J Labour Market Res. (2013) 46:239–252
DOI 10.1007/s12651-013-0133-2

ARTICLE

Replication issues in social experiments: lessons from US labor market programs

Burt S. Barnow · David Greenberg

Published online: 5 July 2013

Abstract When evaluating a pilot or demonstration program, there are risks from drawing inferences from a single test. This paper reviews the experiences of replication efforts from demonstrations using randomized controlled trials in the initial evaluation and the replications. Although replications of promising programs are primarily gathered to increase sample size, replications are also used to learn if the intervention is successful for other target groups and geographic locations, and to vary some of the intervention's features. In many cases, replications fail to achieve the same success as the original evaluation, and the paper reviews reasons that have been suggested for such failures. The paper reviews what has been learned from replications where random assignment was used in six instances: income maintenance experiments, unemployment insurance bonus experiments, the Center for Employment Training program, job clubs, job search experiments, and the Quantum Opportunity Program. The paper concludes by summarizing lessons learned from the review and areas where more research is needed.

Keywords Evaluations · Training programs · Employment

Prepared for the Institute for Employment Research International Conference on Field Experiments in Policy Evaluation, Nuremberg, Germany October 18 and 19, 2012.

B.S. Barnow (✉)
Trachtenberg School of Public Policy and Public Administration, George Washington University, 805 21st ST NW, Washington, DC 20052, USA
e-mail: barnow@jhu.edu

D. Greenberg
Department of Economics, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA

Probleme bei sozialen Experimenten: Lehren aus US-amerikanischen Arbeitsmarktprogrammen

Zusammenfassung Bei der Bewertung eines Pilot- oder Testprogramms besteht die Gefahr, aus einem einzelnen Test Rückschlüsse zu ziehen. In dieser Arbeit werden die Erfahrungen mit Wiederholungen von Testprogrammstudien anhand einer randomisierten, kontrollierten Studie für die erstmalige Auswertung und die Wiederholungen besprochen. Auch wenn Wiederholungsstudien vielversprechender Programme primär zur Erhöhung des Stichprobenumfangs durchgeführt werden, dienen sie auch zum Sammeln von Erfahrungswerten dahingehend, ob die Intervention auch bei anderen Zielgruppen und an anderen geografischen Standorten erfolgreich ist, und um einige der Interventionsmerkmale zu variieren. In vielen Fällen sind Wiederholungsstudien nicht so erfolgreich wie die ursprüngliche Erhebung. In dieser Arbeit werden die für ein solches Fehlschlagen vorgebrachten Begründungen besprochen. Außerdem werden die Erfahrungen aus den Wiederholungsstudien unter Anwendung einer randomisierten Zuweisung in sechs Fällen dargelegt: Experimente zur Einkommenssicherung, Experimente zu Bonuszahlungen bei der Arbeitslosenversicherung, Programm des Center for Employment Training, Job-Clubs, Experimente zur Stellensuche und „Quantum Opportunity"-Programm (Programm für höhere Chancen). Zum Abschluss der Arbeit werden die Erkenntnisse aus der Besprechung zusammengefasst sowie Bereiche aufgezeigt, in denen weitere Forschung notwendig ist.

1 Introduction

When evaluating a pilot or demonstration program, there are risks from drawing policy inferences from a single test. First,

the positive impact could be due to chance, although a large enough sample size makes chance alone an unlikely explanation for a large, statistically significant positive impact estimate. Second, if the treatment is conducted in a single site, then it might not be the treatment itself that accounts for a large difference with the control group. For example, in an educational setting with one class in the treatment group and one in the control group, the treatment group could have an unusually effective teacher, and the treatment of interest, say the curriculum used, might not be the source of the differences between groups.

Most of the replication efforts we have identified sought to replicate an intervention that had promising results in the initial study by increasing the sample size and the number of locations where the intervention was tested. Replications are sometimes used to conduct research on other issues of interest including the following:

- *Additional target groups*. In some instances this includes other racial and ethnic groups, and in others extending the treatment from a general population to individuals with specific characteristics, such as disabilities.
- *Additional geographic locations*. For example, programs tested in urban areas have been replicated in rural areas.
- *Different intervention parameters*. Some programs, such as welfare programs and health insurance, can be characterized by a few specific characteristics for example the benefit reduction rate for earnings in a welfare program or the co-payment rate in a health insurance program. Replications may test a wider range of parameters than the original demonstration, or, if policy interest is more focused, they may test a narrower range of parameters.[1]
- *Additional related treatments*. If the replication effort includes enough observations, it is possible to expand the experiment to test related interventions in an "experiment within the experiment." For example, the income maintenance experiment replication efforts tested the effects of adding day care subsidies in one replication (Gary) and counseling, vocational education, and training in another (Seattle-Denver).

Interestingly, we are not aware of *any* cases in which replications were performed when the original evaluation found no impact. Although this is understandable, with the high cost of social experiments limiting the number that can be tested, it is worth considering whether any promising interventions have been dropped inappropriately because of negative or statistically insignificant findings in the initial evaluation, particularly when the sample size was small.

We use the term replication to designate sequential testing of an intervention, rather than testing an intervention simultaneously in multiple sites. Examples of initial tests in multiple sites using random assignment include the evaluation of the Jobstart demonstration (Cave et al. 1993), which was implemented in 13 sites; the National JTPA study, which was implemented in 16 sites (Orr et al. 1996); and the Employment Retention and Advancement demonstration (Hamilton and Scrivener 2012), which was implemented in 12 sites.[2] The rationale for implementing tests in multiple sites initially is similar to the reason for replicating promising demonstrations—to assure that the findings reflect the impact of the intervention rather than some other factor, to test the intervention in a variety of settings, and obtain more precision in the impact estimates.

Replication is not as simple a matter as it may first appear. A number of the replication efforts described below fail to find impact estimates similar to the original evaluation. Schorr and Farrow (2011) claim that the history of replication and scale-up efforts is "discouraging." They cite seven reasons why replication efforts have been unsuccessful:

- Insufficient understanding of what made the original intervention successful;
- Insufficient care and resources devoted to the quality of implementation and the process of scaling up;
- Insufficient attention to the culture within the helping organization and the regulatory and systems context surrounding it;
- Insufficient attention to local capacity and the organizational environment within which the intervention must be sustained;
- Failure to understand that what works for most children and families may not change outcomes for the children and families who are most at risk;
- Failure to understand the "uptake problem" among local front-line personnel and supervisors;
- Funders' reluctance to devote significant sums to the substantial operational costs of scaling up (Schorr and Farrow 2011, p. 17).

Although all of these factors can be issues, many times it is a failure to understand what made the intervention successful along with a tendency try to reduce expenditures in the replication efforts that lead to problems in replication efforts. In some of the cases described below, the problems are more obvious; for example, in two replication efforts, the Center

[1] Even in situations where the intervention is straightforward, there can be variations in the way that the intervention is explained to members of the treatment group.

[2] The Employment Retention and Advancement demonstration differs from the other two examples in that somewhat different treatments were implemented across the sites in that study, while the sites in the Jobstart demonstration all implemented the same basic model and the sites in the National JTPA study all implemented the JTPA program with some variation.

for Employment and Training (CET) and Quantum Opportunity Program (QOP), there was a failure to implement the program design with fidelity.

Social experiments have a longer history in the United States than in Europe, and replications date back to the 1960s and 1970s for both large and small social experiments in the United States. In recent years, interest in and use of social experiments has grown considerably in Europe, and a recent conference at the Institute for Employment Research focused on recent field experiments conducted in Europe and the United States.[3] As social experiments increase in Europe, the lessons from replications conducted in the United States can provide guidance on when and how to replicate.

