Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments

Howard S. Bloom
Carolyn J. Hill
James A. Riccio

Abstract

This paper addresses the question: How does implementation influence the effectiveness of mandatory welfare-to-work programs? Data from three large-scale, multi-site random assignment experiments were pooled; quantitative measures of program implementation were constructed; and multilevel statistical modeling was used to examine the relationship between program implementation and effects on short-term client earnings. Individual-level data were analyzed for 69,399 sample members and group-level implementation data were analyzed for 59 local programs. Findings indicate that, other things being equal, earnings effects are increased by: an emphasis on quick client employment, an emphasis on personalized client attention, staff caseloads that do not get too large, and limited use of basic education. Findings also show that mandatory welfare-to-work programs can be effective for many types of people, and that focusing on clients who are especially job-ready (or not) does not have a consistent influence on a program's effectiveness. © 2003 by the Association for Public Policy Analysis and Management

INTRODUCTION

Over the past few decades, in large part due to the increased use of random assignment experiments, important advances have been made in building a convincing body of evidence about the effectiveness of social programs. Furthermore, knowledge about program implementation—how social policies and program models operate in the real world—has been expanded greatly by researchers. Meanwhile, much less has been learned about the relationship between implementation and effectiveness.

This disconnect is partly explained by the fact that implementation studies often do not include impact analyses. Thus, although they provide a wealth of information about leadership strategies, operating practices, and patterns of client involvement, they cannot determine how these factors affect client impacts.[1] The prevailing disconnect is also due to the fact that, although most random assignment impact studies include implementation analyses, they seldom have enough sites for a mean-

[1] We are not implying that all implementation studies should analyze the connections to program effects. Examples of informative implementation studies that are not explicity designed to look at impacts include Behn (1991); Brodkin (1997); Hagen and Lurie (1994); Mead (1986); and Meyers, Glaser, and MacDonald (1998).

Manuscript received August 2002; out for review August 2002; review completed November 2002; accepted January 2003.

Journal of Policy Analysis and Management, Vol. 22, No. 4, 551–575 (2003)
© 2003 by the Association for Public Policy Analysis and Management
Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/pam.10154

ingful statistical analysis of how implementation and impacts are related. Thus, researchers can only speculate about which of the many implementation differences they observed caused the variation in impacts they documented.

The present paper presents preliminary findings from a research synthesis study to help bridge this gap. The study pools data from a series of large-sample, multi-site experiments of mandatory welfare-to-work programs conducted by MDRC during the past 15 years. These experiments, which reflect the experiences of 69,399 sample members from 59 locations in seven states, have the unusual (if not unique) virtue of including consistently measured indicators of program implementation for all sites.[2]

Multi-site experiments that measure program implementation and effects in a consistent manner offer a powerful way to "get inside the black box" of social programs to explore why programs perform as well—or as poorly—as they do (Greenberg, Meyer, and Wiseman, 1994; Raudenbush and Liu, 2000). And pooling data across such experiments is a promising way to obtain enough sites for a meaningful research synthesis.[3] However, in order to create opportunities for such a synthesis, an adequate supply of appropriate studies must be built up over time. Thus, in this paper we try to illustrate why researchers and funders from different policy domains and program areas should consider developing long-term research agendas that use random assignment experiments with strong implementation studies and comparable measures and methods to accumulate knowledge about "what works best for whom."

In the following sections we discuss the theoretical and policy background for our research, and describe the programs, samples, data, and analytic framework upon which it is based.[4] Our findings illustrate the profound influence that implementation can have on the effectiveness of social programs.

DIMENSIONS OF IMPLEMENTATION

Policymakers, program administrators, program staff members, and researchers have debated for decades the influence of four sets of implementation-related factors on the effectiveness of welfare-to-work programs: (1) how programs are managed and what frontline practices they emphasize, (2) the kinds of activities that clients participate in, (3) the economic environment in which programs operate, and (4) the characteristics of clients that they serve. Our analysis tests the following hypotheses that grow out of these debates.

Management Choices and Frontline Practice

Many experts contend that how frontline workers interact with clients and the social climate or institutional culture within which they interact can be powerful determinants of a program's success (Bane, 1989; Behn, 1991; and Mead, 1983, 1986).[5] Yet, when it comes to which frontline practices work best, expert opinion

[2] Earlier efforts to explore the statistical relationships between implementation and impacts using a smaller set of welfare-to-work sites were conducted by Riccio and Hasenfeld (1996) and Riccio and Orenstein (1996).

[3] For a similar analysis using a smaller sample of Job Training Partnership Act (JTPA) sites, see Heinrich (2002).

[4] For further background on this study, see Bloom, Hill, and Riccio (2001) and Riccio, Bloom, and Hill (2000).

[5] For example, based on her review of existing research and first-hand experience as a state welfare administrator, Bane (1989, p. 287) argues that the central challenge in building effective programs is "... how to shape an organizational culture that ... delivers a clear message that the goal is jobs, sets a clear expectation that clients can get jobs and that workers are obligated to make that happen, monitors performance, and provides necessary resources."

often differs because of limited evidence. We examine the influence of several such debated practices.

