NO. 01-19 · JUNE 2019

COMMUNITY & ECONOMIC DEVELOPMENT
DISCUSSION PAPER

What Works at Scale:

A Framework to Scale Up Workforce Development Programs

Alexander Ruder
Community and Economic Development Department, Federal Reserve Bank of Atlanta

Primary issue:

Workforce development policymakers have access to a growing evidence base of successful training programs backed by rigorous randomized controlled trials. This evidence base identifies programs that work in specific geographic and temporal contexts but may not necessarily work in new contexts or at a scale sufficient to meet regional workforce needs.

Key findings:

First, the author examines a sample of recent workforce development randomized controlled trials and reports to what extent this body of evidence informs policymakers about what works at scale. The author finds that most evaluations provide limited evidence about what works at scale: evaluations are implemented with relatively few participants, use nonrandom samples from the population of interest, and are concentrated in the most populous urban areas and U.S. states. Second, the author proposes a framework to help regional workforce policymakers apply evidence-based programs in their unique contexts and at a scale sufficient to achieve their strategic goals. The author reviews mechanism mapping, which policymakers can use systematically to identify ways an evidence-based program may fail in a new context. The author then reviews sensitivity analysis, which is commonly used in cost-benefit analysis to assess how violations in key assumptions affect expected program impacts. The author illustrates the framework with an example based on a published evaluation of a workforce development program for youth.

Takeaways for practice:

Policymakers who incorporate mechanism mapping and sensitivity analysis may identify weaknesses in an overall strategic plan, suggest adjustments to program implementation, and ultimately help stakeholders make a more informed decision about the adoption of workforce development programs.

FEDERAL RESERVE BANK of ATLANTA

Follow Atlanta Fed CED on

The Federal Reserve Bank of Atlanta's Community & Economic Development (CED) Discussion Paper Series addresses emerging and critical issues in community development. Our goal is to provide information on topics that will be useful to the many actors involved in community development—governments, nonprofits, financial institutions, and beneficiaries. Find more research, use data tools, and sign up for email updates at **frbatlanta.org/commdev**.

No. 01-19 • June 2019

What Works at Scale?

A Framework to Scale Up Workforce Development Programs

Abstract:

Workforce development policymakers have access to a growing catalog of training programs evaluated with rigorous randomized controlled trials. This evidence base identifies programs that work in specific geographic and temporal contexts but may not necessarily work in other contexts or at a scale sufficient to meet regional workforce needs. The author examines a sample of recent randomized controlled trials of workforce development programs and reports to what extent this body of evidence informs policymakers about what works at scale. The author finds that most programs are implemented at a small scale, use nonrandom samples from the population of interest, and are concentrated in the most populous urban areas and U.S. states. The author then discusses a method to help state and local policymakers, technical colleges, training providers, and other workforce development organizations adopt evidence-based policies in their local contexts and at scale. The two-step method includes a check on the assumptions in a program's theory of change and an assessment of the sensitivity of projected results to violations in assumptions such as program completion rates. The author provides an example of the method applied to a hypothetical metropolitan area that seeks to adopt an evidence-based training program for youth with barriers to employment.

JEL classification: I38, J08, J24

Key words: workforce development, human capital, skills, provision and effects of welfare programs

https://doi.org/10.29338/dp2019-01

About the Author:

Alexander Ruder is a senior adviser in Community and Economic Development at the Federal Reserve Bank of Atlanta, specializing in workforce and economic development policy. Previously, Ruder was an assistant professor (tenure track) in public policy and a Center for Innovation in Higher Education economics of education fellow at the University of South Carolina. He has also held positions as a research project manager at the John J. Heldrich Center for Workforce Development at Rutgers University and as the Illinois workNet business services coordinator at the Illinois Department of Commerce and Economic Opportunity. Ruder's scholarly work has appeared or is forthcoming in *Economics of Education Review, Journal of Public Policy, Presidential Studies Quarterly, Quarterly Journal of Political Science*, and Upjohn Press. He has served in an advisory capacity for several community, economic, and workforce development nonprofits. He holds a PhD from Princeton University, an MPP from the Harris School at the University of Chicago, and a BA from the University of Florida.

Acknowledgments: The author would like to thank Ann Carpenter, Karen Leone de Nie, William Mabe, John Robertson, and Nisha Sutaria. The views expressed here are the author's and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the author's responsibility.

Comments and suggestions to the author are welcome at alexander.ruder@atl.frb.org.

Introduction

The workforce development system seeks to provide businesses with a skilled workforce and help individuals acquire in-demand skills that allow them to obtain stable jobs that offer pathways for career advancement. In response to the increasingly rapid adoption of new technologies, the workforce system also is adapting to a labor market in which workers must periodically reenter training to acquire new skills. To accomplish these goals, workforce development policymakers need to know what skills are in demand, how many regional jobs are available, and what programs efficiently and effectively prepare individuals for these jobs.

The need to identify effective programs leads many professionals in the workforce development field to advocate for the use of evidence-based policymaking and to disseminate information about "what works" (Hendra and Hamilton 2015; Nightengale and Eyster 2018).[1] A growing collection of rigorous randomized controlled trials (RCTs) provides a catalog of effective programs. Policymakers, however, face numerous challenges when seeking to adopt these evidence-based programs. Some RCTs are implemented in very specific geographic and institutional contexts, such as at a single training provider in one urban area, while other RCTs include multiple contexts and are designed to produce results that generalize to a broader population. In workforce development, the relevant question is often not "what works" in general or somewhere else but "what works here and at scale sufficient to meet our region's economic development goals."[2]

A growing literature discusses the challenges to scaling up programs evaluated with RCTs (Banerjee et al. 2017; Kowalski 2018; Muralidharan and Niehaus 2017; Williams 2018).[3] Several features of RCT design or implementation such as context dependence, randomization/site selection bias, and piloting/implementation bias affect how well the results from one RCT study scale to larger populations or to different geographic, socioeconomic, or institutional contexts. These challenges are unfortunately common in RCTs. In a sobering statement, Deaton and Cartwright (2016) consider the pervasive failure of trial results to replicate at scale as "likely to be the rule rather than the exception." This literature includes recommendations for researchers who seek to generalize RCT results (such as meta-analysis) and for policymakers who want to scale up effective programs (like mechanism mapping).

[1] The demand for evidence-based policymaking spans numerous areas of federal government policy, as demonstrated by the Foundations for Evidence-Based Policymaking Act of 2018 (H.R. 4174).

[2] I use the term "region" in a generic sense to refer to any U.S. geographic area (state, county, or metropolitan area) that is the focus of a workforce development program.

[3] Researchers of prekindergarten programs have extensively focused on the challenge of scaling up successful program models that were implemented at a small scale and at a high per-student expense (Bartik and Hershbein 2018; Christina and Nicholson-Goodman 2005; Lipsey, Farran, and Durkin 2018).

Several organizations provide databases of successful programs, but there is limited information on what this body of evidence tells us about what is scalable.[4] To begin to address this need, I review a sample of recent experimental studies in workforce development policy and ask to what extent does this collection of "what works" inform what is scalable to larger populations and outside of the unique context of the original study. In other words, how well does the body of evidence inform us about what works at scale? My review shows that most programs are implemented at a small scale, use nonrandom samples from a population of interest, and are concentrated in several large urban areas and U.S. states. Despite a growing literature that uses rigorous randomized control trials, many workforce development evaluations have study designs that limit application of their findings to other contexts.

My review is not a critique of these RCT studies for not producing results that scale. Many of the studies I reviewed feature rigorous designs, are well executed, and provide internally valid estimates of program impacts. These studies provide critical information about what works and under what conditions. In addition, many of the studies explicitly discuss the limitations to the generalizability of their results or speculate how the results generalize to other contexts. The purpose of the review is to summarize how informative this research is about scalable programs.

This paper proposes a framework to help regional workforce development policymakers apply evidence-based programs in their unique contexts and at scale. I review two analysis methods that aim to check assumptions and estimate how robust the program is to violations in those assumptions. The first method, mechanism mapping, is used systematically to identify ways an evidence-based program may fail in a new context. The second method, sensitivity analysis, is commonly used in cost-benefit analysis to estimate the robustness of projected outcomes (such as employment rates) to changes in key assumptions. Both methods help policymakers adopt evidence-based programs with less uncertainty about their success in a new context. It also allows policymakers to communicate that uncertainty to other stakeholders in order to encourage more informed decision making about program choice and implementation.

Background

A focus on evidence-based policymaking has encouraged significant investment in program evaluations to measure causal effects. To identify causal effects, researchers choose evaluation designs with high internal validity, such as randomized controlled trials (RCTs). By internal validity, I mean that the measured impact of a program is attributable to the program itself and not a spurious cause, such as a biased sample of individuals who enrolled in the program. RCTs reduce threats to internal validity because the random assignment mechanism guarantees that individuals enrolled in a program under study (like a treatment group) are similar, on average, in observed and unobserved factors to those individuals in the comparison group (that is, a control group).