2 Income maintenance experiments

The income maintenance experiments are generally considered the first major social experiments conducted in the United States.[4] These experiments were conducted to test an approach to welfare known as a "negative income tax," that was then popular with both liberal and conservative economists. The primary concern with the approach was that by covering two-parent families under such a plan, there would be a disincentive for the parents to work. The original income maintenance experiment was conducted in four cities in New Jersey and one city in Pennsylvania, and it is referred to as the New Jersey Income Maintenance Experiment. The four New Jersey cities had relatively few whites, so Scranton, Pennsylvania was included to add more whites to the experiment. The New Jersey experiment was sponsored by the US Office of Economic Opportunity (OEO), an independent federal agency established to study poverty issues, and ran from 1968 to 1972. The experiment included 1,357 families and tested variations involving four levels of benefits (50, 75, 100 and 125 percent of the federal poverty level) and three implicit tax rates on earned income (30, 50, and 70 percent of income).

Three replications of the New Jersey income maintenance experiment began before the field work for the New Jersey study was complete. First, OEO was concerned that the results from the largely urban areas covered by the New Jersey experiment might not apply to rural areas, so the agency instituted a similar experiment in three rural counties in North Carolina and Iowa in 1969, the year after the New Jersey experiment commenced. The Rural Income Maintenance Experiment was structured similar to the New Jersey experiment, and it included the same variations in the benefits provided and implicit tax rates; the Rural Income Maintenance Experiment was smaller, including 809 families.

The two remaining income maintenance experiments were set up and administered by the US Department of Health, Education, and Welfare (HEW and now Health and Human Services), and they were also started while the field work for the New Jersey experiment was ongoing. The HEW experiments were conducted in Gary, Indiana, and in Seattle, Washington and Denver, Colorado. Although similar to the original experiments, the Gary and Seattle-Denver experiments included slightly different benefit levels and implicit tax rates.[5] The Gary experiment was fielded from 1971 through 1975, and the Seattle-Denver experiment was fielded from 1971 through 1978. The two HEW experiments were larger than the OEO initiatives, with 1,799 families enrolled in Gary and 4,800 families enrolled in Seattle-Denver.

Greenberg et al. (2003) note that the findings from the income maintenance experiments were relatively similar across the experiments and largely consistent with economic theory as well as with previous nonexperimental estimates. Burtless (1986) estimated that the average effect of the experiments was a reduction in labor supply of 7 percent for husbands and 17 percent for wives in two-parent families.

Replications are usually conducted after the analyses from the original project are complete, so the timing on the income maintenance experiments is unusual. Greenberg et al. (2003) note several reasons for the early replication in this case. First, if welfare reform were to be implemented, it would be HEW that would be responsible for implementing the new welfare program, so HEW believed they should administer their own experiments. Second, the researchers at HEW "thought they could do it better." The HEW experiments were larger than the OEO studies, and, as previously mentioned, they also incorporated several interesting experiments within the experiments. Of course from a knowledge development standpoint, the major contributions of the additional experiments were the replications in additional environments than the original studies and the precision added by the increased sample size.

3 Unemployment insurance bonus experiments

The US unemployment insurance (UI) system has served as a laboratory for experimentation from the 1980s to the

[3]The conference was held at the Institute for Employment Research October 18–19, 2012. Copies of the presentations and papers are available at http://www.iab.de/en/veranstaltungen/konferenzen-und-workshops-2012/field_2012/programm.aspx, accessed 4/27/2013.

[4]The discussion of the income maintenance experiments is based largely on Greenberg et al. (2003).

[5]The benefit levels tested were 77 and 101 percent of the poverty level in Gary and 90, 116, and 135 percent of the poverty level in Seattle-Denver. The implicit tax rates on earnings were 40 and 60 percent in Gary and 50, 70, and 80 percent in Seattle-Denver.

present. The UI system is a federal-state partnership, with the federal government setting broad parameters for the system and funding the infrastructure, with states paying for benefits, primarily through employer payroll taxes, and establishing specific rules for qualification for benefits and the level of benefits received.[6] Like welfare programs, there is a tension in the UI system between providing adequate support to beneficiaries and keeping costs reasonable, and more importantly, UI beneficiaries can affect the benefits they collect by varying the intensity of their job search and the wage rate and type of work they will accept. During the 1980s, social experiments were conducted and replicated on two broad approaches to reducing UI costs.

One approach was to emphasize responsibilities of UI claimants to search for work and accept suitable job offers. Experiments in this category required claimants to attend job search workshops and, in some instances, to increase the number of contacts with employers and/or to provide increased documentation of their job search activities. Although often characterized as a "stick" rather than a "carrot" approach, the job search activities may have provided valuable assistance to claimants in their job search. Nevada and Wisconsin implemented experiments in 1977 and 1983, respectively, and Minnesota implemented its experiment in 1988. The first federally sponsored experiment in this category was carried out in Charleston, South Carolina in 1983, and similar experiments were conducted New Jersey and Washington in 1986 and 1987.[7] (The job search experiments are discussed below.)

An alternative approach to requiring greater job search by claimants is the payment of cash bonuses to claimants who find a job in a specific period of time. The underlying concept is that by offering a cash bonus to claimants who find a job quickly, claimants will search more intensively than they otherwise would. The initial reemployment bonus experiment was conducted by the State of Illinois in 1984. Claimants who found a job within 11 weeks of filing their initial claim and remained employed for at least four months were eligible for a cash payment of $500.00. The experiment included a control group of 3,952 claimants, and there were two treatment groups. In one treatment, 4,186 claimants were eligible to receive the bonus payment (the claimant experiment), and in the second treatment, 3,963 employers were eligible for the bonus (the employer experiment).

The employer experiment was characterized by a relatively low participation rate among eligible claimants, 64 percent according to Woodbury and Spiegelman (1987), compared to 84 percent for the claimant experiment. Overall, there were no statistically significant differences in weeks of benefits or benefits collected in the employer experiment, although Woodbury and Spiegelman (1987) note that weeks of unemployment for white women were reduced by about one week in the employer experiment. Likely because of the low participation rate and the lack of a statistically significant impact, the employer experiment was not included in any of the three replications.

The claimant experiment, however, had a large statistically significant impact on benefit duration in Illinois. Duration was reduced by 1.15 weeks for those eligible for the benefit year and benefits were reduced by $194. With over 9 million initial claims annually projected by the US Department of Labor 2012 through 2017, the potential savings from such a program are very high.[8] Woodbury and Spiegelman (1987) note that there is not a statistically significant difference in earnings between the treatment and control groups, so the treatment group is not taking lower paying jobs. Woodbury and Spiegelman (1987) compute the ratio of benefits to costs from the state perspective for the claimant experiment, and they find the ratio for all groups to be 2.3. Note, however, that for the claimants themselves, the finding of no change in earnings in the year following the initial claim combined with an average receipt of benefits for one week less implies that claimants are worse off from the treatment. Moreover, as Woodbury and Spiegelman (1987) point out, the favorable benefit-cost ratio depends critically on the fact that only 54 percent of those eligible for a bonus actually applied for a bonus. Meyer (1995) notes that if reemployment bonuses become a permanent policy, the proportion of eligible claimants who claim the bonus might well increase, thus driving up the costs of such a policy.

Given the large benefits of the reemployment bonus experiment in Illinois, it is not surprising that the US Department of Labor decided to replicate the experiment in three more states—New Jersey, Pennsylvania, and Washington. The replications were implemented fairly quickly after the results from the Illinois experiment were available: July 1986 to June 1987 for New Jersey, February 1988 to November 1988 for Washington, and July 1988 to October 1989 for Pennsylvania. The replications were more complicated than the Illinois study, with six treatments tested in each replication. Corson and Spiegelman (2001a) explain why the New Jersey experiment is excluded from the Robins and Spiegelman (2001) volume on reemployment bonuses:

> Although the bonus offer treatment in the New Jersey experiment had strong results, we do not believe that

[6] For a description of the UI system, see http://ows.doleta.gov/unemploy/uifactsheet.asp retrieved September 9, 2012 or Corson and Spiegelman (2001a).