Quick Employment

Emphasis on quick employment reflects how forcefully a program urges its clients to move quickly into jobs—even low-paying ones. Advocates of this approach believe that almost any job is a positive first step that can promote future advancement through the acquisition of work experience, job-related skills, and a track record of employment. Opponents believe that welfare recipients would do better by increasing their human capital through education and training so that they can qualify for better jobs before looking for work.

Programs can manifest these contrasting philosophies in many ways. For example, the prevailing philosophy can be evident in the initial activity that is encouraged or mandated: job search versus education or training. In addition, it can be reflected by how strongly job search activities stress rapid employment instead of holding out for a better job. Furthermore, it can be reflected by how long participants who are assigned to education or training are allowed to wait before looking for work. Therefore, efforts (or lack of efforts) to promote rapid employment can pervade staff interactions with clients regardless of the program activities to which they are assigned.[6]

In examining this issue, several past random assignment experiments have found that counties with the largest employment impacts were places where, among other things, staff strongly encouraged quick client employment (Hamilton et al., 2001; Riccio, Friedlander, and Freedman, 1994). Hence, there is some prior evidence on this issue.

Personalized Client Attention

A second practice that we examine is the extent to which frontline staff get to know their clients' personal situations, needs, and goals; arrange services that support these needs and goals; continue to work with clients over time to assure their success; and adjust client service plans to meet their changing situations. Many managers believe that such a personalized approach is more effective than one in which clients are handled in a narrowly prescribed way (where "one size fits all"). They emphasize that "getting close to the customer" is key to properly addressing clients' aspirations and situations. Others see less payoff from this investment of time and scarce program resources. Although this issue has not been widely studied, one past random assignment experiment casts doubt on the benefits of increased personalized attention (Riccio and Orenstein, 1996).

Closeness of Client Monitoring

Monitoring clients' participation in mandatory welfare-to-work programs is believed by many to be important for enforcing participation requirements and for

[6] For example, Riccio, Friedlander, and Freedman (1994, p. xxv) described efforts in California's Riverside County GAIN program, which epitomized the quick employment philosophy, "to communicate a strong 'message' to all registrants..., at all stages of the program, that employment was central, that it should be sought expeditiously, and that opportunities to obtain low-paying jobs should not be turned down."

helping clients to get the most from a program. Careful monitoring can help staff learn whether clients are showing up for their assigned activities and whether they are progressing in them. Based on what staff learn, they may start formal enforcement proceedings if participation obligations are being ignored; initiate assistance with personal problems or circumstances that might be interfering with clients' progress; or consider alternative client services and activities.[7] Thus, monitoring may contribute to a program's performance in various ways. However, close monitoring can be difficult for programs where employment-related activities occur at many different institutions or locations, and this can be especially problematic when tracking systems are deficient (Freedman et al., 2002; Riccio, Friedlander, and Freedman, 1994; Wiseman, 1987).

Consistency of Staff Views

Program performance may suffer when staff members are divided—whether due to confusion or disagreement—over what a program should be doing or how it should be done. Thus, it is frequently hypothesized that managers can improve program performance by focusing staff efforts on a common purpose and instilling in them a strong organizational culture (Behn, 1991; Miller, 1992; Nathan, 1993).

Size of Staff Caseload

It is often argued that large caseloads prevent program staff members from spending enough time with their clients to be effective (Gueron and Pauly, 1991). The only direct evidence on this issue is from one small-scale experiment (Riccio, Friedlander, and Freedman, 1994). It compared client outcomes for a standard caseload, averaging 97 clients per worker, and a reduced caseload, averaging 53 clients per worker. No differences were found.

Program Activities

The three main types of job-related program activities tested here are job search assistance, basic education, and vocational training. The relative effectiveness of these activities has been debated for many years as part of a broader philosophical controversy over how best to promote economic self-sufficiency for welfare recipients.

Job search assistance has been a staple of welfare-to-work programs since their inception. However, during the past two decades, federal and state governments have begun to invest more heavily in basic reading and math classes, preparation for the General Educational Development (GED) degree, and courses in English as a Second Language (ESL). To a lesser extent, they also have begun to invest more in vocational training.

Other activities provided by some welfare-to-work programs include unpaid work experience positions through which clients work at public or not-for-profit jobs in exchange for their welfare grants ("workfare"), and enrollment in four-year colleges or associate's degree programs at community colleges. Neither of these activities is highly prevalent, however.

[7] Unfortunately, we do not have a consistent measure of the degree of enforcement across all the offices in this study.

Findings from an early welfare-to-work experiment in California cast doubt on the efficacy of immediately assigning welfare recipients with especially weak educational backgrounds to basic education classes (Freedman et al., 1996; Riccio, Friedlander, and Freedman, 1994).[8] Partly in response to this finding, a later six-state welfare-to-work experiment directly compared (in three of its sites) a labor force attachment approach (which emphasized job search first) against a human capital development approach (which emphasized education and training first) (Hamilton, 2002; Hamilton et al., 2001). This study did not find the anticipated advantage of human capital development over 5 years of follow-up.[9] However, both prior studies found that programs with the largest earnings impacts were ones that took a mixed approach, allowing clients to choose education and training or job search as their initial activity.

Local Economic Conditions

That local economic conditions can affect the performance of welfare-to-work programs seems almost self-evident. Nevertheless, there are two competing views about the likely direction of this effect. One view is that program performance will be better when unemployment rates are lower because low unemployment rates imply more jobs for clients. Thus, if a program can motivate and prepare clients for these jobs, it can increase their employment appreciably.