[4] For example, see the Clearinghouse for Labor Evaluation and Research (clear.dol.gov/), the What Works Clearinghouse (ies.ed.gov/ncee/wwc/), and the Employment Strategies for Low-Income Adults Evidence Review (https://employmentstrategies.acf.hhs.gov/). All three sites accessed March 27, 2019.

A study with internal validity shows whether or not a program is effective under the specific circumstances of the study; however, a study with high internal validity alone does not guarantee that the results generalize to a context other than the original study sample. This concern is referred to as external validity.

Banerjee et al. (2017) discuss several features of RCTs that may limit external validity: context dependence, randomization/site selection bias, and piloting/implementation bias.[5] Context dependence means that the program's impact depends on social, economic, geographic, and political characteristics that vary across locations. Promising program results obtained in one location may change when an organization implements the program in another location. To reduce context dependence, Banerjee et al. (2017) recommend that researchers replicate the RCT in different contexts together with systematic reviews and meta-analysis.[6]

Randomization or site selection bias occurs when the study sites or individuals that agree to participate in RCTs are not representative of the population of interest. Heckman and Smith (1995) discuss how some evaluations modify participation criteria or recruitment efforts in order to recruit a sufficient number of participants. This modified strategy results in a study sample that likely differs from the population that would enroll in the program under normal operational circumstances (that is, absent the study). Sianesi (2014) examines randomization bias in the large-scale Employment Retention and Advancement study in the United Kingdom. She documents two ways that randomization itself changed the characteristics of program participants relative to participation in absence of the program. First, some eligible individuals refused to participate in the random assignment. Second, caseworkers are rewarded by moving individuals into work. They thus were incentivized to divert more employable individuals away from the random assignment pool and keep these more-employable individuals for their own caseloads. She finds that up to 30 percent of eligible individuals are not included in the study sample.[7]

Heckman and Smith (1995) and Allcott (2015) offer two examples of site selection bias. Heckman and Smith (1995) review the experimental evaluation of various job training programs funded by the Job Training Partnership Act (JTPA). A key implementation challenge was that training center participation in the evaluation was voluntary rather than mandatory. A national search for training center sites to participate—with extensive financial resources to incentivize participation—resulted in over a 90 percent training center refusal rate and a nonrandom sample of only 16 sites (Heckman and

[5] Market equilibrium effects and political reactions are two additional challenges to scalability of programs. Both are important concerns, but I do not include them in the review.

[6] For example, Greenberg, Michalopoulos, and Robins (2003) conduct a meta-analysis of 31 evaluations of 15 government-funded job training programs and find that effects are largest for women, smaller for men, and negligible for youth. They also find that the effects of training programs for any group are rarely large, raising earnings for the typical trainee by less than $2,000 per year.

[7] See Olsen, Bell, and Nichols (2018), who propose a method to reduce a particular type of randomization bias called applicant inclusion bias.

Smith 1995). Allcott (2015) studies the multisite evaluation of the Opower energy conservation program, which seeks to reduce energy consumption by mailing detailed reports about power usage to residential energy customers. Alcott finds that the program was significantly more effective in the first 10 utilities to participate than in the next 101 utilities to participate. He argues that the utilities that first adopted the program were different from late adopters in characteristics related to likelihood of program effectiveness, such as having an environmentally-conscious customer base.

Piloting bias or implementation bias means that program effects may differ when implemented at a larger scale than the original pilot study. Small nonprofits may operate a pilot study while a larger governmental agency may operate a program at scale. These larger organizations may differ in their capacity to implement the program effectively. Vivalt (2016), for example, finds that RCTs run by government organizations typically have smaller effect sizes than similar RCTs conducted by nonprofits and researchers.

Recent evaluations provide some examples of these threats to external validity. Hendra et al. (2016) report on a rigorous RCT evaluation of a workforce development program model implemented by four providers in three different U.S. states. Each provider undertakes a variety of training and placement services that aim to prepare participants for jobs in a targeted industry sector. In their discussion of the results, the authors note several features of the program that may limit its capacity to scale. For example, the expertise and effort required in program implementation may suggest piloting bias:

> [The Program] was a multicomponent program that required all providers to go beyond their normal operations. All providers received extensive technical assistance during the study period, to ensure that research procedures were being followed and to ensure that the providers were delivering the strongest program possible, according to the way it was designed. [The evaluator] and several consultants provided technical assistance (Hendra et al. 2016, page 16).

The authors also discuss the nonrandom site selection process, which may lead to site selection bias. The following quote illustrates several features of the selected programs directly related to the site's capacity and expertise to carry out the evaluation. To what extent will the program scale when the new implementing organizations do not have these same features?

> A primary factor in selection decisions was whether a provider could demonstrate that it was currently, or had the capability to be, firmly grounded in a targeted sector; this included in-depth knowledge of and strong relationships with employers who provided letters of support. Applicants had to demonstrate current or potential capacity to operate at the intended scale, to carry out an advancement-focused approach, and to work with a range of lower-income individuals—rather than only those who would be easiest to place in jobs. Additional selection criteria included overall organizational capabilities (including appropriate fiscal and data management capacity and the ability to comply with federal funding requirements), clear commitment to the program model, and a willingness and ability to participate in a random assignment study and to help raise matching local funds (Hendra et al. 2016, page 13).

Another example of site selection bias is in the Quantum Opportunity Program Demonstration (QOP), a program for at-risk youth that sought to increase high school graduation rates and enrollment in

postsecondary education and training (Shirm et al. 2003). Seven community-based organizations located in Ohio, Texas, Tennessee, and Washington, DC, operated the QOP demonstration programs. These organizations coordinated the four primary components of the QOP model: case management and mentoring, education, developmental activities, and community service. In the evaluation report, the authors explicitly state how site selection bias limits the external validity of their findings:

> Once we have obtained impact estimates, we face the question of whether we can generalize our findings beyond the seven CBOs [community-based organizations] in the QOP demonstration, that is, are the findings "externally valid." The answer is no. The CBOs in the demonstration were not selected by using any type of probability sampling (Shirm et al. 2003, page 19).

Finally, the 2016 early impact study of the Workforce Investment Act Gold Standard Evaluation provides an example of context dependence on local or national economic conditions (McConnell et al. 2016). The Workforce Investment Act Gold Standard Evaluation is a national study of the effectiveness of Workforce Investment Act (WIA)-funded job training and staff assisted employer services. Twenty-eight workforce investment areas recruited and randomly assigned participants into various study groups between November 2011 and spring 2013:

> The weak economy of the study period may influence the effectiveness of the WIA Adult and Dislocated Worker programs. The evaluation occurred as the nation was emerging from the major recession . . . (McConnell et al. 2016, page 25).

In workforce development, programs often rely on a network of partners to implement a program. This leads to a special form of context dependence called "embeddedness." A pilot program is more likely to be more embedded in a local context the more local organizations it depends on for implementation. These partner organizations are unlikely to be found exactly the same in a new context, and of course policymakers cannot transport a partner organization like they transport a program model. Thus, when policymakers transport an embedded program to a new location they have to re-create or develop a network of providers that is similar to the network in the original pilot study.

Recent evaluations provide many examples of program embeddedness. For example, the Valley Initiative for Development and Advancement (VIDA) is an education and training program for low-income adults in the Texas Lower Rio Grande Valley. A recent evaluation study describes how the program relies on several regional partners for funding, training, and supportive services (Rolston, Copson, and Gardiner 2017): a nonprofit organization operates VIDA and partners with economic development corporations, cities and counties, workforce development boards and American Job Centers, three community colleges, and two universities (Rolston, Copson, and Gardiner 2017, page 31). An organization seeking to implement VIDA in another region would have to find or develop a similar network of partners with comparable expertise and capacity.

Workforce Development and Scale

In the workforce development context, discussions of scale typically involve advice for scaling up specific strategies. For example, King and Prince (2015) review challenges to scale up sectoral and career

pathways strategies.[8] They note difficulties such as a tendency for policymakers to stick with "business as usual," restrictive funding sources that impede innovation, the need for employer and other stakeholder buy-in, and broader state policy contexts. Hendra and Hamilton (2015) consider a program model called WorkAdvance. The WorkAdvance model prepares individuals for employment in specific industry sectors, such as information technology, with five program components: intensive screening of applicants, preemployment and career readiness services, occupational skills training, job development and placement, and retention and advancement services. Since providers must provide all these complex services, Hendra and Hamilton (2015) argue that the model may be difficult to replicate in new locations and other organizations. For instance, sector strategies are tailored to a local labor market; these labor market conditions may not be the same in another location, so results also may differ.

Some authors consider more general approaches to scale successful programs. Berman (2015) describes the New York City Center for Economic Opportunity's (CEO) approach to scale up successful city initiatives. CEO pilots initiatives with a mix of public and private funding, rigorously evaluates these initiatives to determine their effectiveness, and then continues successful programs with a goal to bring them to scale. Unsuccessful programs are discontinued. One way that CEO brings a program to scale is by integrating it into the city's larger workforce development system rather than keeping the program as an independent entity. The integration helps CEO leverage the finances, staff, and service delivery of the workforce development system.