[7] See Corson et al. (1985) for the evaluation of the Charleston experiment; Corson and Haimson (1996) for the evaluation of the New Jersey experiment.

[8] Obtained from https://ows.doleta.gov/unemploy/pdf/MSR.pdf retrieved on September 29, 2012.

> this experiment provided much guidance for policy, because the bonus-offer treatment was not replicable. In the New Jersey experiment, bonus offers were made only after seven weeks of insured unemployment, and the pending offer was unknown to the selected participants prior to that time. Such a situation could not be replicated in a real program, as knowledge of the pending offer would be available to all claimants from the start of their benefit year... This knowledge can be expected to critically affect job-search behavior during the first seven weeks of the benefit year, as well as during the period in which the bonus was available (p. 14).

The New Jersey replication also differed from the other bonus experiments because the treatment groups were required to participate in a job search assistance component and the size of the bonus was more variable.[9] In the other three replications, there was some variation in the amount of the bonus, but Meyer (1995) notes that the bonus available was generally close to the $500 offered in the Illinois program. All three replications differed from the Illinois program in excluding from bonus eligibility claimants who returned to their previous employer.

The three replications had similar sample sizes to the Illinois experiment, but the manner in which sample sizes were determined varied across the states, according to Corson and Spiegelman (2001b). In Illinois, sample size was determined by estimating how many participants could be enrolled to exhaust the $750,000 the state had available for the bonus pool. In Washington, the sample size was set using the results of the Illinois experiment to determine the sample size needed to produce statistically significant estimates of the reduction in benefit weeks for each treatment cell if the true effect were the same as in Illinois (1.15 weeks). The Pennsylvania budget was based on the budget developed for Washington.

As noted previously, the reduction weeks of benefits received was 1.15 weeks in the original Illinois experiment. As is often the case, impacts were not as impressive in the replication studies. In New Jersey, benefit weeks were reduced by 0.9 weeks for the bonus plus job search assistance, but Meyer (1995) points out that the reduction for job search alone was 0.5 weeks, so the marginal impact of the bonus was 0.4 weeks. In Pennsylvania and Washington, most of bonus treatments had modest impacts that were generally not statistically significant, but Meyer (1995) shows in his summary that in both states the treatment that included a high bonus and a long qualification period had a fairly large and statistically significant negative impact on duration, 0.9 weeks in Pennsylvania and 0.7 weeks in Washington.

Although the Illinois experiment created great interest among economists and policy makers, the smaller impacts in the replications and changes in policy tastes led to a lack of continued interest in such measures. Meyer (1995) believes that an ongoing reemployment bonus program would have higher bonus take-up rates, making the benefit-cost analysis less favorable from the viewpoint of the UI program. Corson and Spiegelman (2001a) are more optimistic about the potential role of reemployment bonuses. More recently, however, policies providing reemployment services have been more popular with policy makers. The American Recovery and Reinvestment Act of 1999 (ARRA) included nearly $247 million for reemployment services for UI claimants. The US Department of Labor has provided support to states for Reemployment and Eligibility Assessment (REA) activities, which combine mandatory participation in activities that are intended to increase the likelihood of reemployment with strict enforcement of the work test.

4 Center for Employment Training

Center for Employment Training (CET) is a nonprofit community based program that was founded in 1967 in San Jose, California.[10] CET originally focused on serving migrant and seasonal farm workers, but the program expanded its target groups to include welfare recipients, high school dropouts, displaced workers, and other groups experiencing problems in the labor market. In the 1980s, CET distinguished itself by participating in two randomized controlled trials, JOBSTART and the Minority Female Single Parent Demonstration (MFSPD), and showing large, statistically significant impacts on employment and earnings in both efforts. Importantly, CET was the only site in these multi-site demonstrations to show large impacts on the outcomes of interest. Consequently, the US Department of Labor decided to replicate the CET program in 12 additional sites.

JOBSTART was implemented between 1985 and 1988 in 13 sites that included community based organizations, schools, and Job Corps centers. The program enrolled 17 to 21 year old economically disadvantaged high school dropouts who participated in comprehensive programs offering vocational training, supportive services, and job placement assistance. Sites offered the services either concurrently or sequentially. The average length of stay in the program was 6.8 months, but about 16 percent of the participants remained over one year. The demonstration included

[9] Meyer (1995) notes that the bonus was initially equal to 50 percent of the remaining UI benefit entitlement, but declining by 10 percent per week.

[10] Background information on CET was obtained from their Internet site http://cetweb.org/about-us/mission-and-history/ retrieved August 27, 2012.

1,941 participants who were randomly assigned to treatment or control status. Nearly the entire treatment group received some education and training (94 percent), and over one-half the control group received such services (56 percent).

Cave et al. (1993) report that overall the program had little impact on employment and earnings in the years following participation. Although earnings were about 10 percent higher for the treatment group than the control group in the fourth year after random assignment ($5,592 for the treatment group and $5,182 for the control group), the difference in earnings was never statistically significant. The treatment group had smaller gains in hours worked and percentage ever employed during the year, and these differences were never statistically significant.

Of the 13 JOBSTART sites, only CET had a statistically significant impact on earnings for participants. For the third and fourth years after random assignment combined, the CET treatment group earned $20,808 on average, and the control group earned $14,721; the earnings impact for the next best performing site was only $2,251 and was not statistically significant. Although the CET site clearly had a much greater impact than the other sites, it is difficult to see why this was the case based on observable characteristics of the program. CET had lower than average hours of participation. The operating costs per participant were less than half the costs at 10 of the 12 other sites. Cave et al. (1993) attempted to adjust for differences in programs and participants across sites, but CET still appeared to be the only effective site after making the statistical adjustments. Cave et al. (1993) conclude that "it is very difficult to identify what features in the CET/San Jose approach led to its strong impacts".

During roughly the same period, 1982 through 1988, CET also participated in the Minority Female Single Parent (MFSP) demonstration. This demonstration, funded by the Rockefeller Foundation, specifically targeted minority single parents and used community based organizations in four large cities to test "whether comprehensive employment-training and support services could enhance the self-sufficiency of minority single mothers and reduce their dependence on welfare" (Burghardt et al. 1992, p. xiii). The four cities and CBOs were Atlanta Urban League (AUL) in Atlanta, Georgia; Opportunities Industrialization Center (OIC) in Providence Rhode Island; Wider opportunities for Women (WOW) in Washington, DC; and the Center for Employment Training (CET) in San Jose, California. A total of 3,965 women were randomly assigned to the treatment and control groups across the four sites.

The four sites all provided education and training services as well as supportive services, but there were differences in the length of training, the supportive services provided, and the manner in which education and training were linked. AUL provided all training through outside vendors, CET and WOW provided all services in-house, and OIC provided training both in-house and through outside vendors. AUL, CET, and OIC provided training for specific occupations, while WOW provided an introduction to skills needed for nontraditional occupations for women. The length of vocational training varied within and across sites. For example, AUL courses lasted 8 weeks to two years, CET training was four to nine months, OIC courses lasted six to nine months, and WOW courses were 10 or 20 weeks long. All sites except CET tested participants for reading and math skills, and those with low scores received basic skills instruction; at CET, there was no testing, and all participants were immediately enrolled in vocational training, where they received basic skills instruction if needed.

In all four sites, a large proportion of the treatment group received training in absolute terms, as well as relative to the control group, with over 70 percent of the treatment group receiving education and/or training, while about 30 percent of the control group in each site received such services. As in the JOBSTART demonstration, the results at the CET site were much stronger than at other sites. Over the 30 months following random assignment, earnings of the CET treatment group exceeded the earnings of the control group by $2,000; the earnings differences between the treatment and control groups were small and insignificant at the other three sites. The only site with somewhat encouraging impacts other than CET was WOW, and while it had statistically significant employment gains, the treatment group did not earn more than the control group in the post-program period; moreover, while the cost-benefit analysis of CET indicated a positive net present value during the five years following random assignment, neither WOW nor any of the other sites indicated benefits to society that exceed the costs. Thus, as in the JOBSTART evaluation, CET appeared to be the only effective program.