An opposing view is that programs perform less well when unemployment rates are lower because it is easier for welfare recipients to find jobs without extra help. Thus, even though a program may have a high placement rate, it might be producing little success beyond what would have been achieved in its absence. Furthermore, welfare recipients who are not working when unemployment rates are low may be especially hard to employ, thus making it particularly difficult for programs to increase employment for them.

The empirical evidence on this issue is limited because few past studies have been able to compare site-level impact estimates from random assignment experiments to measures of local economic conditions. An exception is Heinrich (2002), who found a negative but statistically insignificant relationship between the local unemployment rate and program impacts in JTPA. Other attempts to measure the relationship between local economic environment and site impacts (for example, Riccio and Orenstein, 1996), are based on small numbers of sites, which limits their ability to control for other local differences.

Client Characteristics

Programs that serve different types of clients may have greater or lesser success not because of what they do, but, rather, because of whom they are trying to serve. Thus,

[8] Counties in that evaluation that most strongly emphasized basic education did not produce the consistently larger earnings impacts for that subgroup of clients, and some had no statistically significant effect on their earnings at all over a 5-year follow-up period.

[9] This study found that among clients who did *not* have a high school diploma or GED at the time of random assignment, the labor force attachment strategy had larger impacts on earnings over the 5-year follow-up period than did the human capital development strategy, which emphasized basic education activities. In contrast, for sample members who entered the study with a high school credential the human capital development strategy, which emphasized vocational training or post-secondary education for this subgroup, was about as effective as the labor force attachment approach-albeit substantially more expensive (and, therefore, less cost-effective).

understanding variation in program effectiveness requires taking into account cross-program variation in client characteristics that reflect their employment potential and employment barriers.[10] The most widely used indicators of this client characteristic are formal education, prior employment experience, and past welfare receipt. Formal education and prior employment experience represent individual human capital; past welfare receipt predicts future reliance on welfare. Other indicators include race and ethnicity (to reflect potential labor market discrimination), number and age of children (to reflect alternative demands on clients' time), and physical and mental health status (to reflect clients' abilities to participate in the labor market).

The limited research that exists on this issue suggests that there is no simple relationship between program impacts and client characteristics. Some evidence indicates that many welfare-to-work programs of the mid-1980s and 1990s were effective with a broad range of clients (Michalopoulos, Schwartz, with Adams-Ciardullo, 2001). At the same time, some subgroups fared much better than others in certain programs. These findings suggest that it is important to incorporate client characteristics into any analysis of program effectiveness.

PROGRAMS, SAMPLES, AND DATA

Our analysis is based on three MDRC random assignment evaluations of mandatory welfare-to-work programs: the Greater Avenues for Independence (GAIN) program conducted in 22 local offices in six California counties (Riccio and Friedlander, 1992), Project Independence (PI) conducted in 10 local offices in nine Florida counties (Kemple and Haimson, 1994), and the National Evaluation of Welfare-to-Work Strategies (NEWWS) conducted in 27 local offices in 10 counties in California, Georgia, Michigan, Ohio, Oklahoma, and Oregon (Hamilton, 2002; Hamilton et al., 2001). These initiatives were operated as each state's version of the federal Job Opportunities and Basic Skills Training (JOBS) program funded by the Family Support Act of 1988.

The programs studied through GAIN, PI, and NEWWS included varying mixes of work-promoting activities such as job-search assistance, basic education, and vocational training. They also provided clients with support services such as childcare and transportation. Clients were assigned to local staff members who arranged for them to attend program activities, helped them gain access to support services, and monitored their participation and progress. Participation in the programs was mandatory, and failure to attend assigned activities without "good cause" could result in reduction of a family's welfare grant. The original reports from these evaluations document in detail how programs were implemented and evaluated.

Each evaluation measured program impacts on client employment, earnings, and welfare receipt by comparing post-random-assignment outcomes for individuals randomly assigned to the program with those for individuals randomly assigned to a control group. Program group members were required to enroll in the program. Control group members were exempted from these requirements and excluded

[10] Understanding how program performance varies with client characteristics is also important for wisely targeting program resources and setting performance standards. Program resource targeting decisions are usually based on two criteria—equity and efficiency. Equity concerns lead to targeting in accord with clients' need for assistance. Efficiency concerns lead to targeting in accord with clients' ability to benefit. Evidence that client characteristics actually do influence program impacts would encourage giving higher priority to serving individuals who are most likely to benefit from the program and establishing performance standards that took the composition of the program's caseload into account.

from program activities and services. However, they could seek assistance from other sources.

By randomly determining which sample members were assigned to the program and which were assigned to control status, each evaluation created two groups that in large samples are comparable in all ways.[11] Hence future outcomes for the control group provide valid indicators of what these outcomes would have been for the program group without the program. In other words, they identify the program group counterfactual. The difference between the actual experience of program group members and their counterfactual is a valid measure of the impact of the program (that is, how the program changed the outcome).

For the present study we pooled comparable data from these three evaluations, yielding a sample of 69,399 program and control group members from 59 local welfare-to-work offices. Although some of the original sample members were men, we focus only on female single parents to create a homogeneous sample. This sample includes 46,977 women from 27 local NEWWS offices, 18,126 women from 22 local GAIN offices, and 4,296 women from 10 local PI offices. Sample members per local office range from 177 to 4,418 and average 1,176.