External Validity and Scale

External validity has a close connection to the concept of program scale-up. Williams (2018) illustrates this connection in two ways. In one sense, scale-up means expanding a successful program to a larger population. The external validity concern is how will the initial program results generalize to this larger population. In a second sense, scale-up refers to the transport of a successful program from one context to another. For program transport, the external validity concern is how the results obtained in one context generalize to a new and different context.

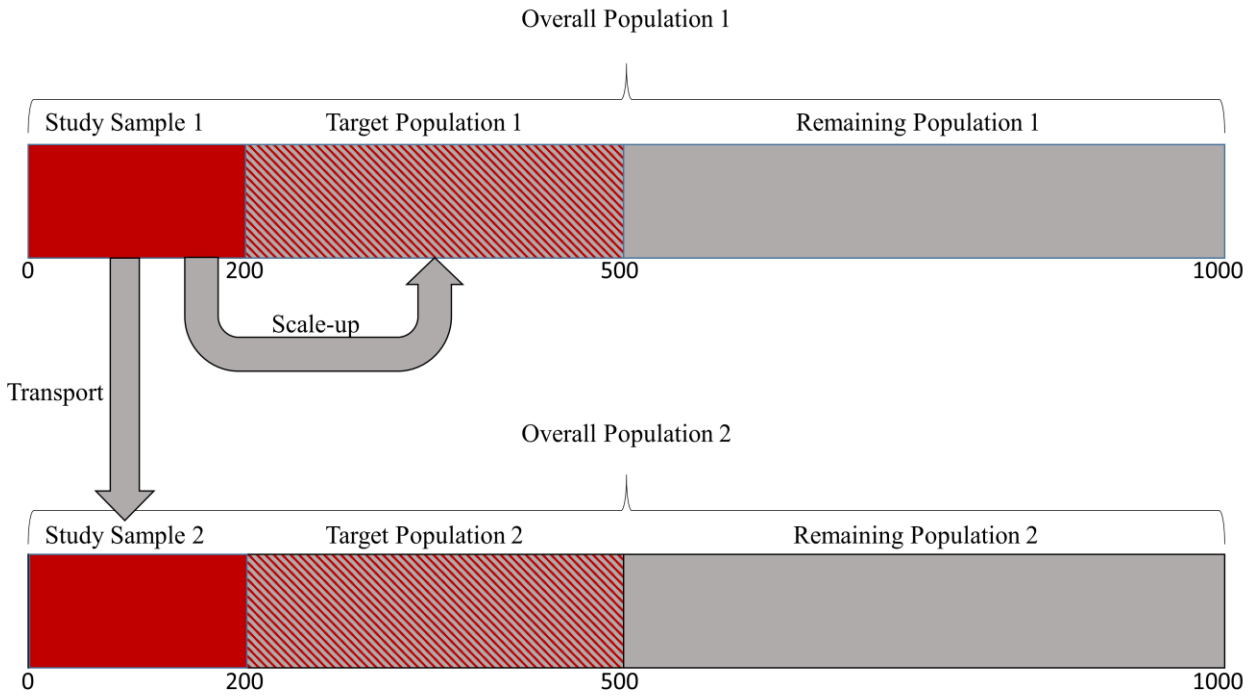
Figure 1 illustrates these two ways to represent the scale-up of a program. The initial study sample (Study Sample 1 and 2) is a subset of the target population. The study sample is a random or nonrandom sample from the target population (Target Population 1 and 2); researchers intend to use the study sample to generalize the pilot results to the target population and expand a successful pilot study to the entire target population. The study sample could be a small pilot study, while the target population could be all low-income, unemployed adults in a U.S. state.

The upper panel of the figure represents program scale-up, or expanding the program to the full target population. The grey curved arrow indicates that the goal is to apply the results from the study

[8] Sector strategies are typically employer-driven initiatives that provide training for a set of skills common to employers in a specific industry sector. Career pathways, which are typically occupationally focused, create sequential steps of education and employment to facilitate individual career advancement. King and Prince (2015) discuss each in detail and describe their integration.

sample to the entire target population. The lower panel of figure 1 represents program transport. The grey arrow pointing straight downward indicates that the goal is to apply the results from Study Sample 1, which was conducted with a sample from another target population, to a sample from a different target population. Researchers then intend to generalize the results from Study Sample 2 to the larger population in Target Population 2.

Figure 1: Program Scale-Up and Program Transport



Notes: Two hypothetical scale scenarios: program scale-up (upper bar) and program transport (lower bar). Numeric values indicate number of individuals out of the total population of 1,000 individuals. Shaded arrows point to the target population to which the study sample results intend to apply under the scale strategy.

Source: the author, based on Williams (2018)

Review: Scale of Recent Workforce Development RCTs

For policymakers seeking to apply examples of successful programs, the relevant question is not just did the program work, but how will the program work in contexts—institutional, socioeconomic, and economic—different from that of the original study? In this section, I conduct a descriptive analysis of recent RCTs of workforce development programs. I adapt the methodology of Muralidharan and Niehaus (2017) by using several metrics to summarize the scale of recent RCTs: 1) the size of the population that the sample represents; 2) the number of units treated; 3) the selection of sites and individuals into the training program; and 4) the geography of program implementation.

I draw a sample of RCTs from the U.S. Department of Labor's (DOL) Clearinghouse for Labor Evaluation and Research (CLEAR).[9] While there are other sources to find workforce-related RCTs, CLEAR is intended to be a resource for evidence specifically on workforce development interventions and promotes the use of evidence for decision making and policy development.[10] CLEAR collects evaluations of interest to the DOL and rates each according to its strength of casual evidence; that is, the rating protocol focuses on the study's internal validity, not the external validity.

The sample in this paper includes all RCTs studies in CLEAR published between 2008 and 2018. I exclude 21 studies that fail to meet any of the following criteria: 1) program has a focus on education, training, and employment for low- or middle-skill jobs; 2) program serves those with barriers to employment; 3) evaluation was published in a peer review journal or as a final evaluation report by an evaluator; and 4) program is an evaluation of an intervention that is a candidate for scale-up.[11] The sample is not intended to be exhaustive of recent RCTs in workforce development, but to assess a sample of prominent studies on a federally funded resource for evidence-based policymaking in workforce development.

The data are organized at the evaluation-site level. I define an "evaluation" as an RCT of a workforce development program model. I define a "site" as an RCT conducted at a specific geographic location in which the evaluators seek to measure impact results *within* that location rather than aggregating results *across* locations to measure an overall program impact. I use the evaluation-site level because some evaluations include multiple sites and study the impact independently at each site. In the review, I find that many evaluations have multiple sites because they are testing different program models within a single evaluation, rather than a single program model across different sites.

In contrast, if the evaluation is testing a single program model across sites, I treat this as a single evaluation. In my analysis, this type of program counts as one evaluation site. Nevertheless, I record the geographic location of the different sites within an evaluation since one of the measures (described below) is the geographic location of the study sites. The final sample includes 56 independent evaluations with 99 evaluation sites.

Random Sampling from Target Population

When the study sample is randomly drawn from a larger population, generalizing the results of a study is relatively straightforward: the program effect estimates in the sample will be unbiased

[9] See clear.dol.gov/, accessed October 5, 2018.

[10] Two ongoing or recently completed RCT evaluations are not yet listed on CLEAR but are listed on other areas of the Department of Labor's Chief Evaluation Office webpage.

[11] I follow Muralidharan and Niehaus (2017) and omit studies that seek to test theoretical mechanisms. I also only review RCT studies since the policymaking community considers these designs the most rigorous way to determine program effectiveness.

estimates for the program effects in the target population. Still, as labor economists have long recognized, samples are infrequently randomly drawn from larger populations. Program sites and individuals often voluntarily choose to participate in RCTs.

Few studies in my review include random sampling from a larger population. I find that 11 out of 99 studies involve random sampling of either program sites or individuals from a larger population. Furthermore, 10 of these 11 studies with random sampling still require either the sites or the individuals to consent to participate in the study. Selection bias thus can still occur if the sites or participants that choose to participate are different in observed or unobserved ways than the sites or participants in the original pre-consent random sample.

Scale of Population Represented

To scale up a program from a pilot study to a larger population, possibly a U.S. state or the entire United States, policymakers would want to find evidence that the program's RCT results generalize to such a large population. For instance, was the evaluation sample of individuals randomly drawn from the state's population, or was it drawn from a smaller convenience sample of willing individuals at a few select job training centers? The size of the population represented informs policymakers about how large of a population the results generalize to; study results that only generalize to relatively small populations provide limited evidence of programs that will work at a large scale.