Unlike JOBSTART, however, the MFSP evaluators offered some potential reasons why the CET program was effective and the other programs were not. As noted above, CET was the only MFSP site that did not test participants and integrated basic skills with vocational training. Providing contextualized basic skills training remains a popular concept and is a key component in Washington State's Integrated Basic Education and Skills Training (IBEST).[11] Burghardt et al. (1992) conclude that:

> CET used an unusual open-access, integrated training design. Its design was distinguished by two features:

[11] In a nonexperimental evaluation using propensity score matching, regression analysis, and difference in difference analyses, Zeidenberg et al. (2010) found statistically significant impacts of I-BEST on six of seven educational outcome measures, but they found no statistically significant impact on labor market outcomes, which they hypothesize could be due to the recession that began during the post-program period.

> that women would enter job training immediately, regardless of their previous educational attainment and that remedial education would be integrated directly into training for a specific job, rather than provided prior to job training or concurrently in a separate class. Job training at CET focused on competencies required by employers for specific jobs; it emphasized training in occupations in which jobs were plentiful, as well as immediate placement in jobs after training. (p. XV)

The evaluators note that although they can be confident that the MFSP program was effective in the CET site, the small number of sites made it difficult to judge whether it was the curriculum itself that accounted for the program's success or some other feature of CET or local economic conditions that led to success at that site. Thus, the researchers recommended replication in other sites.

Based on the promising results of CET in the JOBSTART and MFSP demonstration, The US Department of Labor funded a replication of CET in 12 sites. Between 1995 and 1999, 1,485 out-of-school youth were enrolled in CET replications in 12 sites and randomly assigned to receive the CET treatment or the control group, where they were not allowed to enroll in CET but could participate in other employment and training programs. The CET replications took place in six CET offices in western states that had already implemented the CET model and six eastern and Midwestern locations that were established for the replication demonstration. Two rounds of follow-up were conducted, at 30 and 54 months after random assignment, and the evaluation results are reported in Miller et al. (2003, 2005).

Unlike the original evaluations of CET, the results in the replication are not simple to describe and interpret. First, Miller et al. (2003) report that the CET model was not instituted with strong fidelity in all sites. In the four sites that had been part of the CET system for a number of years, the model was instituted with high fidelity, but Miller et al. (2003) report that in the sites that were added, the program model was implemented with medium fidelity in six sites, and in two sites, the model was implemented with low fidelity. In particular, the authors report that it was difficult to implement the program with the intensive participation in training and strong organizational stability that characterized the original CET program in San Jose.

Impact estimates varied by whether the participants were in a high-fidelity site or one of the other sites, by sex, and by follow-up period. Impact estimates for employment and earnings for the medium and low replication sites were consistently very small or negative and were never statistically significant. The findings for the high fidelity sites were more complex. In the 30-month follow-up evaluation, women in high fidelity sites had positive impacts across most employment and earnings outcomes, although the impact estimates were usually not statistically significant. For men in high fidelity sites and both sexes in sites with low or medium fidelity, impact estimates in the 30-month follow-up analysis were usually zero or negative.

In the 54-month follow-up evaluation, the results were even less promising. Although CET participants received more education and training than the control group over the follow-up period, there were no statistically significant differences in employment and earnings between the treatment and control groups. In high fidelity sites, the findings by sex in the 54-month follow-up were reversed relative to the 30-month follow-up: treatment group women earned less than control group women, while treatment group men now earned more than men in the control group. We would not, however, emphasize subgroup differences. As Miller et al. (2005) concluded,

> "Access to CET did not lead to better outcomes than youth would have had on their own, either by enrolling in other training programs or by gaining experience in the labor market" (p. iii).

Miller et al. (2005) offered three hypotheses on why the CET replications failed to match the impressive impacts of CET in the original JOBSTART and MFSP demonstrations. First, the participants in the replication sites may not have needed the program as much as in the original demonstrations. JOBSTART in particular was targeted on more disadvantaged youth, and the economy had improved by the time the replications were implemented. The second hypothesis was that the treatment group failed to take advantage of the training and credentials received; as evidence for this hypothesis, they note that treatment group members often failed to report the credentials they received. Finally, they suggest that the strong features of CET were more commonly available in other programs during the replication period as other training providers learned from CET and other successful training organizations. Although these explanations are all plausible, Miller et al. (2005) could not identify which, if any, were correct.

The CET experience illustrates the challenges and limitations of replicating a promising practice. The strong impact findings in not one but two earlier multi-site demonstrations point to a very strong likelihood that CET was a highly effective program, and DOL was wise to try to replicate the program. But as Hollister (1990) notes, there were many features in CET that may have contributed to its large impact on youth. It is impossible to know which features are the key ones required for success. Moreover, Miller et al. (2003 and 2005) document the difficulties in replicating the CET model—only four of the 12 replications sites were able to replicate the model with high fidelity, and these were the sites that were already part of the CET network. And as Miller et al. (2005) note, the impact also may depend on the environment, particularly local economic conditions, as well as the merits of the program.

5 Job Clubs

Although group job search activities, often known as "job clubs" or job search workshops, are commonly used by workforce and welfare programs today, they were very rare prior to the 1970s. In a series of social experiments conducted by Nathan Azrin and colleagues, group job search was tested against the more conventional individual-oriented approach. The first such study, Azrin et al. (1975), tested the job club approach on 120 individuals searching for a job and referred by sources such as a public employment agency, employer personnel departments, and word of mouth. Half the group was randomly assigned to participate in job club activities, and half received no job finding services. Although the analysis was not up to the standards commonly used today, the results were very strong—within two months, 90 percent of the treatment group had a job, but only 55 percent of the control group was employed.[12] Azrin replicated his efforts with two specialized target groups. Azrin and Philip (1979) tested the job club approach against an individualized assistance approach for individuals with disabilities. Once again, the results were striking—95 percent of the treatment group was employed at the six-month follow-up point compared to only 28 percent of the control group. With funding from the US Department of Labor, Azrin et al. (1980) conducted a much larger scale experiment to determine whether job clubs are effective for welfare recipients. This study was carried out in five cities, and nearly 1,000 welfare recipients were randomly assigned to receive either group job search through job clubs or the usual employment services provided to welfare recipients in the city. Once again, the results were striking—at the 12-month follow-up, 87 percent of the job club sample had obtained jobs compared to 59 percent of the control sample. There were statistically significant differences favoring the treatment group in all five sites and for every subgroup examined. The Department of Labor then retained the Manpower Demonstration Research Corporation (now MDRC) to replicate the demonstration on a large scale in Louisville, Kentucky, a site that had agreed to serve as a laboratory to test promising strategies for welfare recipients. Welfare recipients who registered in Louisville between October 1980 and May 1981 were eligible to participate in the experiment. Wolfhagen and Goldman (1983) find that a total of 750 individuals participated, and two quarters after random assignment, 49 percent of the treatment group was ever employed compared to 34 percent of the control group; earnings for the treatment group over this period were $550 for the treatment group and $144 for the control group. The Azrin studies and the MDRC replication helped change the workforce community's perception about whether job seekers should be served individually or in groups, and group activities are now a widely used strategy.

[12] Possible concerns with the analysis include failure to control on participant characteristics, which could be important given the small sample size, and the fact that participants who attended four sessions or less (and their matched pair person) were excluded from the analysis. (The published article does not indicate what proportion of the treatment group was excluded because of this criterion.) Also, the control group received no job search services, so while the study establishes that group job search is superior to no counseling, it does not compare group job search to individualized services. Finally, the study was carried out in a single small town in Southern Illinois, so there could be external validity concerns.