Intake forms provide information on all program and control group members' socioeconomic backgrounds, which we used to measure client characteristics.

Administrative records from state unemployment insurance (UI) agencies provide data on all program and control group members' quarterly earnings during the first 2 years after random assignment, which we used to measure client outcomes.

Staff surveys of 1,225 caseworkers and 194 supervisors from local program offices provide data on how programs were run, which we used to measure their implementation. The number of respondents per office ranged from 1 to 83 caseworkers and from 0 to 14 supervisors, and averaged 21 and 3 per office, respectively. Completion rates exceed 90 percent for most offices.

Follow-up surveys of a random subsample of 15,235 program and control group members provide data that we used to measure participation in employment-related activities sponsored by the program or other local organizations. Subsamples range from 27 to 2,159 individuals per office and average 258. Response rates range from 70 to 93 percent across counties in the studies.

County-level statistics on unemployment rates from the U.S. Bureau of Labor Statistics and California Employment Development Department were used to measure the local economic environment.

VARIATION IN PROGRAM EFFECTS

The dependent variable—the measure of local program effectiveness—is the estimated impact on sample members' mean total earnings (measured in constant 1996 dollars) for the first 2 years after random assignment.[12]

The average program increased clients' earnings by $879 during their 2-year follow-up period, or by 18 percent of what these earnings would have been otherwise. This is a sizable impact relative to those documented for other welfare-to-work programs (Gueron and Pauly, 1991).

[11] Strictly speaking, randomization produces groups whose expected values are equal for all variables. The larger the sample randomized the more closely it approximates this ideal. Given the especially large samples for the present analysis, the program and control groups were quite similar.

[12] Earnings were expressed in constant 1996 dollars using the CPI-U (Economic Report of the President, 2000).

More importantly for the present analysis, the variation across offices in unconditional[13] impact estimates is large (see Figure 1), ranging from -$1,412 to $4,217. Of the office-level impact estimates, 13 are negative, although not statistically significant at conventional levels. In contrast, 24 of the positive impact estimates are statistically significant, and many are quite large. Hence, the variation in impact estimates across local offices is highly statistically significant (beyond the 0.001 level).

Overall, these results show that there is plenty of impact variation to model, and that this variation reflects true differences in program impacts, not just estimation error.[14] The magnitude of the variation also underscores the importance from a policy perspective of trying to account for why some offices performed so much better than others.

MULTILEVEL MODELS USED TO EXPLAIN VARIATION IN EXPERIMENTAL IMPACT FINDINGS

To explore what created this variation in effects, we attempt to isolate the independent influences on it of the implementation factors described above. This requires accounting for the multilevel structure of the data in which sample members (level 1) are grouped by local program office (level 2). To do so we estimate a two-level hierarchical model (Raudenbush and Bryk, 2002).[15]

Level 1 of the model comprises a linear regression for individual sample members, which serves two purposes. It indicates how client characteristics influence program impacts, which is of direct substantive interest. And it produces estimates of conditional impacts for each program office (holding client characteristics constant), which is the main dependent variable for level 2.

Three linear regressions for local program offices compose level 2 of the model. The first regression represents how conditional program effects depend on program implementation, activities, and environment. The parameters of this regression are our core findings. The second regression represents how the conditional mean control group outcome for each office (holding client characteristics constant) varies with local economic environment. This provides an estimate of the counterfactual for each office. The third regression (which is an artifact of how the original experiments were conducted) accounts for the fact that several sites changed the ratio of their program and control groups over time.[16]

By estimating these relationships as a two-level model, all parameters at both levels are determined simultaneously and each parameter represents the influence of a single factor holding constant the others.[17] For example, estimates of the influence of clients' prior education on program effectiveness hold constant all other client characteristics in level one plus all factors in level 2. Likewise, estimates of the influence of a particular feature of program implementation hold constant all

[13] Unconditional impacts refer to impacts estimated without controlling for cross-office variation in measured client characteristics. They are distinguished from conditional impacts discussed later, which do control for these characteristics.

[14] Raudenbush and Bryk (2002, pp. 63–64) describe the $\chi2$ test that we used to assess the statistical significance of the variation in program impacts across local offices.

[15] Hierarchical models—also called random effects models, mixed models, or variance component models—are a major advance in the analysis of data where observations are grouped within aggregate units such as students in schools, employees in firms, residents in neighborhoods, and clients in programs.

[16] See Bloom, Hill, and Riccio (2001, p. 23) for further details.

[17] This is accomplished through a combination of maximum likelihood and weighted least squares procedures (Raudenbush and Bryk, 2002).

