitate the work is essential. The initiative created and modified a set of communication tools that supported the partners. During the time the initiative was active, a number of electronic forms of communication, such as websites, tele-learning and e-newsletters became common, and these were used in addition to print reports and other resources to share information. These communication tools helped VC cities create new approaches, experiment with new ideas, and share their strategies and implementation lessons with others. The evaluators made the following recommendations for future projects:

- **Intermediaries and grantees** should recognize that no single remedy will address the complexities of poverty, and successful strategies must be customized for the community in which they are working. A comprehensive and mutually reinforcing approach is essential for achieving successful outcomes. Also, different kinds of supports and guidance are likely to be needed at different times during the process.
- **Partnerships are essential for successful community change.** Strong skills are needed to build partnerships. These relationships require time and nurturing to develop but can bring additional support and momentum to the initiative as it unfolds.
- **A Community of Practice** has clear benefits for organizations and they should take the time to invest in them. Peer learning can be a valuable source of ideas.
- **Funders** should direct resources to communication supports that will enable intermediaries to learn and foster their creativity. The varied communication methods described above were extremely useful in sharing lessons learned and new ideas, as well as providing sounding boards for program operators. The evaluators also recommend that it can be beneficial for the funder to take an active role in this type of work so they, too, learn and grow in the process.
- **Local groups and communities** should encourage peer-to-peer learning to generate new and innovative ideas and create support for the implementation process.

Sources

Gamble, Jamie. 2010. "Evaluating Vibrant Communities: 2002–2010." Waterloo, Ontario: Tamarack—An Institute for Community Engagement.

Gamble, Jamie. 2012. "Inspired Learning: An Evaluation of Vibrant Communities' National Supports 2002-2012." Waterloo, Ontario: Tamarack—An Institute for Community Engagement

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 104 (490): 493–505.

Athey, Susan, and Guido W. Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32.

Bartlett, Susan, Jacob Klerman, Lauren Olsho, et al. 2014. *Evaluation of the Healthy Incentives Pilot (HIP): Final Report.* Prepared by Abt Associates for the US Department of Agriculture, Food and Nutrition Service.

Bloom, Howard. 2006. "The Core Analytics of Randomized Experiments for Social Research." Oakland, CA: MDRC.

Bloom, Howard, James A. Riccio, and Nandita Verma. 2005. "Promoting Work in Public Housing: The Effectiveness of Jobs-Plus." Oakland, CA: MDRC.

Brown, Michael H. 2013. "Urban Health Initiative: Working to Ensure the Health and Safety of Children." Princeton, NJ: Robert Wood Johnson Foundation.

Brown, Prudence. 1996. "Comprehensive Neighborhood-Based Initiatives." *Cityscape* 2 (2): 161–76.

Card, David, and Alan Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84 (4): 772–93.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2015. "The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment." Cambridge, MA: National Bureau of Economic Research.

Chigas, Diana, Madeline Church, and Vanessa Corlazzoli. 2014. "Evaluating Impacts of Peacebuilding Interventions: Approaches and Methods, Challenges and Considerations." London, UK: Conflict, Crime and Violence Results Initiative (CCVRI).

Cornwall, Andrea, and Rachel Jewkes. 1995. "What Is Participatory Research?" *Social Science & Medicine* 41 (12): 1667–76.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *The Quarterly Journal of Economics* 128 (2): 531–80.

Davies, Ian C. 1999. "Evaluation and Performance Management in Government." *Evaluation* 5 (2): 150–9.

Deaton, Angus, and Nancy Cartwright. 2016. "Understanding and Misunderstanding Randomized Controlled Trials." Cambridge, MA: National Bureau of Economic Research.

Dehejia, Rajeev H., and Sadek Wahba. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84 (1): 151–61.

Diaz, Jose Y., Sarah Gehrig, Ellen Shelton, and Cael Warren. 2015. *Prospective Return on Investment of the Northside Achievement Zone.* St. Paul, MN: Wilder Research.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2008. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, vol. 3, edited by T. Schultz and John Strauss. Amsterdam: North Holland Publishing Co.

Fiester, Leila. 2006. "'Looking for Shadows': Evaluating Community Change in the Plain Talk Initiative." Baltimore: Annie E. Casey Foundation.

Fiester, Leila. 2011. "Measuring Change While Changing Measures: Learning in, and from, the Evaluation of Making Connections." Baltimore: The Annie E. Casey Foundation.

Gamble, Jamie. 2010. "Evaluating Vibrant Communities: 2002–2010." Waterloo, Ontario: Tamarack—An Institute for Community Engagement.

Gamble, Jamie. 2012. "Inspired Learning: An Evaluation of Vibrant Communities' National Supports 2002–2012." Waterloo, Ontario: Tamarack—An Institute for Community Engagement.

Grossman, Jean Baldwin, Karen E. Walker, Lauren J. Kotloff, and Sarah Pepper. 2001. "Adult Communication and Teen Sex: Changing a Community." Baltimore: The Annie E. Casey Foundation.

Hatry, Harry P. 2013. "Sorting the Relationships Among Performance Measurement, Program Evaluation, and Performance Management." *New Directions for Evaluation* 2013 (137): 19–32.