6 Job search experiments

An important set of demonstrations was conducted in the 1970s and the 1980s to test the effects of mandatory services for unemployment insurance claimants on their receipt of unemployment insurance and their employment and earnings. The US Department of Labor sponsored a demonstration in Charleston, South Carolina in 1983 that required UI claimants to participate in job search activities and undergo stronger enforcement of the work test requirement that claimants be available and actively searching for work.[13]

The Charleston experiment included 1,428 individuals assigned to the control group and 4,247 claimants assigned to one of three treatment groups. All three treatment groups received enhanced enforcement of the UI work search requirements where claimants were required to register with the Employment Service to receive benefits, and claimants' job search activities were monitored more closely than was usually the case. Treatment group 1 also received enhanced placement activities and job search workshops. The enhanced placement activities consisted of a placement interview that was expected to result in either a job referral or efforts to develop a job for the claimant, and claimants were taught how to use the job listings maintained by the agency. The job search workshop lasted three hours and taught participants job search strategies, interviewing techniques, and how to access labor market information. Treatment group two received the enhanced enforcement of the work search requirements and enhanced placement efforts, but not the job search workshop. Treatment group 3 only received the enhanced enforcement of work search requirements.

The evaluation found statistically significant reductions in weeks of UI benefits for all three treatment variations. Corson et al. (1985) report that the reduction in weeks of UI ranged from 0.5 weeks to about 0.75 weeks; Meyer (1995) summarizes the results in similar terms. In terms of impacts on workers, Corson et al. (1985) report that their

[13] The description of the Charleston demonstration is based on Corson et al. (1985).

analysis "did not provide any strong indications of the effects of the demonstration treatment on claimants' reemployment success" (p. 67).[14] Corson et al. (1985) report that the Charleston demonstration resulted in savings per claimant of $46 to $56; although this might not sound like a large figure, new UI claims currently exceed 300,000 per week, so the potential for cost savings is quite large. Meyer (1995) performs a rough cost-benefit analysis and finds that the Charleston demonstration produced net benefits to claimants, government, and society.[15]

The Department of Labor conducted subsequent reemployment demonstrations in New Jersey and Washington in 1986 and 1987. These demonstrations were not replications of the Charleston demonstration, but they tested alternative combinations of reemployment services and work test enforcement.[16] Although the specific treatments varied in the three demonstrations, Meyer (1995) notes that the three evaluations

> "all show reductions of about one-half of a week in UI receipt with more intensive services and oversight." (p. 116)

In addition to the Department of Labor replications of the reemployment services demonstrations, several states also conducted their own demonstrations and evaluations of reemployment services for claimants, generally with similar results.[17]

The weight of the evidence in the claimant reemployment services demonstrations provides a strong case for providing intensive services to claimants along with strict enforcement of the work test.[18] In 2005, the US Department of Labor began providing grants to states to provide Reemployment and Eligibility Assessment (REA) to claimants.[19] The REA effort began with $18 million awarded to 21 states in 2005 and has grown to $65.5 million in 40 states in 2012.[20] Although the initial impact evaluations reported in Benus et al. (2008) were based on nonexperimental methods in nine states, the later evaluation by Poe-Yamagata et al. (2011) uses experimental designs in four states. The impact evaluations based on random assignment in the Poe-Yamagata et al. (2011) study found large decreases in benefit weeks for three states (1.74 weeks in Florida, 1.14 weeks in Idaho, and 2.96 weeks in Nevada), but no impact in the fourth state, Illinois, which they attribute to inconsistent implementation and a small sample size. Although participation in the REA program is optional for states, a majority participate, and participating states are required to maintain a random sample of those eligible as a control group. This will permit the US Department of Labor to update its evaluations on an ongoing basis and see which variations are most successful in reducing benefits and assisting claimants.

7 Quantum opportunity program

As previously discussed, fidelity to model design was a major problem with the Center for Employment and Training (CET) replications. Another program with replication problems was the Quantum Opportunity Program (QOP). QOP was an after-school program intended to provide services to at-risk youth entering high school. In the original evaluation by Hahn et al. (1994), the evaluators concluded that the program was instituted with such low fidelity in one site that they did not perform the impact evaluation in that site. QOP was replicated with funding from the US Department of Labor and the Ford Foundation from 1995 through 2001 in seven sites. Maxfield et al. (2003) note that the replications were also plagued with replication issues. The evaluators concluded that two sites deviated substantially from the QOP model, and they found that most sites failed to implement some of the key features of QOP; only two of the sites offered the specified educational, developmental, and community service activities at appropriate levels.

8 Concluding thoughts on replications

The examples above show that there have been replications of a wide variety of demonstrations that included random assignment. It appears, however, that decisions on whether and how to implement replications is done mostly on an ad hoc

[14] There is ambiguity on whether the demonstration might be expected to increase or decrease participant earnings. If effective, the reemployment services could lead to a reduction in time unemployed and help claimants find a better job. On the other hand, stricter enforcement of the work test may have led claimants to accept jobs they would otherwise have passed up due to low wages.

[15] The cost-benefit analysis is not precise because the earnings data used to capture benefits for participants is not precise, a fact noted by Meyer (1995) and in the Corson et al. (1985) evaluation.

[16] See Corson and Haimson (1996) for the evaluation of the New Jersey demonstration and Johnson and Klepinger (1991) for the evaluation of the Washington demonstration.

[17] See Meyer (1995) and Meyer (1992) for discussions of the state demonstrations conducted in Wisconsin, Nevada, and Minnesota. Meyer (1995) notes that the state evaluations were generally not as carefully done as the Department of Labor evaluations.

[18] Ashenfelter et al. (2005) review four state experiments where the only treatment was strict enforcement of the work test with no concomitant reemployment services. They conclude "we found no evidence that verification of claimant search behavior led to shorter claims or lower total benefit payments" (p. 70).

[19] For a discussion of the history of REA and its predecessors, see Wandner (2010).

[20] REA figures for 2010 are from Benus et al. (2008), and information for 2012 are from a May 7, 2012 news release issued by the Employment and Training Administration accessed at http://www.dol.gov/opa/media/press/eta/ETA20120916.htm on September 19, 2012.

basis. In this section we propose some concepts that should be more systematically considered in deciding about replications and discuss the implications of this research for other nations.

8.1 Deciding when replication is desirable

As noted earlier, replications are generally performed for demonstrations judged to be successful, but that is a fairly vague criterion. Beyond the simple value of adding additional observations, replications can be used to test the intervention in different geographic regions, under different economic conditions, and on different target groups. Expanding implementation to different types of individuals and different environments can answer some of the concerns that often arise regarding the external validity of the findings from an evaluation. Can the results from an evaluation ever be so positive that no replication is needed? The evaluation of deworming in Kenya using randomized controlled trials and the accompanying cost-benefit analysis provided such powerful evidence of the efficacy of programs to eliminate intestinal worms, that after the articles on the success of the Kenya program by Miguel and Kremer (2004) and Kremer and Miguel (2007), a major international initiative was started to replicate the program without major efforts to replicate the initial success.[21,22] Alternatively, consistent statistically significant findings across a substantial number of sites in the initial experiment might eliminate the need for replication.

8.2 Efforts to assure fidelity

Several of the social experiments described here suffered from fidelity issues. Examples noted above include the CET program, Quantum Opportunity Program, and one of the states implementing a reemployment and eligibility assessment program. As we noted earlier, sometimes it is not known which components of a model should be considered crucial in the replication effort. Conducting a process study in conjunction with the impact evaluation is essential to learn whether a program was implemented with fidelity. Unfortunately, there is no simple way to know which aspects of a program are critical to its success without conducting replications that systematically omit specific aspects of the program.