Muralidharan and Niehaus (2017) use two methods to code this variable, depending on how the evaluation selects participants from the target population. When evaluations report that the sample is randomly drawn from a larger target population, Muralidharan and Niehaus (2017) code the size of the target population. This method is appropriate for random samples without attrition or opt-out of the study, as the results from the randomly drawn sample generalize to the target population. However, for nonrandom samples, the sample does not offer an unbiased representation of the target population. Rather, the results of the evaluation only generalize to the sample itself. As such, Muralidharan and Niehaus (2017) report the size of the target population whenever available; for other cases, they report the sample size alone.

I follow Muralidharan and Niehaus (2017) and report for nonrandom samples the size of the sample as the measure of the scale of the population represented. For one study in my sample that randomly drew subjects from a larger target population and did not require consent to participate, I report the size of the target population given in the evaluation.[12] For all others, I define the sample as the total number of participants in the study. Even for the 10 of the 11 studies that randomly sample sites or participants from a larger target population, I still report the sample size and not the size of the target population from which the random sample is drawn. In these 10 studies, I am unable to determine the precise size of the target population; in addition, these 10 studies require site or

[12] This study evaluated the impact of an offer to accept a disability benefits loss offset on various employment outcomes.

participant consent, which introduces the possibility of randomization bias such that results may not generalize to any larger population.

For example, the Workforce Investment Act Gold Standard Evaluation randomly selected 30 out of 487 local workforce areas to participate in the study (McConnell et al. 2016). Ultimately, 26 of 30 local areas agreed to participate, and the evaluator replaced two of the four that declined to participate. Furthermore, individuals at each site still had to consent to participate in the study. Do the evaluation results generalize to the individuals at the 487 local areas in the target population? Given the possibility of randomization bias, I cautiously code this evaluation as generalizing only to the sample of individuals at the 28 local workforce areas, not to the target population.[13] I recognize that to the extent these results generalize to a larger population, my analysis in Table 1 understates the size of the population represented for these 10 studies.

Table 1 shows descriptive statistics for the median, maximum, and minimum sample size for all studies by CLEAR topic area.[14] Across all studies, the median sample size is 1,155. For comparison, Muralidharan and Niehaus (2017) report a median size of 10,885 for non-U.S. development economics studies. Table 1 also shows considerable variation across topic areas, with median sample sizes ranging from 57 for Job Search Assistance, Opportunities for Youth and 15,821 for Reemployment. The sample size ranges from over 6.5 million individuals to as small as 50 individuals.

[13] Small samples, even if drawn at random from a larger target population, can produce noisy estimates of the program impact in the target population. For example, evaluating a program on a random sample of 10 American Job Centers from the population of American Job Centers produces an unbiased estimate of the program impact; however, the same evaluation conducted on a new sample of 10 different American Job Centers may produce significantly larger or smaller impact estimates.

[14] The CLEAR website assigns the topic area to each study. For the studies not published on CLEAR, I assign a CLEAR topic area based on the study's population of interest. For details about topic areas assignment, see CLEAR Policies and Procedures, Version 3.1, June 1, 2016, available at clear.dol.gov, accessed March 26, 2019.

Table 1: Scale of Recent Workforce Development Evaluation Samples

Topic Area	N Studies	Median Sample Size	Maximum Sample Size	Minimum Sample Size
All	99	1,155	6,526,888	50
Low-Income Adults	41	1,217	35,665	58
Disability Employment Policy	22	822	6,526,888	50
Opportunities for Youth	10	1,838	264,075	195
Community College	13	898	13,555	369
Reemployment	7	15,821	80,531	1,935
Career Academies	1	1,764	1,764	1,764
Entrepreneurship and Self-Employment	1	4,197	4,197	4,197
Behavioral Insights	1	747	747	747
Job Search Assistance	1	102	102	102
Job Search Assistance, Opportunities for Youth	1	57	57	57
Women in Science, Technology, Engineering, & Math (STEM)	1	252	252	252

Note: Table 1 shows median, maximum, and minimum evaluation sample size for all studies and by topic area. Study N indicates the number of evaluations reviewed by topic area.

Source: author's calculations of literature review sample

Scale of Program Treatment

I next report the scale at which the program delivers the main intervention, which I follow the literature on casual impact analysis and call the "treatment." There are at least two reasons why results obtained with small treatment groups may not generalize to larger populations. First, smaller pilot studies are often run by expert nonprofits with motivated staff or with the aid of significant technical assistance. Thus, smaller studies may feature more effective program monitoring and treatment provision than large programs run by other organizations. A governmental organization may not as effectively or efficiently provide the treatment, so results at scale may differ from the smaller-scale pilot.

Second, scaling up a small program may require subtle changes to treatment delivery that affect program effectiveness. For example, at a small scale, a career counselor may have only 150 clients to assist with job training, supportive services, and placement. At a larger scale, and in order to control costs, the program may have to increase the counselor caseload to 1,500, which means less one-on-one time with each client.

I record the number of individuals in each study who received the primary treatment.[15] Table 2 shows descriptive statistics for the median number of individuals who receive the main program treatment overall and by CLEAR topic area. The median group size treated across all studies is 585. For comparison, Muralidharan and Niehaus (2017) report 5,340 as the median number treated. The scale of the treatment provision varies across topic areas, from 33 in Job Search Assistance, Opportunities for Youth to 2,553 in Reemployment. The range of treatment group size is from 25 in Disability Employment Policy to over 100,000 in Opportunities for Youth.

Table 2: Scale of Recent Workforce Development Evaluation Treatment Groups

Topic Area	N Studies	Median Treatment Group Size	Maximum Treatment Group Size	Minimum Treatment Group Size
All	99	585	116,919	25
Low-Income Adults	41	626	31,304	29
Disability Employment Policy	22	420	77,115	25
Opportunities for Youth	10	744	116,919	97
Community College	13	505	8,049	224
Reemployment	7	2,553	38,600	560
Career Academies	1	959	959	959
Entrepreneurship and Self-Employment	1	2,094	2,094	2,094
Behavioral Insights	1	372	372	372
Job Search Assistance	1	53	53	53
Job Search Assistance, Opportunities for Youth	1	33	33	33
Women in Science, Technology, Engineering, & Math (STEM)	1	145	145	145

Note: Table 2 shows median, maximum, and minimum treatment group size for all studies and by topic area. Study N indicates the number of evaluations reviewed by topic area.

Source: author's calculations of literature review sample

Context Dependence

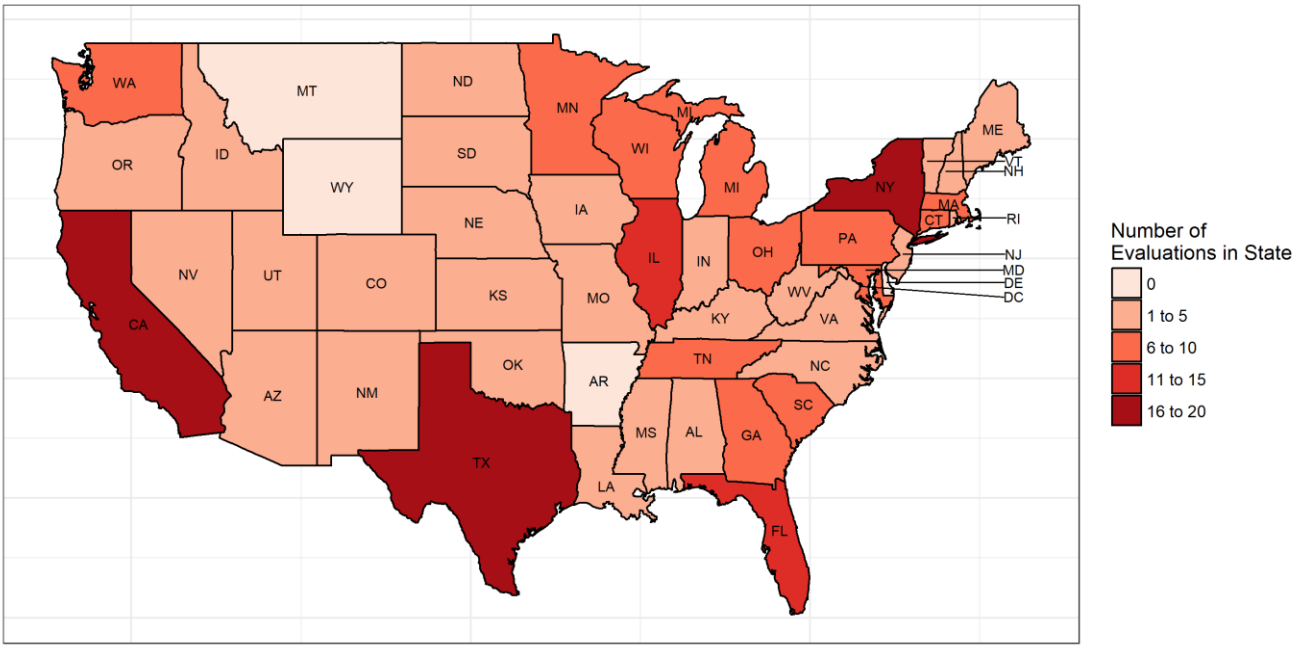
An additional barrier to external validity is context dependence associated with the geographic location in which a given study is conducted. As noted by Banerjee et al. (2017), observed or unobserved

[15] I define "primary" as the treatment of interest. In the case of more than two treatment conditions I code the number of units in the most intensive treatment, reasoning that the more intensive treatment would be more challenging to scale up.

characteristics of a specific location may influence study results. When a program is transported to another location than the pilot study (such as a neighborhood, city, state, or country), these location-specific contextual factors may affect program performance at the new site compared with the original pilot site. To capture this geographic context dependence, I report the geographic location of an evaluation site by state and county of implementation.[16] For every evaluation site, I code the study's geographic location as defined in the published evaluation. The location is defined as the geographic unit(s) in which the workforce development program (the "treatment") is delivered to pilot study participants. The location can be a single city, county, or state, but it can also be a combination of cities, counties, and states.