He, Feng J., Caryl Anne Nowson, Mn Lucas, and Graham A. MacGregor. 2007. "Increased Consumption of Fruit and Vegetables Is Related to a Reduced Risk of Coronary Heart Disease: Meta-Analysis of Cohort Studies." *Journal of human hypertension* 21 (9): 717.

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

Hulsey, Lara, Andrea Mraz Esposito, Kimberly Boller, and Sarah Osborn. 2015. *Promise Neighborhoods Case Studies.* Princeton, NJ: Mathematica Policy Research.

Idzelis Rothe, Monica, Ellen Shelton, and Greg Owen. 2014. *Northside Achievement Zone: 2013 Community Survey Results: A Follow-Up to the 2010 Baseline Survey*. St. Paul, MN: Wilder Research.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2): 615–35.

Leeuw, Frans L. 2012. "Linking Theory-Based Evaluation and Contribution Analysis: Three Programs and a Few Solutions." *Evaluation* 18 (3): 348–63.

Ludwig, Jens, Jeffrey R. Kling, and Sendhil Mullainathan. 2011. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives* 25 (3): 17–38.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *The Review of Economic Studies* 60 (3): 531–42.

Mayne, John. 2001. "Addressing Attribution through Contribution Analysis: Using Performance Measures Sensibly." *The Canadian Journal of Program Evaluation* 16 (1): 1–24.

Morgan-Lopez, Antonio, and Anupa Bir. 2017. *Unpacking the "Black Box" of Programs and Policies: A Conceptual Overview of Mediation Analysis*. OPRE Report #2017-01. Washington, DC: US Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

National Association for State Community Services Programs (NASCSP). 2018. "FFY 2017 CSBG Highlights: From the FFY Community Services Block Grant Information System (CSBG IS) Survey." Washington, DC: National Association for State Community Service Programs.

Nichols, Austin. 2013. "Evaluation of Community-Wide Interventions." Washington, DC: Urban Institute.

Popkin, Susan J., Bruce Katz, Mary K. Cunningham, Karen D. Brown, Jeremy Gustafson, and Margery A. Turner. 2004. "A Decade of HOPE VI: Research Findings and Policy Challenges." Washington, DC: Urban Institute.

Preskill, Hallie, Marcie Parkhurst, and Jennifer S. Juster. 2014. "Guide to Evaluating Collective Impact: Part 2: Assessing Progress and Impact." Washington, DC: Collective Impact Forum, FSG.

Prinz, Ronald J., Matthew R. Sanders, Cheri J. Shapiro, Daniel J. Whitaker, and John R. Lutzker. 2016. "Population-Based Prevention of Child Maltreatment: The US Triple P System Population Trial." *Prevention Science* 17 (3): 410–16.

P16 Plus. 2017. "Community Impact Report Card." San Antonio, TX: Council of Greater Bexar County.

Robins, James M., and Miguel A. Hernan. 2009. "Estimation of the Causal Effects of Time-Varying Exposures." In *Longitudinal Data Analysis*, edited by Garrett Fitzmaurice, Marie Davidian, Geert Verbeke, and Geert Molenberghs. New York: Chapman and Hall/CRC Press.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.

Schorr, Lisbeth, B., and Frank Farrow. 2011. "Expanding the Evidence Universe: Doing Better by Knowing More." Washington, DC: Center for the Study of Social Policy.

Shelton, Ellen, Cael Warren, and Sarah Gehrig. 2017. "NAZ 2016 Annual Report, Including Accomplishments Over Five Years as a Promise Neighborhood." St. Paul, MN: Wilder Research.

Strive Together. 2017. "Better Systems for Better Outcomes." Cincinnati, OH: Strive Together.

Strive Together. 2017. "8 Million Students, One Vision." Cincinnati, OH: Strive Together.

US Department of Health and Human Services (HHS). 2015. "Community Services Block Grant: Report to Congress, Fiscal Year 2014." Washington, DC: HHS.

Walker, Chris, Chris Hayes, George Galster, Patrix Boxall, and Jennifer Johnson. 2002. "The Impact of CDBG Spending on Urban Neighborhoods." Washington, DC: Urban Institute.

Weitzman, Beth C., Diana Silver, and Keri-Nicole Dillman. 2002. "Integrating a Comparison Group Design into a Theory of Change Evaluation: The Case of the Urban Health Initiative." *American Journal of Evaluation* 23 (4): 371–85.

About the Authors

William J. Congdon is a principal research associate in the Center on Labor, Human Services, and Population. His research focuses on issues in labor market policy and social insurance, and his recent work emphasizes the perspective of behavioral economics and the role of experimental methods for understanding economic outcomes and developing public policy.

Margaret C. Simms is a nonresident fellow at the Urban Institute. Until April 2018, she was an Institute fellow and director of the Low-Income Working Families project. A nationally recognized expert on the economic well-being of African Americans, her current work focuses on low-income families, with an emphasis on employment and asset building.

Carol De Vita is a former Urban Institute associate in the Center on Nonprofits and Philanthropy, specializing in the role, capacity, and financial well-being of community-based organizations, including faith-based organizations. Her current research focuses on capacity-building efforts to improve the performance of nonprofit organizations and government programs through outcome measures and performance management.

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.

STATEMENT OF INDEPENDENCE

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org