8.3 How many sites and how many participants in the replication

Many of the replications appear to have been set up on an ad hoc basis, typically with the budget available largely dictating the sample size and number of sites and convenience and willingness to participate often appearing to be the key determinants of where the replications occur. A notable exception described in Corson and Spiegelman (2001b) was that the Washington replication of the Illinois UI bonus experiment used information from the Illinois experiment to determine the number of participants to include in the experiment. Perhaps this practice is carried out more than we were able to detect, but replications should be able to make use of data from the initial experiment to determine the minimum sample size that is needed to obtain results that are useful for policy decisions.

8.4 Expansion to other target groups

In some cases, the intervention in the original demonstration was conducted on a single target group or a very general population. If the intervention is successful, it is natural to see if the promising results hold up with other populations. The original job club demonstration evaluation by Azrin et al. (1975) was carried out on a general population of job seekers, but the replications in Azrin and Philip (1979) and Azrin et al. (1980) focused on people with disabilities and welfare recipients, respectively. The National Supported Work demonstration, in operation from 1975 through 1980 and summarized by the Manpower Demonstration Research Corporation (1980), replicated the Vera Institute of Justice's experimental evaluation of its Wildcat supported work project for former substance abusers by adding three additional target groups—long-term welfare recipients, disadvantaged youth, and ex-offenders. The Vera evaluation showed a positive impact for former substance abusers, while the national supported work demonstration found strong impacts on employment and earnings for welfare recipients, more modest gains for the former substance abusers, and no significant gains for the ex-offenders and disadvantaged youth.

8.5 Lessons for other countries

The use of social experiments for evaluating promising labor market policies has grown dramatically in recent years. A conference on social field experiments sponsored by the Institute for Employment Research in October 2012 included 10 presentations on field experiments conducted in Europe, as well as several papers about US and Canadian

[21] The deworming effort was replicated in India, but the effort also included interventions to reduce iron deficiency anemia. See Bobonis et al. (2006).

[22] In some situations, random assignment is not needed at all. Since 1986, the Carter Foundation has sought to eliminate Guinea worms and the disease they cause by breaking their reproductive cycle. The number of cases worldwide has dropped from 3.5 million in 1986 to 1,058 in 2011. See https://www.cartercenter.org/health/guinea_worm/mini_site/index.html accessed on September 19, 2012.

experiences.[23] As knowledge based on social experiments grows, it will be important to decide when it is useful to replicate promising approaches demonstrated through the experiments. In fact, Rosholm (2012) describes a series of experiments conducted in Denmark involving meetings with caseworkers. Rosholm (2012) describes how, after achieving success in their initial experiment (employment was three weeks greater in the follow-up period for the treatment group), which included a variety of strategies for assisting the unemployed, successive experiments that have been conducted or are planned were designed to learn more about how the policies work and for whom they work best. Variations included using group meetings versus individual meetings with caseworkers, frequency of meetings with caseworkers, targeting workers at different times in their unemployment spells, and targeting specifically on youth and people with disabilities. These variations are similar to replication efforts observed in the United States, but it is important for countries to evaluate policies in their own setting rather than rely on findings from other nations. Variations in culture, education and labor market policies, and assistance programs make it possible that results can vary significantly across countries.

Executive summary

When evaluating a demonstration program, there are risks from drawing inferences from a single test. This paper reviews the experiences of replication efforts from demonstrations using randomized controlled trials in the initial evaluation and the replications. Replications are undertaken for several reasons, including (1) offering the treatment to additional target groups, (2) adding additional geographic locations, (3) varying some of the treatment design parameters to better understand what aspects of a treatment are responsible for its success, and (4) testing additional interventions related to the original treatment. Replications of six randomized controls are reviewed, and all but one (the Quantum Opportunity Program) are summarized here.

The income maintenance experiments are generally considered the first major social experiments in the United States. These experiments were conducted to learn more about the labor supply responses of families to welfare programs that guaranteed a certain level of income regardless of work effort. The New Jersey Income Maintenance Experiment was conducted from 1968 through 1972, and three replications were initiated before the field work was complete. The additional studies were undertaken for several reasons: to add additional locations and thereby increase external validity, to increase the sample size in the evaluations, and (to some extent) due to rivalry between two federal agencies. The original experiment and the replications had fairly similar findings that were consistent with economic theory—labor supply was reduced by about 7 percent for husbands and 17 percent for wives in two-parent families—reinforcing confidence in the results.

The Unemployment Insurance (UI) Bonus Experiments were implemented to test whether offering claimants a bonus for returning to work within a reasonable period would reduce costs to the UI system. The original bonus experiment was conducted in Illinois in 1984 and provided a bonus of $500 to claimants who found a job within 11 weeks of filing a claim. The experiment was very successful, with UI duration reduced by 1.15 weeks and payments reduced by $194. Given the success of the original experiment, replications were conducted in New Jersey, Pennsylvania, and Washington, with disappointing results. The reduction in weeks claimed was 0.5 weeks in New Jersey, and the impacts in Pennsylvania and Washington were generally smaller and not statistically significant. The lack of strong results in the replication states reduced interest in the employment bonus strategy.

The Center for Employment and Training (CET) is a nonprofit community based program in San Jose California that serves groups with labor market problems. During the 1980s, CET participated in two randomized controlled trials (RCTs) and was the only site in either demonstration to achieve labor market success. Based on these promising findings in two separate RCTs, the US Department of Labor replicated the CET program in 12 sites. The findings in the replication sites were disappointing, with no statistically significant findings 54 months after random assignment. An important limitation of the replication efforts was that only four of the 12 sites implemented the CET model with high fidelity. The evaluators offered several hypotheses on why the replications failed to reproduce the strong findings in the original studies, but there was no way to test the hypotheses.

In the 1970s, Nathan Azrin conducted a series of small experiments to test whether group job search activities conducted through "job clubs" led to improved labor market outcomes for job seekers over traditional individual approaches. Azrin found that the group activities had statistically significant large impacts; in his first study, for example, he found within two months, 90 percent of the treatment group had a job, but only 55 percent of the control group was employed. The US Department of Labor then funded several larger experiments and again found strong gains. Group job search is now widely used.

The Department of Labor supported three experiments in the 1980s that tested a Reemployment and Eligibility Assessment program that included job search requirements for

[23] Papers and presentations from the conference are available at http://www.iab.de/en/veranstaltungen/konferenzen-und-workshops-2012/field_2012/programm.aspx,accessed 4/27/2013.

UI claimants and provision of job search assistance and labor market information to help them find jobs. Both the original RCT in Charleston, South Carolina and replications in New Jersey and Washington found reductions of claim duration by about one-half week with no loss of earnings for claimants. The Department of Labor now supports similar efforts in 40 states, and additional evaluations of these through RCTs have also shown cost savings.

Replications of workforce programs have proven a valuable tool in developing workforce policy. In several cases, the original findings were not sustained and the interventions have been discontinued. (However, sometimes the lack of significant outcomes results from a lack of fidelity in the replications.) In other instances, such as for job clubs and job search experiments, the replications have demonstrated that a particular type of intervention is worthwhile and the standard mix of services for job seekers has been changed.

Kurzfassung

Bei der Bewertung eines Testprogramms besteht die Gefahr, aus einem einzelnen Test Rückschlüsse zu ziehen. In dieser Arbeit werden die Erfahrungen mit Wiederholungen von Testprogrammstudien anhand einer randomisierten, kontrollierten Studie für die erstmalige Auswertung und die Wiederholungen besprochen. Wiederholungsstudien werden aus unterschiedlichen Gründen durchgeführt, z. B. (1) um eine Erweiterung auf zusätzliche Zielgruppen durchzuführen, (2) um zusätzliche geografische Standorte einzubeziehen, (3) um einige der Verfahrenseckdaten zu variieren, um ein besseres Verständnis für die Aspekte eines Verfahrens zu erlangen, die für dessen Erfolg verantwortlich sind, und (4) um zusätzliche, mit dem ursprünglichen Verfahren im Zusammenhang stehende Interventionen zu testen. Wiederholungsstudien von sechs randomisierten, kontrollierten Studien werden hier besprochen und bis auf das „Quantum Opportunity“-Programm auch zusammengefasst.