Figure 2 shows the total number of evaluations conducted in each state.[17] No evaluations were conducted in Arkansas, Montana, or Wyoming. The most evaluations were conducted in three of the most populous states: California, New York, and Texas. Each of these states has between 16 and 20 evaluations. States in parts of the Midwest, the Mountain West, and the Southeast have relatively few evaluations. The majority of states have between one and five evaluations.

Figure 2: Workforce Development Evaluations by U.S. State



Source: author's calculations of literature review sample

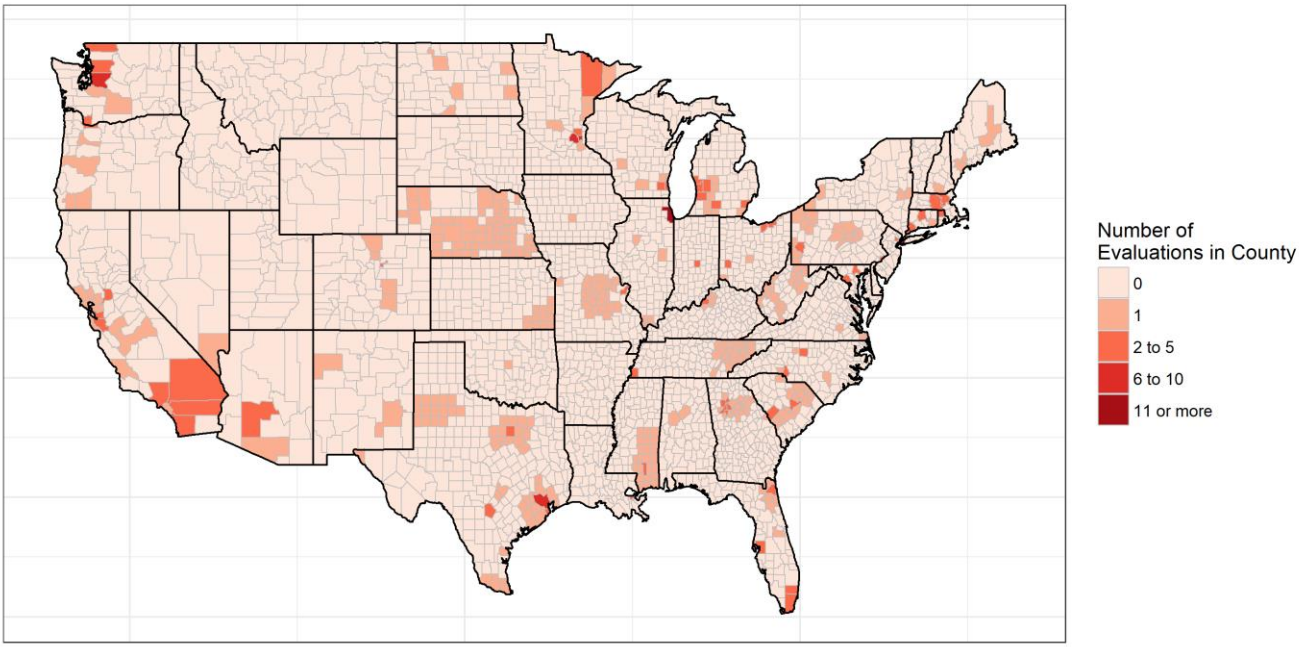
[16] Many location-specific factors influence program performance. The use of geographic location to measure local context is a simplification. However, this measure likely captures relatively fixed characteristics of a location such as political institutions, infrastructure, networks of workforce organizations, and certain socioeconomic variables.

[17] One study was conducted in Honolulu, Hawaii.

Variation at the county level reveals more detail about the geographic distribution of evaluations. Figure 3 shows U.S. counties by the number of evaluations. The figure reveals that considerable variation exists within states and the majority of counties have had zero evaluations. In the states with the most evaluation sites—California, New York, and Texas—evaluations are likely to be in counties surrounding major metropolitan areas such as Los Angeles, New York City, and Houston. Overall, counties with the most evaluations are surrounding New York City, Chicago, Houston, Minneapolis/St. Paul, and Seattle.

Evaluations have also been conducted in certain large, mostly rural multicounty regions; these groups of counties typically represent a single evaluation rather than several separate evaluations in neighboring counties. For example, nearly every county in Nebraska is highlighted, but this is because one study was implemented at a near statewide level. Similarly, in West Virginia, Mississippi, South Carolina, and other states, individual studies include several mostly rural counties in order to obtain a sufficient sample size.[18]

Figure 3: Workforce Development Evaluation Sites by U.S. County



Source: author's calculations of literature review sample

While not a comprehensive analysis of all workforce development RCTs, this review shows that many recent evaluations have design or implementation features that limit the scalability of results to contexts outside of the original study. At the same time, some evaluations in the sample adopt methods to increase generalizability, such as conducting the evaluations in multiple sites that are intentionally

[18] Since the focus on this map is within-state variation, the figure omits 16 studies conducted at the state or national level.

selected to be more representative of the target population. Whether the evaluation was conducted at a single pilot site or across numerous sites, policymakers who seek to adopt successful programs must still consider how the results will scale to their own unique context. The following section offers a framework to guide policymakers and program administrators on adopting or expanding programs that are successful in other contexts.

Scale Analysis

In this section, I briefly review a sequential method to help practitioners scale up successful workforce strategies. I then apply the method using an example from a nationally recognized workforce development program that targets youth with barriers to employment.

Two Steps of Scale Analysis

In workforce development strategic planning, one of the most difficult steps involves going from broad goals to specific programs with numeric targets for success. For example, city leaders in a major metropolitan area commission a labor market analysis and find a skills gap in the information technology sector. The report finds a labor supply shortfall of 1,000 information technology workers. How does the group of city leaders choose a combination of programs that can meet this goal? How do they convince community stakeholders and potential funders (such as elected officials, philanthropy, and nonprofits) that the proposal has the potential to succeed?

Evidence-based policymaking presumes that policymakers adopt or expand programs that are based on rigorous evidence of success. However, because programs have worked in one place or time does not guarantee they will work in another. This is the fundamental problem of expanding an evidence-based program: evidence suggests programs work in a given context, but policymakers need them to work in a new context in which evidence of success does not yet exist (Williams 2018).

Scale analysis is a method to help policymakers conduct background research, list assumptions, and test the sensitivity of projected employment outcomes to violations in the core assumptions. The analysis helps policymakers identify weaknesses and strengths in their efforts to develop programs to meet specific program goals and objectives.

Scale analysis begins after policymakers choose specific evidence-based training programs as candidates for local adoption. It then proceeds in two sequential steps: mechanism mapping and sensitivity analysis.

1. Mechanism mapping

Mechanism mapping was developed by Williams (2018) as a method to check assumptions when scaling up a program. Pilot programs typically follow a theory of change, or a sequence of steps that links program inputs (like instructional faculty) to program outputs (such as graduate employment). For policymakers seeking to scale up a program, the relevant question is how the theory of change applies in the new context. Specifically, what assumptions were required for the pilot study to succeed, and are those assumptions likely to hold in the new context?

In brief, the process involves several simple steps, although the difficulty of the process depends on the complexity of the program's theory of change.[19] First, identify the program's theory of change. The theory of change includes the program's inputs, activities, outputs, intermediate outcomes, and final outcomes. Second, list the contextual assumptions that underlie the causal logic of the theory. What assumptions must be maintained for the program inputs to be available, activities to be performed, and outputs to be produced? What assumptions connect the outputs to the intermediate outcomes, and how do the intermediate outcomes affect the final outcomes of interest?

For example, many workforce initiatives must begin with a simple question about the target population: who and where are the potential workers we seek to train? To answer this question, workforce policymakers must investigate core assumptions about the number of potential workers available, their interest in the program, and their readiness for training and employment.