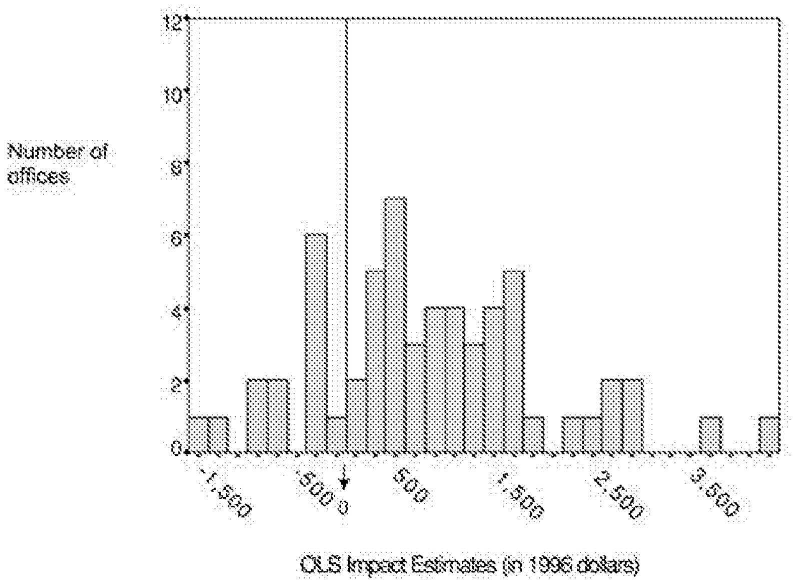

Figure 1. The distribution of unconditional office-level impact estimates.

other factors in level 2 plus all client characteristics in level 1. Our model was specified as follows.

Level 1, for sample members:

$$Y_{ji} = \alpha_j + \beta_j P_{ji} + \sum_k \delta_k CC_{kji} + \sum_k \gamma_k CC_{kji} P_{ji} + \kappa_j RA_{ji} + \varepsilon_{ji} \qquad \text{(Eq. 1)}$$

where:

Y_{ji} = the outcome measure for each sample member,
P_{ji} = a zero/one program group indicator for each sample member,
CC_{kji} = client characteristic k for each sample member (grand mean centered),[18]
RA_{ji} = a zero/one random assignment cohort indicator for each sample member,
α_j = the conditional control group mean outcome (counterfactual) for each local office,
β_j = the conditional program impact for each local office,
δ_k = the effect of client characteristic k on the control group mean outcome,
γ_k = the effect of client characteristic k on the program impact,
κ_j = a random assignment cohort coefficient (which has no substantive meaning) for each office,
ϵ_{ji} = a random component of the outcome for each sample member.

Level 2, for local offices:

$$\beta_j = \beta_0 + \sum_m \pi_m PI_{mj} + \sum_n \phi_n PA_{nj} + \psi EE_j + \mu_j \qquad \text{(Eq. 2)}$$

[18] Raudenbush and Bryk (2002) pp. 31–35 describe how different ways of centering variables affect the interpretation of their coefficients in a hierarchical model.

$$\alpha_j = \alpha_o + \lambda EE_j + \upsilon_j \qquad \text{(Eq. 3)}$$

$$\kappa_j = \kappa_0 + \eta_j \qquad \text{(Eq. 4)}$$

where:

PI_{mj} = program implementation feature m for each office (grand mean centered),
PA_{nj} = program activity n for each office (grand mean centered),
EE_j = economic environment for each office (grand mean centered),
β_0 = the grand mean impact,
π_m = the effect of program implementation feature m on program impacts,
ϕ_n = the effect of program activity n on program impacts,
ψ = the effect of economic environment on program impacts,
μ_j = a random component of program impacts for each office.
and
α_o = the grand mean control group earnings,
λ = the effect of economic environment on control group earnings,
υ_j = a random component of control group mean earnings for each office,
and
κ_0 = the grand mean random assignment cohort coefficient, and
η_j = a random component of the cohort coefficient for each office.

CONSTRUCTION AND ASSESSMENT OF THE IMPLEMENTATION MEASURES

The primary independent variables for each local office are measures of program implementation constructed from average staff survey responses.[19] Hence these measures reflect the perceptions of frontline workers. Survey questions were based on hypotheses about what works, drawn from the research literature and experience in the field. The first staff survey (for GAIN) was developed by MDRC and used by Riccio and Hasenfeld (1996) and Riccio and Orenstein (1996) to explore operational correlates of program impacts. This analysis was complemented by in-depth fieldwork to document local practices in order to better understand what was happening on the ground. Later surveys (for PI and NEWWS) evolved to reflect changing views about best practices. However, a common core of questions on issues that remain in the forefront of programmatic discussions was maintained. It is from these questions that we constructed our six measures of program implementation. Table 1 lists the questions used to do so.

Types of Measures

Program Practices

Three measures of program practices are multi-question scales that were standardized to have mean values of zero and standard deviations of one across

[19] Bloom, Hill, and Riccio (2001, p. 88) describe how these averages were regression-adjusted to control for office differences in staff characteristics. Although these adjustments were minimal they help to hold constant differences that may exist in the perceptions of different types of staff members.

Table 1. Staff survey questions for scales of program implementation.

Scale and Questions[a]
Emphasis on quick job entry for clients
• Does your unit emphasize helping clients build basic skills, or moving them quickly into jobs?
• Should your unit emphasize helping clients build basic skills, or moving them quickly into jobs?
• What would be your personal advice to a client who can either take a low-skill, low-paying job OR stay on welfare and wait for a better opportunity?
• What advice would your supervisor want you to give to such a client?
Emphasis on personalized client attention
• Does your program emphasize the quality of its services more than the number of clients it serves?
• During intake, does your unit spend enough time with clients?
• During intake, do staff make an effort to learn about clients' family problems?
• During intake, do staff make an effort to learn about clients' goals and motivation to work?
• How well is your program tailoring services to clients' needs?
Closeness of client monitoring
• How closely are staff monitoring clients?
• If a client has been assigned to adult basic education but has not attended, how soon would staff find out?
• If a client has been assigned to vocational education but has not attended, how soon would staff find out?
• How closely is your agency monitoring whether clients quit or lose part-time jobs?
• Once your agency learns a client lost a part-time job, how soon would she be assigned to another activity?
Staff caseload size
• How many clients are on your caseload today?