Die Experimente zur Einkommenssicherung werden im Allgemeinen als erste wichtige Sozialexperimente in den USA betrachtet. Diese Experimente wurden durchgeführt, um mehr über die arbeitsmarkttechnischen Reaktionen von Familien auf Sozialprogramme zu erfahren, die unabhängig von der Arbeitsleistung ein bestimmtes Einkommensniveau garantierten. Das New Jersey Income Maintenance Experiment (Experiment zur Einkommenssicherung in New Jersey) wurde zwischen 1968 und 1972 durchgeführt. Bevor die Feldphase abgeschlossen war, wurden drei Wiederholungsstudien initiiert. Die zusätzlichen Studien wurden aus verschiedenen Gründen durchgeführt: um zusätzliche Standorte einzubeziehen und somit die Aussagekraft zu erweitern, um den Stichprobenumfang der Erhebungen zu erhöhen und (bis zu einem gewissen Grad) aufgrund der Rivalität zwischen zwei bundesstaatlichen Agenturen. Das ursprüngliche Experiment und die Wiederholungsstudien führten zu recht ähnlichen Ergebnissen, die mit der ökonomischen Theorie übereinstimmten – das Arbeitsangebot reduzierte sich bei Familien mit zwei Elternteilen um ca. 7 % für Ehemänner und 17 % für Ehefrauen – womit das Vertrauen in die Ergebnisse bekräftigt wurde.

Die Experimente zu Bonuszahlungen bei der Arbeitslosenversicherung wurden durchgeführt, um festzustellen, ob die Zahlung eines Bonus für Antragsteller bei Rückkehr in den Job innerhalb einer angemessenen Zeitspanne die Kosten des Arbeitslosenversicherungssystems reduzieren würde. Das ursprüngliche Bonusexperiment wurde 1984 in Illinois durchgeführt. Dabei wurde Antragstellern, die innerhalb von 11 Wochen nach Antragstellung eine Arbeitsstelle gefunden hatten, ein Bonus von 500 $ gewährt. Das Experiment war sehr erfolgreich; die Dauer des Arbeitslosenversicherungsbezuges reduzierte sich um 1,15 Wochen und die Zahlungen um 194 $. Aufgrund des Erfolges des ursprünglichen Experiments wurde es in New Jersey, Pennsylvania und Washington wiederholt. Die Ergebnisse waren jedoch enttäuschend. Die Reduktion der Dauer lag in New Jersey bei 0,5 Wochen, in Pennsylvania und Washington waren die Auswirkungen im Allgemeinen noch geringer und daher statistisch nicht relevant. Die fehlenden aussagekräftigen Ergebnisse in den Wiederholungsstaaten minderten das Interesse an der Beschäftigungsbonusstrategie.

Das Center for Employment and Training (CET) basiert auf einer gemeinnützigen Gemeinschaft und ist ein Programm in San Jose, Kalifornien, für Personengruppen, die auf dem Arbeitsmarkt Probleme haben. In den 1980er Jahren nahm das CET an zwei randomisierten, kontrollierten Studien teil und war in beiden Erhebungen die einzige Institution, die Erfolge auf dem Arbeitsmarkt verzeichnen konnte. Auf Basis dieser vielversprechenden Ergebnisse in zwei unabhängigen randomisierten, kontrollierten Studien wiederholte das US-amerikanische Arbeitsministerium das CET-Programm an zwölf Standorten. Die Ergebnisse der Wiederholungsstudien waren enttäuschend. 54 Monate nach der randomisierten Zuweisung waren keine statistisch relevanten Ergebnisse zu erkennen. Eine wichtige Einschränkung der Wiederholungsstudie war die Tatsache, dass nur vier der 12 Standorte das CET-Modell originalgetreu übernahmen. Die Bewerter hatten unterschiedliche Hypothesen über die Gründe für die schlechten Ergebnisse im Vergleich zur ursprünglichen Studie, die jedoch nicht überprüft werden konnten.

In den 1970er Jahren führte Nathan Azrin eine Reihe von kleinen Experimenten durch, um festzustellen, ob die Stellensuche in der Gruppe im Rahmen von „Job-Clubs“ im Vergleich zu den herkömmlichen, individuellen Strategien zu verbesserten Arbeitsmarktergebnissen für Arbeitssuchende

führte. Azrin fand heraus, dass die Gruppenaktivitäten statistisch relevante, tiefgreifende Auswirkungen hatten. In seiner ersten Studie zeigte sich beispielsweise, dass innerhalb von zwei Monaten 90 % der Studienteilnehmer eine Arbeitsstelle gefunden hatten, in der Kontrollgruppe nur 55 %. Das US-amerikanische Arbeitsministerium förderte im Anschluss mehrere größere Experimente, die wiederum aussagekräftige Ergebnisse lieferten. Die Stellensuche in der Gruppe wird heute in großem Umfang genutzt.

Das Arbeitsministerium unterstützte in den 1980er Jahren drei Experimente, in denen ein Bewertungsprogramm zur Wiedereinstellung und Eignung getestet wurde, das Voraussetzungen für die Stellensuche für Antragsteller der Arbeitslosenversicherung sowie Hilfe bei der Stellensuche und Informationen zum Arbeitsmarkt beinhaltete, um die Teilnehmer bei der Stellensuche zu unterstützen. Sowohl in den ursprünglichen randomisierten, kontrollierten Studien in Charleston und South Carolina als auch in den Wiederholungsstudien in New Jersey und Washington wurde eine Reduktion der Anspruchsdauer um ca. eine halbe Woche festgestellt, ohne dass die Antragsteller Einbußen beim Einkommen hinnehmen mussten. Das Arbeitsministerium unterstützt heute ähnliche Programme in 40 Staaten. Außerdem haben auch zusätzliche Auswertungen anhand von randomisierten, kontrollierten Studien Kosteneinsparungen aufgezeigt.

Wiederholungsstudien für Arbeitsmarktprogramme haben sich als wichtiges Werkzeug für die Entwicklung von Arbeitsmarktstrategien erwiesen. In mehreren Fällen wurden die ursprünglichen Ergebnisse nicht bestätigt, und die Interventionen wurden eingestellt (in einigen Fällen liegt das Fehlen signifikanter Ergebnisse jedoch daran, dass die Wiederholungsstudien sich nicht nah genug an den Originalstudien orientierten) In anderen Fällen, z. B. bei Job-Clubs und Experimenten zur Stellensuche, zeigten die Wiederholungsstudien, dass ein bestimmter Interventionstyp sinnvoll ist, und der standardmäßige Dienstleistungspool für Arbeitssuchende wurde entsprechend angepasst.