2. Sensitivity analysis

Policymakers make numerous assumptions when projecting the impact of workforce programs in a new context. These predictions are more sensitive—or robust—to changes in some assumptions than in others. Sensitivity analysis allows policymakers to assess how robust their projections are to changes in key assumption. Workforce programs typically depend on key assumptions about program recruitment, retention, graduation, and postgraduate employment rates. Policymakers should begin this step by performing sensitivity analysis on the assumptions underlying projections about these key outcomes. How, for example, does the estimated number of individuals trained and employed change under different assumptions about participant recruitment and enrollment?

Example: Information Technology Sector Strategy

I now present an example of mechanism mapping and sensitivity analysis using an information technology (IT) sector strategy. I assume that a group of policymakers in a large metropolitan area has begun a strategic planning process. They commissioned a labor market analysis, which identified a projected shortage of 500 IT workers per year.

Rather than innovate and develop a new IT program, the policymakers seek to adopt a program that has a demonstrated record of success and focuses on youth with barriers to employment. The policymakers want to transport the program to their metropolitan area. In the example, I use a well-known workforce development program that has been implemented in several major U.S. cities. A recent evaluation of the program includes the theory of change and the results from a multisite evaluation. I use these elements of the evaluation to conduct mechanism mapping and sensitivity analysis and call this program *IT Excellence* for the purposes of this example.[20]

---

[19] Williams (2018) introduces mechanism mapping as a method to improve scale-up efforts, while Gertler et al. (2016) review the importance of checking assumptions and following a theory of change in program evaluation more broadly.

[20] The actual program details, theory of change, and sample are described in Fein and Hamadyk (2018).

The policymakers decide *IT Excellence* will form one part of a strategy to meet the strategic goal of 500 workers. *IT Excellence* will be the program focused on low-income youth with barriers to employment. Given the capacity constraints of the program, the policymakers determine that it needs to produce 100 trained and employed graduates per year. The typical cohort size of *IT Excellence* is 105 participants per cohort, with two cohorts starting each year. Under the assumption that all enrollees start and complete the program and obtain employment in the IT sector, the program will produce 210 IT workers per year. However, the policymakers realize that all the students are unlikely to complete training and find employment. A recent national multisite evaluation of the program suggests that they can expect approximately 58 percent of enrolled individuals to find a job in the IT sector. Can they expect similar performance in their context?

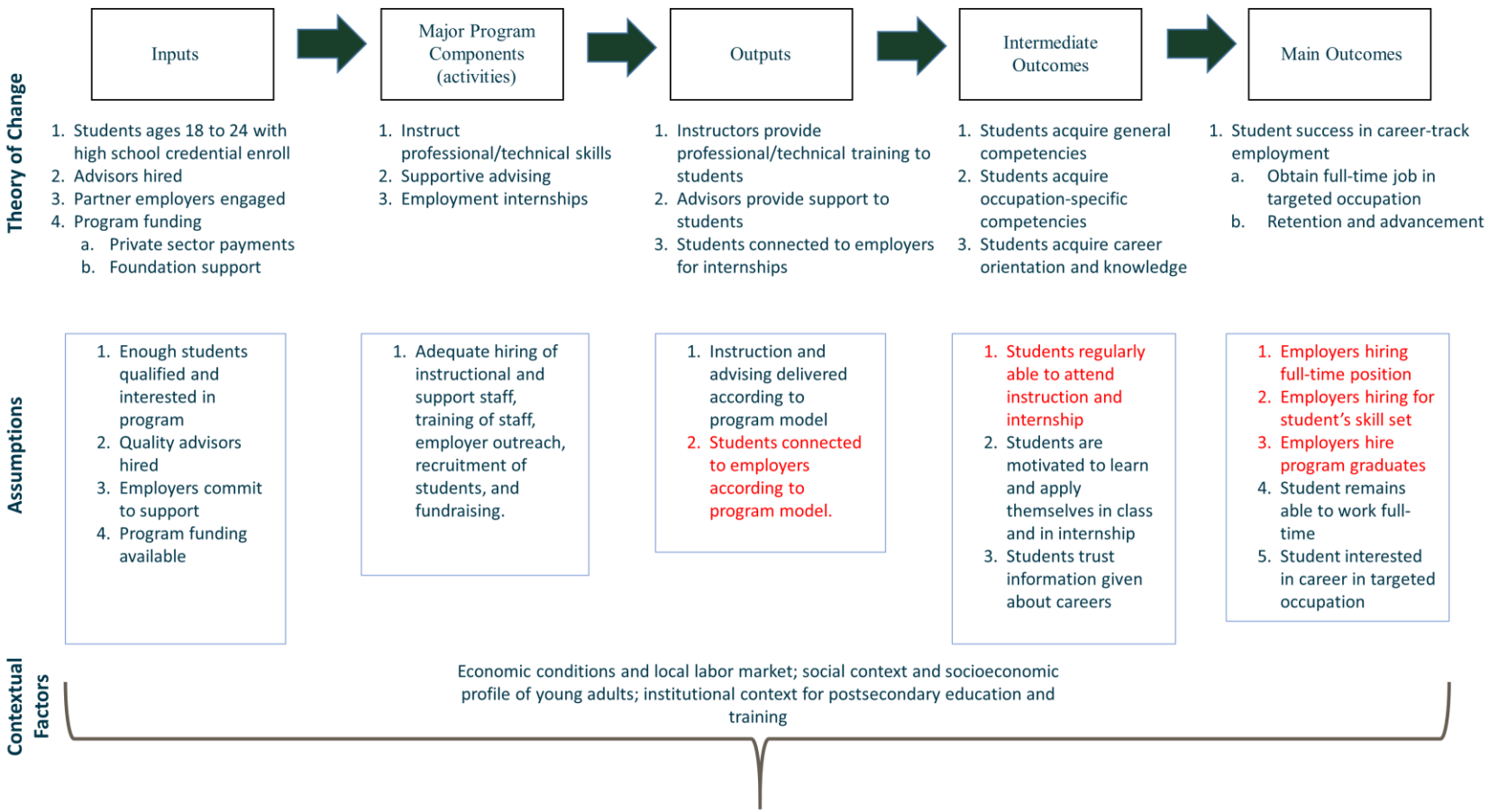
To examine how closely implementation at the proposed new site corresponds to the conditions in the original multisite evaluation, the policymakers decide to use mechanism mapping with the program's published theory of change. Then, they will use results from the multisite evaluation to measure uncertainty over how many students graduate from *IT Excellence* and find employment in the IT sector.

Figure 4 shows the program's theory of change along with key assumptions.[21] The top part of the figure presents the theory of change adapted from the published evaluation study. Beneath each component of the theory of change are some assumptions necessary for the theory to work in practice. The arrows represent the sequential steps that connect the inputs to the main program outcomes. The bottom of the figure lists broader contextual assumptions about the economy and the socioeconomic characteristics of the region.

The first two parts of the theory of change—inputs and activities—represent the implementation of the program. The inputs required for program implementation include qualified students, advisers, partner employers, and program funding. Assumptions required at this point include the availability of enough students, advisers, interested employer partners, and funders. The next step in the theory of change is the major program components, or activities. These activities include teaching the students the required technical and professional skills, supporting students with academic advising, and organizing internships. For the program to implement these activities successfully, some assumptions are necessary: the program is able to recruit qualified students and able to hire and train a sufficient number of teachers and advisers; outreach to employers results in the recruitment of partner employers; and the necessary funds are obtained.

[21] The published theory of change includes several more inputs and major program activities. For the purposes of the example, I simplify the theory of change and focus on a few select inputs and activities.

Figure 4: Theory of Change for IT Excellence Program



Note: Red text indicates the examples of violated assumptions as discussed in the main body of the paper.

Source: author's adaptation of theory of change in Fein and Hamadyk (2018)

The final three parts of the theory of change are the outputs, the intermediate outcomes, and the main outcomes. Together, these parts represent the impact of the program (Williams 2018). Outputs are the "goods and services, produced and delivered, under the control of the implementing agency" (Gertler et al. 2016, page 35). The outputs of *IT Excellence* include the academic staff teaching students, the advisers providing counseling, and the advisers connecting and enrolling students in internships with employers. Some of the key assumptions at this step are that the implementing organization provides instruction and internships according to the program model.

The intermediate and main outcomes are the measurable impacts of the program toward the goal of training 100 new IT employees. According to the theory of change, instruction causes students to acquire general professional skills and occupation-specific competencies; the advising gives students career knowledge and an orientation to the IT sector. Key assumptions at this step include students regularly attend classes and the internship, and students trust the career and labor market information given in advising.[22] The theory of change states that these intermediate outcomes lead to the main outcomes of full employment in the IT sector as well as retention and advancement in the sector. A key assumption at this stage is that IT sector employers are hiring, they consider *IT Excellence* graduates to have the sufficient skills required for those openings, and they are willing to hire the program graduates.