[a] The questions in this table paraphrase each staff survey question. Response categories generally took the form of a 5-point or 7-point Likert scale.

offices. A first measure—emphasis on quick job entry for clients—reflects the employment message conveyed to clients at each office—how much they were encouraged to take a job quickly, or to be more selective and wait for a better job or pursue education and training to improve their future prospects. A second measure—emphasis on personalized client attention—reflects the emphasis placed by each office on gaining in-depth understanding of clients' personal histories and circumstances to better accommodate their individual needs and preferences when making program assignments. A third measure—closeness of client monitoring—reflects how closely staff at each office tracked client participation in assigned activities to keep abreast of their progress, their changing needs, and their involvement in the program.

A fourth measure of program practices—staff caseload size—reflects the average number of clients for which each frontline staff member was responsible. This measure ranged across offices from 70 to 367 clients per staff member and averaged 136.

A fifth measure of program practice—staff disagreement—indicates the variability within each office of frontline staff responses to the first three sets of questions, while a sixth measure—frontline staff and supervisor disagreement—indicates the difference between average frontline staff and supervisor answers to these questions. Hence, these final two measures (which also were standardized to have means of zero and standard deviations of one) reflect the degree to which each office had a common vision of its program.

Participation in Activities

The next set of office-level independent variables concerns participation in employment-related activities. In constructing these measures, we take into account the fact that control group members often can obtain similar services in their communities without any help from the welfare-to-work program that is being evaluated (Gueron and Pauly, 1991). Differences in program and control group participation rates are estimated for the most common activities: job-search assistance, basic education, and vocational training.[20]

Job-search assistance includes self-directed individual job search and participation in group-focused job clubs. Basic education includes adult education classes, GED preparation, and ESL courses. Vocational training includes classroom training in basic occupational skills along with several less commonly used activities: on-the-job training, unpaid work experience, and post-secondary education or training. Differences in client participation rates, expressed in percentage points, represent the degree to which local programs increased exposure to each type of activity. These measures were constructed from follow-up survey data obtained roughly two years after random assignment for a random subsample of clients from each office. Appendix Table A.1 shows descriptive statistics for these estimated differences.

Economic Environment

The final office-level independent variable, representing the economic environment for each local office, was the average county unemployment rate during the client follow-up period for the office. Because sample enrollment often took several years at an office, its unemployment rate is an average of different periods for different sample members.

Client Characteristics

Independent variables for sample members include their education, number of children, age, race/ethnicity, recent past welfare receipt, and recent past earnings. These measures were constructed from data recorded on sample intake forms. Appendix Table A.2 shows descriptive statistics for these characteristics for the full sample, as well as across the 59 offices.

Variability, Reliability, and Validity

Before using the preceding independent variables we assessed them in terms of three requirements for a good measure: variability, reliability, and validity. Variabil-

[20] Bloom, Hill, and Riccio (2001, p. 91) describe how these participation differences were regression-adjusted to control for minor differences in the background characteristics of program and control group members at each office. This was done to increase the precision of program activity measures and to estimate them in a way that is consistent with the estimation of program impacts.

ity in a measure is required in order to determine its relationship with program impacts. For example, only if implementation varies across offices can one determine how impacts vary with implementation. Fortunately, the variation in eight of our ten office level measures is statistically significant at beyond the 0.001-level (Bloom, Hill, and Riccio, 2001, p. 109). The significance of variation for our staff disagreement measure could not be determined, and that for our staff/supervisor disagreement measure is low.[21] Nevertheless both measures were maintained because of their conceptual importance.

A reliable measure has a high signal-to-noise ratio. Thus variation in its values reflects systematic differences in the subjects being observed, not just random error. This is necessary for an independent variable in order to obtain precise and unbiased estimates of its relationship to a dependent variable.[22] An unreliable measure of program implementation would produce an imprecise and distorted view of the true relationship between implementation and impacts.

Our multi-question implementation measures—emphasis on quick job entry, emphasis on personalized attention, and closeness of monitoring—have two separate dimensions of reliability. One dimension, inter-question reliability (often called inter-item reliability), reflects the consistency of answers to different questions used to construct each scale. High inter-question reliability refers to a scale whose component questions are highly correlated. The second dimension, inter-respondent reliability, reflects the consistency of responses from different staff members at each office. High inter-respondent reliability refers to a scale whose values for different staff members at the same office are highly correlated. Fortunately, our measures are reliable in both regards, with inter-question reliability coefficients between 0.76 and 0.84 and inter-respondent reliability coefficients between 0.76 and 0.83.[23]

A valid measure is one that represents what it is intended to represent. Thus systematic variation in its values reflects true variation in its intended construct. This is necessary for an independent variable to obtain unbiased estimates of its relationship to a dependent variable. Thus an invalid measure of program implementation would produce a misleading view of the relationship between implementation and impacts.