References

Ashenfelter, O., Ashmore, D., Olivier, D.: Do unemployment insurance recipients actively seek work? Evidence from randomized trials in four US states. J. Econom. **125**, 53–75 (2005)

Azrin, N.H., Flores, T., Kaplan, S.J.: Job-finding club: a group-assisted program for finding employment. Behav. Res. Ther. **13**, 17–27 (1975)

Azrin, N.H., Philip, R.A.: The job club method for the handicapped: a comparative outcome study. Rehabil. Couns. Bull. **23**, 144–155 (1979)

Azrin, N.H., Philip, R.A., Thienes-Hontos, P., Besalel, V.A.: Comparative evaluation of the job club program with welfare recipients. J. Vocat. Behav. **16**, 133–145 (1980)

Benus, J., Poe-Yamagata, E., Wang, Y., Blass, E.: Reemployment and eligibility assessment (REA) study: FY 2005 initiative final report. Columbia, MD: IMPAQ International (2008)

Bobonis, G.J., Miguel, E., Charma, C.P.: Anemia and school participation. J. Hum. Resour. **41**, 692–721 (2006)

Burghardt, J., Rangarajan, A., Gordon, A., Kisker, E.: Evaluation of the Minority Single Parent Demonstration Vol. I: Summary report. Princeton, NJ: Mathematica Policy Research (1992)

Burtless, G.: The work response to a guaranteed income: a survey of experimental evidence. In: Munnell, A.H. (ed.) Lessons from Income Maintenance Experiments. Federal Reserve Bank of Boston, Boston (1986)

Cave, G., Bos, H., Doolittle, F., Toussaint, C.: Jobstart: Final Report on Program for School Dropouts. Manpower Research Demonstration Corporation, New York (1993)

Corson, W., Haimson, J.: The New Jersey unemployment insurance reemployment demonstration project. Washington, DC: US Department of Labor, Employment and Training Administration, Unemployment Insurance, occasional paper 95-2, revised edition (1996)

Corson, W.A., Long, D., Nicholson, W.: Evaluation of the Charleston claimant placement demonstration. Washington, DC: US Department of Labor, Employment and Training Administration, Unemployment Insurance, occasional paper 85-2 (1985)

Corson, W.A., Spiegelman, R.G.: Introduction and background of the reemployment bonus experiments. In: Robins, P.K., Spiegelman, R.G. (eds.) Reemployment Bonuses in the Unemployment Insurance System: Evidence from Three Field Experiments. W.E. Upjohn Institute for Employment Research, Kalamazoo (2001a)

Corson, W.A., Spiegelman, R.G.: Design of three field experiments. In: Robins, P.K., Spiegelman, R.G. (eds.) Reemployment Bonuses in the Unemployment Insurance System: Evidence from Three Field Experiments. W.E. Upjohn Institute for Employment Research, Kalamazoo (2001b)

Greenberg, D., Linksz, D., Mandell, M.: Social Experimentation and Public Policymaking. The Urban Institute Press, Washington (2003)

Hahn, A., Leavitt, T., Aaron, P.: Evaluation of the Quantum Opportunities Program: Did the Program Work? Brandeis University, Heller Graduate School, Waltham (1994)

Hamilton, G., Scrivener, S.: Increasing employment stability and earnings for low-wage workers: Lessons from the employment retention and advancement (ERA) project. OPRE report 2012-19. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, US Department of Health and Human Services (2012)

Hollister, R.: New Evidence About Effective Training Strategies. Rockefeller Foundation, New York (1990)

Johnson, T.R., Klepinger, D.H.: Evaluation of the impacts of the Washington alternative work search experiment. occasional paper 91-4, Washington, DC: US Department of Labor, Employment and Training Administration, Unemployment Insurance (1991)

Kremer, M., Miguel, E.: The illusion of sustainability. Q. J. Econ. **122**, 1007–1065 (2007)

Manpower Demonstration Research Corporation Board of Directors: Summary and Findings of the National Supported Work Demonstration. Ballinger Publishing Company, Cambridge (1980)

Maxfield, M., Schirm, A., Rodriguez-Planas, N.: The Quantum Opportunity Program Demonstration: Implementation and Short-Term Impacts. Mathematica Policy Research, Washington (2003)

Meyer, B.D.: Policy Lessons from the US unemployment insurance experiments. National Bureau of Economic Research (1992). Working paper no. 4197

Meyer, B.D.: Lessons from the US unemployment insurance experiments. J. Econ. Lit. **33**, 91–131 (1995)

Miguel, E., Kremer, M.: Worms: identifying impacts on education and health in the presence of treatment externalities. Econometrica **72**, 159–217 (2004)

Miller, C., Bos, J.M., Porter, K.E., Tseng, F.M., Doolittle, F.C., Tanguay, D.N., Vencil, M.P.: Working with Disadvantaged Youth:

Thirty-Month Findings from the Center for Employment Training Replication Sites. MDRC, New York (2003)
Miller, C., Bos, J.M., Porter, K.E., Tseng, F.M., Abe, Y.: The Challenge of Repeating Success in a Changing World: Final Report on the Center for Employment Training Replication Sites. MDRC, New York (2005)
Orr, L.L., Bloom, H.S., Bell, S.H., Doolittle, F., Lin, W., Cave, G.: Does Training for the Disadvantaged Work? Evidence from the National JTPA Study. Urban Institute Press, Washington (1996)
Poe-Yamagata, E., Benus, J., Bill, N., Carrington, H., Michaelides, M., Shen, T.: Impact of the reemployment and eligibility assessment (REA) initiative. ETA occasional paper 2012-08, Washington, DC: US Department of Labor, Employment and Training Administration (2011)
Robins, P.K., Spiegelman, R.G. (eds.): Reemployment Bonuses in the Unemployment Insurance System: Evidence from Three Field Experiments. W.E. Upjohn Institute for Employment Research, Kalamazoo (2001)
Rosholm, M.: The importance of meeting your caseworker: evidence from several field experiments (2012). Denmark: Aarhus University Department of Economics and Business. Retrieved 4/27/2013 from http://doku.iab.de/veranstaltungen/2012/field_2012_Rosholm_pres.pdf
Schorr, L.B., Farrow, F.: Expanding the evidence universe: doing better by knowing more (2011). Washington, D.C.: Center for the Study of Social Policy. Retrieved September 24, 2012, from http://www.cssp.org/publications/harold-richman-public-policy-symposium/Expanding-Evidence-the-Evidence-Universe_Doing-Better-by-Knowing-More_December-2011.pdf
Wandner, S.A.: Solving the Reemployment Puzzle. W.E. Upjohn Institute for Employment Research, Kalamazoo (2010)
Wolfhagen, C., Goldman, B.S.: Job Search Strategies: Lessons from the Louisville WIN Laboratory. Manpower Demonstration Research Corporation, New York (1983)
Woodbury, S.A., Spiegelman, R.G.: Bonuses to workers and employers to reduce unemployment: randomized trials in Illinois. Am. Econ. Rev. **77**, 513–530 (1987)
Zeidenberg, M., Cho, S., Jenkins, D.: Washington State's Integrated Basic Education and Skills Training Program (I-BEST): New evidence of effectiveness. New York: Community College Research Center, Teachers College, Columbia University, CCRC working paper no. 20 (2010)

Burt S. Barnow is the Amsterdam Professor of Public Service and of Economics at the Trachtenberg School of Public Policy and Public Administration at George Washington University. Dr. Barnow has over 30 years of experience as an economist conducting and managing research in the fields of workforce investment, program evaluation, performance management, and labor market analysis. Prior to coming to George Washington University, Dr. Barnow was Associate Director for Research at Johns Hopkins University's Institute for Policy Studies, where he worked for 18 years. Prior to that, he worked for 8 years at the Lewin Group and 9 years at the US Department of Labor, including 4 years as Director of the Office of Research and Evaluation in the Employment and Training Administration. Prior to these positions, he was an assistant professor of economics at the University of Pittsburgh. He has a B.S. degree in economics from the Massachusetts Institute of Technology and M.S. and Ph.D. degrees in economics from the University of Wisconsin at Madison.

David Greenberg is professor of economics emeritus at the University of Maryland–Baltimore County (UMBC). Before coming to UMBC in 1982, he worked for the Rand Corporation, SRI International, and the US Department of Health and Human Services. Much of his research focuses on social experiments and the evaluation of government programs targeted at the low-income population, especially public assistance, employment, and training programs. He has written a guide for conducting and using cost analyses of employment and training programs, and he is the coauthor of the *Digest of Social Experiments, Third Edition* (Urban Institute Press) and a textbook on cost-benefit analysis. He has a B.A. degree in economics from Southern Methodist University and a Ph.D. degree in economics from the Massachusetts Institute of Technology.