Violations in any of the assumptions in figure 4 could cause failures in either program implementation or impact. I highlight in red some examples of assumptions that may fail in the new context. For example, the theory of change assumes that the program has strong relationships with employers, who are used to raise funds and create internships and employment opportunities for students. In a new context, a program may lack these strong employer relationships and fail to connect students to employers according to the program model (an output), which means that students will not regularly attend an internship (intermediate outcomes). Without the experience and connections gained in an internship experience, students are less likely to find full-time employment in the IT sector.

The example in figure 4 shows the importance of employer relationships to the program's theory of change. Together with a critical examination of the other assumptions, this process helps policymakers identify core assumptions *before* program implementation. When policymakers take steps to increase the plausibility of these assumptions, the program is more likely to follow the theory of change and deliver the desired main outcomes.

A complementary way to examine program assumptions is sensitivity analysis. Sensitivity analysis, which is a method common in applications such as cost-benefit analysis, can help workforce development policymakers: 1) evaluate the sensitivity of performance projections under different contextual circumstances; and 2) communicate the credibility of performance projections to internal and external stakeholders in order to make more informed decisions and increase stakeholder commitment to the program.

[22] For example, some students may not believe that local labor market statistics provide a meaningful predictor of how much they individually will earn.

Strategic plans often give a point estimate projection, such as "We will train and employ 100 information technology workers." However, point estimates do not communicate to stakeholders the uncertainty of the estimate. For instance, under different assumptions, how large is the range of possible outcomes? Are policymakers reasonably certain that the number of trained individuals will be between 90 and 110, or is the range between 50 and 150? The latter range suggests much more uncertainty over program outcomes than the former.

Sensitivity analysis helps policymakers quantify this uncertainty by creating a range of projected outcomes under different assumptions. For example, the projection of 100 workers assumes that the unemployment rate is approximately 3 percent and the labor market is tight, which means that 100 percent of program graduates obtain employment. Sensitivity analysis asks how that projected number changes if the assumption about the unemployment rate is incorrect. How many graduates will the program place if the unemployment rate increases to 5 percent?

Policymakers can conduct sensitivity analysis by using information in the evaluations of the program that they seek to adopt. For example, an evaluation shows that at some sites the completion rate of students was 100 percent, while at others it was only 70 percent. Using sensitivity analysis, policymakers can project program performance under both scenarios, which allows them to determine how sensitive the projection of 100 workers is to variations in the completion rate.

I use the published results for *IT Excellence* to illustrate sensitivity analysis. The upper section of table 3 shows the published average study results across sites. The evaluation found that, on average across sites, 75 percent of students complete *IT Excellence* and 58 percent of completers find employment in the targeted IT sector.

Nevertheless, these results vary across the eight different sites in the evaluation. The middle section in table 3 shows the range of best and worst case results across sites. The lowest completion rate is 61 percent and the highest is 84 percent. The lowest employment rate was 49 percent and the highest was 75 percent.

Table 3: Evaluation Results and Sensitivity Analysis Parameters for IT Excellence

Average Evaluation Results across Sites	Participants (per year)	210
	Completion rate	75%
	Employment rate	58%
Best Case/Worst Case Evaluation Results across Sites	Completion rate low	61%
	Completion rate high	84%
	Employment rate low	49%
	Employment rate high	75%
Projections for New Site	Individuals enrolled	100
	Individuals employed (Base Case)	91
	Individuals employed (Worst Case)	63
	Individuals employed (Best Case)	132

Notes: The table section "Average Evaluation Results across Sites" presents the average results across all sites in the evaluation. The section "Best Case/Worst Case Evaluation Results across Sites" presents the results from the worst and best performing sites in the evaluation. The section "Projections for New Site" presents policymaker estimates for program performance in the new location for *IT Excellence*. The number of participants assumes that two cohorts of 105 individuals per year begin the program.

Source: author's calculations and analysis of published evaluation results (Fein and Hamadyk 2018)

I use these results to calculate the projected number of employed graduates under the worst- and best-case scenarios of completion and employment rates.[23] The lower section of table 3 shows the projected number of employed graduates at a new site under three different scenarios: 1) the base case in which the results correspond to the evaluation's average program completion and employment rates; 2) the worst case in which the program results correspond to the lowest site completion and employment rates; and 3) the best case in which the program results correspond to the best site completion and employment rates.

The goal of meeting the target of 100 employed persons differs across these three scenarios. If the program achieves the average results, then 91 graduates enter employment in the IT sector. However, if the program results correspond to the worst site in the evaluation, then only 63 graduates enter employment. The results under the best-case scenario—132 employed individuals—exceed the target number.

[23] Worse- and best-case analysis is described in Boardman et al. (2011).

Worst- and best-case sensitivity analysis has several limitations. First, actual program completion and employment rates in a new site can equal many plausible values. For example, a new site can have poor completion rates but average or excellent employment rates; it is often prohibitively costly to examine all combinations of plausible values. Second, the results do not measure the policymakers' uncertainty over the final outcomes: policymakers are unable, for example, to estimate the probability that the program will train and employ *at least* 100 new IT workers.

I apply a technique called Monte Carlo sensitivity analysis to address these limitations. Monte Carlo sensitivity analysis incorporates the uncertainty in results by assuming that there is a probability of achieving any result in a given range of values. A random draw from this probability distribution represents a realization of a program outcome. By drawing thousands of random draws, Monte Carlo sensitivity analysis creates a distribution of outcomes with which policymakers can calculate the probability of any outcome occurring.

In the *IT Excellence* example, I assume that any completion rate or employment outcome between the worst- and best-case scenarios is equally likely.[24] Using these values, I conduct the Monte Carlo sensitivity analysis by repeating these two steps 5,000 times:

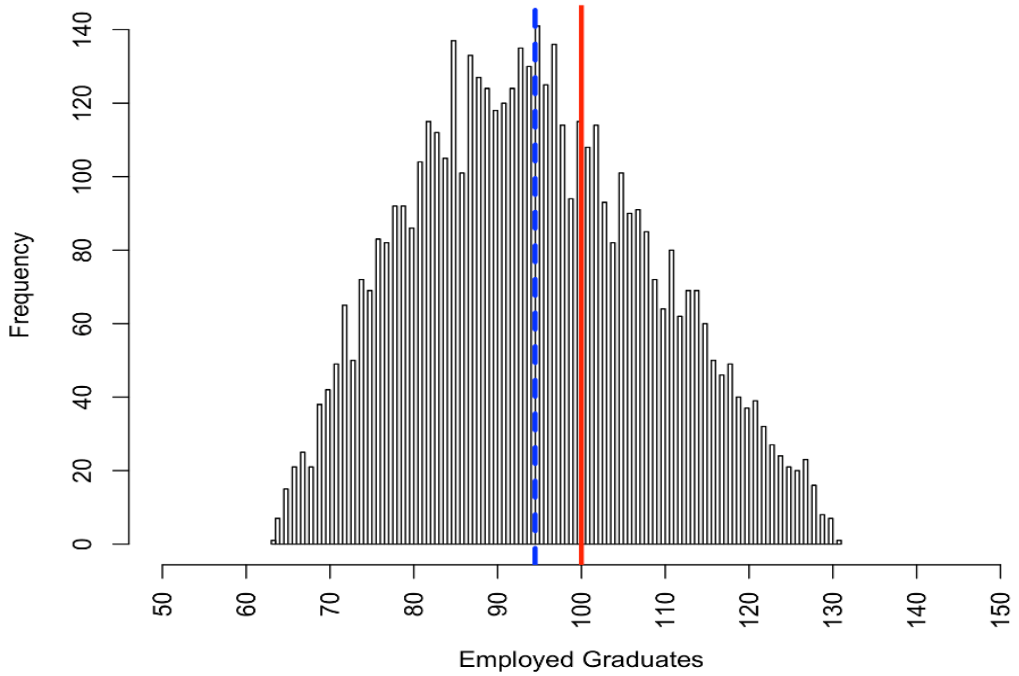
1. Randomly draw a program completion rate from between 61 percent and 84 percent. Calculate expected number of program completers by multiplying completion rate by number of enrolled participants (210).

2. Randomly draw a program employment rate from between 49 percent and 75 percent. Calculate expected number of employed graduates by multiplying employment rate by number of expected completers from Step 1.

The results are shown in figure 5, which shows the distribution of the expected number of program graduates who find employment in the IT sector within four months after graduation. Each grey bar represents the frequency that a certain number of graduates are employed out of 5,000 random draws. The vertical red line represents the target employment number of 100, and the vertical blue line shows the average of all simulated employment numbers, which is 94.

[24] Boardman et al. (2011) recommend draws from a uniform distribution when theory or empirical evidence does not suggest a particular distribution. Thus, an assumption of this analysis is that the policymakers believe all outcomes between the worst- and best-case scenarios are equally likely.