Two types of validity were considered when assessing office-level independent variables: face validity and construct validity. Face validity is the degree to which the survey questions used for each independent variable appear, on their face, to convey the meaning that the variable is intended to convey. In other words, the components of the variable must accord with common sense. We believe that our variables meet this standard. Moreover, findings from the field research conducted across program offices for the original evaluations also support the face validity of the staff survey measures (Bloom, Hill, and Riccio, 2001).

Construct validity asks whether the measures correlate with each other in ways that would be expected for the office-level constructs they are supposed to represent. We find this to be the case.[24] For example, program emphasis on quick job entry is positively correlated with increased participation in job-search assistance and negatively correlated with increased participation in basic education or vocational training. In addition, program emphasis on personalized client attention and closeness of client monitoring are positively correlated with each other and

[21] This probably is because there were only one or two supervisors per office.
[22] See Greene (1993, pp. 435-440) for a discussion of this problem.
[23] See Bloom, Hill, and Riccio (2001, pp. 88-89) for further details.
[24] See Bloom, Hill, and Riccio (2001, pp. 94, 95, and 108) for further details.

negatively correlated with average staff caseload size. Furthermore local unemployment rates are negatively correlated with average control group earnings.

Thus it was possible to estimate our model with an office-level dependent variable that had substantial variation and a series of office-level independent variables that were variable, reliable, and valid.

FINDINGS

Table 2 presents estimates of how program implementation, activities, and environment affect program impacts. The regression coefficients in column one indicate the change in program impacts per unit change in each variable listed, holding constant all others in the model. The partially standardized regression coefficients in column two indicate the change in program impacts per standard deviation change in each variable, holding constant all of the others.[25] Note that the unstandardized and standardized regression coefficients are the same for some of our program implementation scales because these scales were defined to have a standard deviation equal to one. The *p*-values in column 3 of the table indicate the statistical significance of each coefficient estimate and the standard

Table 2. The effects of program implementation, activities, and environment on program impacts.

Program Characteristic	Regression Coefficient	Partially Standardized Regression Coefficient	Statistical Significance (*p*-value)	Standard Error
Program Implementation				
Emphasis on quick job entry	$ 720***	$ 720***	2 x 10^{-6}	$134
Emphasis on personalized service	428***	428***	0.0002	107
Closeness of monitoring	– 197	– 197	0.110	121
Staff caseload size	– 4***	– 268***	0.003	1
Staff disagreement	124	124	0.141	83
Staff/supervisor disagreement	– 159 *	– 159*	0.102	96
Program Activities				
Basic education	– 16 **	– 208 **	0.017	6
Job search assistance	1	12	0.899	9
Vocational training	7	71	0.503	11
Economic Environment				
Unemployment rate	– 94 ***	– 291***	0.004	30

Regression coefficients are reported in 1996 dollars per unit change in each independent variable. Partially standardized regression coefficients are reported in 1996 dollars per standard deviation change in each independent variable. These coefficients are estimated simultaneously with those reported in Table 3. The grand mean impact is $879 or 18 percent of the counterfactual. Statistical significance is indicated by * for the 0.10-level, ** for the 0.05-level and *** for the 0.01-level.

[25] The partially standardized regression coefficient equals the original coefficient multiplied by the standard deviation of the independent variable that it represents.

errors in column four indicate their precision. These findings indicate the following important points.

A Strong Employment Message Is a Powerful Medium

The emphasis placed by programs on quick client employment has by far the largest, most statistically significant, and most robust effect on program impacts of all that we observed. Both the unstandardized and standardized coefficients for this multi-question survey scale indicate that when it increases by one unit (which by definition equals one standard deviation of its distribution across local offices), program impacts increase by $720, other things equal. To place this in perspective, recall that the grand mean program impact is $879. This is the estimated impact when all variables in the model are at their mean values. If the quick employment scale increases by one unit and all other variables remain constant, the estimated program impact increases to $1,599. In percentage terms this represents an increase from 18 percent of the average counterfactual to 33 percent.[26] The third column in the table indicates that the *p*-value for the quick client employment coefficient estimate equals 2×10^{-6}, which is well beyond conventional levels of statistical significance.

In addition, sensitivity tests demonstrate that this finding is highly robust to a wide range of variation regarding which program offices are included in the analysis.[27] For example, the finding did not change materially when as many as 10 local offices with the most extreme program impacts (the dependent variable) or the most extreme emphases on quick employment (the independent variable) were eliminated from the analysis. Hence the finding does not appear to be confined to a small group of sites.

Further sensitivity tests were conducted to assess the extent to which the finding reflects only differences in the over-arching programs examined (GAIN, PI, and NEWWS), or the state programs in the sample instead of variation across local offices within these larger aggregates.[28] However, even when these sources of variation are removed, the basic finding stays the same.

Getting Close to the Client Counts

Findings for personalized client attention are also striking, statistically significant, and robust. The unstandardized and standardized regression coefficients for this variable indicate that increasing it by one unit (one standard deviation of its distribution across office) increases program impacts by $428, other things equal. Thus if all variables in the model are at their mean value and the personalized attention scale is increased by one unit, the estimated program impact would increase from $879 to $1,307, or from 18 percent to 27 percent of the average counterfactual. The *p*-value for the coefficients is 0.0002, which indicates that they are highly statistically significant. In addition, sensitivity tests demonstrate that this finding is robust to variations in sample composition and whether or not cross-program or cross-state variation is included in the analysis. Hence the finding strongly suggests that personalized atten-

[26] The counterfactual (control group conditional mean earnings) was $4,871.