Figure 5: Monte Carlo Sensitivity Analysis Results



Source: author's calculations

The results show similarities to the worst- and best-case analysis in table 3. The range of outcomes is between 64 and 131 and the average outcome is 94. The distribution of results provides a measure of uncertainty over the outcome. It also allows policymakers to calculate the probability of any outcome occurring. To calculate the probability that the program graduates and employs more than 100 graduates, one simply divides the number of simulated outcomes above 100 by the total number of simulations. Under the assumptions of this Monte Carlo sensitivity analysis, the probability the program achieves at least its goal is 0.33. Some workforce development partners may consider a roughly 30 percent chance of success to be too risky, while others may consider program modifications or further research that can help reduce this uncertainty.

The Monte Carlo sensitivity analysis itself makes several assumptions. For example, it assumes that 210 individuals per year participate. However, policymakers may face considerable uncertainty over how many individuals they will recruit into the program, particularly during its early years. Policymakers could model this uncertainty in the sensitivity analysis by also drawing the number of program participants from a distribution. Intuitively, uncertainty over program participation will add even greater overall uncertainty to the predicted number of employed individuals in the IT sector.

Discussion

To meet ambitious workforce development goals, policymakers need to apply evidence-based practices at scale. Policymakers can choose from a growing list of successful programs, but they still face uncertainty over whether or not a program will work in a specific context. What works in one geographic location or time may fail in another.

Policymakers adopt evidence-based programs because evaluations show they are effective, which to some extent reduces uncertainty about whether or not the adopted program will achieve its goals; however, they should also adopt programs that are shown to be scalable or applicable in other contexts than that of the original evaluation. To frame this discussion about scalability, I review a sample of recent workforce development RCTs and report to what extent this body of research informs policymakers about what works at scale and what programs are transportable to other contexts. Most of the studies in my review feature nonrandom samples, small treatment groups, and they are conducted in specific geographic locations with unique political, economic, demographic, and institutional contexts.

The literature review suggests there is limited evidence for programs that work at scale and in contexts other than that of the original study. Thus, policymakers must carefully evaluate how an evidence-based program will operate at scale or in a new context. I seek to provide policymakers with a systematic way to incorporate evidence-based programs into their workforce development strategic planning process. Policymakers begin by specifying the program's theory of change. Using a method called mechanism mapping (Williams 2018), policymakers identify the assumptions required for the theory of change to work in practice. That is, they identify the assumptions that are required for the program inputs to produce the program outputs as the theory specifies. Policymakers then assess which assumptions are likely to be valid when they implement the program. Results of the mechanism mapping exercise help policymakers identify weaknesses and make adjustments to the program before implementation begins.

Policymakers also need a way to access the uncertainty over projected program outcomes. Point predictions such as "We will train and employ 100 graduates in the IT sector" provide no measure of uncertainty. Useful measures of uncertainty include the range of possible outcomes and the probability that the program will train equal to or greater than the target number. Policymakers can use measures of uncertainty to compare different programs and to better inform stakeholders about the program's expected outcomes.

The paper illustrates two ways to measure uncertainty using sensitivity analysis: worst- and best-case analysis and Monte Carlo sensitivity analysis. The example uses results from a real evaluation of a multisite employment and training program that focuses on youth with barriers to employment. The results reveal that the expected number of program graduates who find employment in the target sector is close to the policymakers' goal. However, there is considerable variation around the average, and there is a small probability that the program trains only approximately half the target number.

Of course, the methods reviewed cannot eliminate uncertainty. The complexity of workforce programs in particular creates more opportunities for unforeseen challenges and implementation failures. In addition, the recommended simulation methods themselves rely on untested assumptions, such as the number of participants per year.

The proposed framework provides guidance to policymakers at several stages of the program adoption process. Policymakers who incorporate mechanism mapping and sensitivity analysis may identify weaknesses in an overall strategic plan, suggest adjustments to program implementation plans, and ultimately allow stakeholders to make a more informed decision about the adoption of workforce development programs.

References

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130 (3): 1117–1165.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *Journal of Economic Perspectives* 31 (4): 73–102.

Bartik, Timothy J., and Brad Hershbein. 2018. "Pre-K Effectiveness at Scale." W.E. Upjohn Institute Policy Brief.

Berman, David S. 2015. "Piloting and Replicating What Works in Workforce Development: Using Performance Management and Evaluation to Identify Effective Programs." In *Transforming Workforce Development Policies for the 21st Century*, edited by Carl Van Horn, Tammy Edwards, and Todd Greene. W.E. Upjohn Institute for Employment Research.

Boardman, Anthony E., David H. Greenberg, Aidan R. Vining, and David L. Weimer. 2011. *Cost-Benefit Analysis: Concepts and Practice*. Fourth ed. Prentice Hall.

Christina, Rachel, and JoVictoria Nicholson-Goodman. 2005. "Going to Scale with High-Quality Early Education: Choices and Consequences in Universal Pre-Kindergarten Efforts." Santa Monica, CA: RAND Corporation.

Deaton, Angus, and Nancy Cartwright. 2016. "Understanding and Misunderstanding Randomized Controlled Trials." National Bureau of Economic Research Working Paper 22595.

Fein, David, and Jill Hamadyk. 2018. "Bridging the Opportunity Divide for Low-Income Youth: Implementation and Early Impacts of the Year Up Program." OPRE Report #2018-65, Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch. 2016. *Impact Evaluation in Practice*. Second ed. World Bank Group.

Greenberg, David H., Charles Michalopoulos, and Philip K. Robins. 2003. "A Meta-Analysis of Government-Sponsored Training Programs." *Industrial and Labor Relations Review* 1 (57): 31–53.

Heckman, James J., and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9 (2): 85–110.

Hendra, Richard, David H. Greenberg, Gayle Hamilton, Ari Oppenheim, Alexandra Pennington, Kelsey Schaberg, and Betsy L. Tessler. 2016. *Encouraging Evidence on a Sector-Focused Advancement Strategy: Two Year Impacts from the WorkAdvance Demonstration*. MDRC.

Hendra, Richard, and Gayle Hamilton. 2015. "Improving the Effectiveness of Education and Training Programs for Low-Income Individuals: Building Knowledge from Three Decades of Rigorous Experiments." In *Transforming Workforce Development Policies for the 21st Century*, edited by Carl Van Horn, Tammy Edwards, and Todd Greene. Kalamazoo, Michigan: W.E. Upjohn Institute for Employment Research.

King, Christopher T., and Heath J. Prince. 2015. "Moving Sectoral and Career Pathway Programs from Promise to Scale." In *Transforming Workforce Development Policies for the 21st Century*, edited by Carl Van Horn, Tammy Edwards, and Todd Greene. Kalamazoo, Michigan: W.E. Upjohn Institute for Employment Research.

Kowalski, Amanda E. 2018. "How to Examine External Validity within an Experiment." National Bureau of Economic Research Working Paper 24834.

Lipsey, Mark W., Dave C. Farran, and Kelley Durkin. 2018. "Effects of a State Prekindergarten Program on Children's Achievement and Behavior through Third Grade." *Early Childhood Research Quarterly* 4 (45): 155–176.

McConnell, Sheena, Kenneth Fortson, Dana Rotz, Peter Schochet, Paul Burkander, Linda Rosenberg, Annalisa Mastri, and Ronald D'Amico. 2016. "Providing Public Workforce Services to Job Seekers: 15-Month Impact Findings on the WIA Adult and Dislocated Worker Programs." Evaluation report submitted to U.S. Department of Labor, Employment and Training Administration, Office of Policy Development and Research. Mathematica Policy Research, Social Policy Research Associates, MDRC, and Corporation for a Skilled Workforce.

Muralidharan, Karthik, and Paul Niehaus. 2017. "Experimentation at Scale." *Journal of Economic Perspectives* 31 (4): 103–124.

Nightengale, Demetra Smith, and Lauren Eyster. 2018. "Results and Returns from Public Investments in Employment and Training." In *Investing in America's Workforce*, edited by Stuart Andreason, Todd Greene, Heath Prince, and Carl E. Van Horn. Kalamazoo, Michigan: W.E. Upjohn Institute for Employment Research.

Olsen, Robert B., Stephen H. Bell, and Austin Nichols. 2018. "Using Preferred Applicant Random Assignment (PARA) to Reduce Randomization Bias in Randomized Trials of Discretionary Programs." *Journal of Policy Analysis and Management* 37 (1): 167–80.

Rolston, Howard, Elizabeth Copson, and Karen Gardiner. 2017. "Valley Initiative for Development and Advancement: Implementation and Early Impact Report." U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation Report 2017-83.

Shirm, Allen, Nuria Rodriguez-Planas, Myles Maxfield, and Christina Tuttle. 2003. "The Quantum Opportunity Program Demonstration: Short-Term Impacts." Mathematica Policy Research, Inc, Reference 8279-093.

Sianesi, Barbara. 2014. "Evidence of Randomisation Bias in a Large-Scale Social Experiment: The Case of ERA." *Journal of Econometrics* 198: 41–64.

Vivalt, Eva. 2016. "How Much Can We Generalize from Impact Evaluations?" Working Paper.

Williams, Martin J. 2018. "External Validity and Policy Adaptation: From Impact Evaluation to Policy Design." Working Paper.