[27] See Bloom, Hill, and Riccio (2001, Appendix D).

[28] Cross-program variation was removed from the analysis by adding dummy variables for GAIN and PI (with NEWWS the omitted category) to equations 2, 3, and 4. Doing so eliminated the fixed effects of these overarching programs. Cross-state variation was removed by adding dummy variables for six of the seven states to equations 2, 3, and 4. This eliminated state fixed effects. The only major finding to change in either case was that discussed later for caseload size.

tion can make a big difference for clients above and beyond whatever services they receive and above and beyond other features of a program and its environment.

Monitoring Alone Is Not Enough

As noted earlier, knowing in a timely fashion how clients are progressing in their assigned program activities is presumably essential if frontline staff are to provide helpful case management or enforce participation mandates. It is therefore surprising that we find offices that more closely monitor clients have smaller effects, other things equal. Specifically, the unstandardized and standardized regression coefficients for this variable in Table 2 indicate that increasing monitoring by one unit (a standard deviation) reduces program impacts by $197. However, these coefficients are not quite statistically significant ($p = 0.110$) and are not robust to sensitivity tests.

When interpreting this finding, note that the monitoring measure we use focuses on the timeliness of staff knowledge about client participation and progress. It does not focus on efforts to enforce compliance or provide assistance. Thus local offices that took a very tough stance or a very lenient stance toward enforcement could have rated high on this scale. Perhaps then what really matters is not just staff awareness of client behavior, but what staff members do with this information.

Large Caseloads Can Produce Small Impacts

Staff members' client caseload has a large and statistically significant negative effect on program impacts. The regression coefficient for this variable indicates that office impacts decline by $4 per additional client per caseworker, other things equal. And its p-value of 0.003 is highly statistically significant. It is more helpful, however, to view this finding through the lens of a partially standardized regression coefficient. This parameter implies that increasing caseload size by one standard deviation of its distribution across offices (67 clients) reduces program impacts by $268, which is substantial.

Sensitivity tests indicate that this finding is robust to variations in the mix of local offices included. Thus it is pervasive and not just confined to a few unusual sites. However, the finding is sensitive to controlling for the evaluation in which sample members were originally involved. Close inspection of this result does not suggest a clear interpretation for it, however.[29]

Although consistent with conventional wisdom, our finding that increased caseload reduces program impacts conflicts with prior results from the Riverside GAIN caseload experiment (Riccio, Friedlander, and Freedman, 1994), which found no difference in earnings impacts between sample members randomly assigned to staff with a standard caseload (averaging 97 clients per caseworker) versus those assigned to staff with a reduced caseload (averaging 53 clients per caseworker). However, our analysis examines caseloads that typically are much larger and vary much more than those for Riverside GAIN. The mean caseload size for a program office in the present study is 136 and its standard deviation across offices is 67. Thus, plus-or-minus one standard deviation from the mean spans a range from 69 clients per caseworker to 203 clients per caseworker. It therefore stands to reason that program impacts

[29] When a dummy variable for GAIN was added to equations 2, 3, and 4 the coefficient for caseload size in equation 2 dropped almost to 0. However, the pattern of correlations for this dummy variable with program impacts, caseload size, and other local office features did not produce a clear explanation of why its presence in the model affected the caseload size coefficient.

would erode substantially when caseloads begin to approach the higher end of this range, where staff may have very little time to devote to each client.

The Importance of Consistency in Staff Views Is Ambiguous

Findings are mixed for our two final measures of implementation—staff-versus-staff and staff-versus-supervisor disagreement about how to provide client services. The regression coefficient for staff-versus-supervisor disagreement is statistically significant ($p = 0.102$) and indicates that program impacts decline by $159 as this form of disagreement increases by one standard deviation, other things equal. This is what we hypothesized.

However, the regression coefficient for staff-versus-staff disagreement is not statistically significant and thus cannot be distinguished from random error. Therefore, on balance it is not clear whether these findings support or challenge the widely held organizational imperative that managers should instill a common sense of mission and method among their staff.

Increasing Basic Education Reduces Short-Run Effects

Findings in Table 2 indicate that programs that increase client use of basic education produce smaller than average effects on short-run earnings gains. The regression coefficient for this variable is negative and statistically significant ($p = 0.017$). It implies that program impacts decline by $16 for each one-point increase in the program-induced percentage of clients who receive basic education, other things equal. The partially standardized regression coefficient indicates that program impacts decline by $208 when the program-induced percentage of clients who receive basic education increases by one standard deviation (13 percentage points).

Although this short-run effect for basic education is consistent with the original findings from the GAIN and NEWWS evaluations (Hamilton, 2002; Riccio, Friedlander, and Freedman, 1994), it is not clear why vocational training does not also depress short-run impacts, because it too imposes an opportunity cost of time required in the classroom that might have been spent producing earnings in the workplace. However basic education often does not have a clear employment focus or a direct connection to the world of work, whereas vocational training usually has both of these features. In addition, program clients are often required to attend basic education as an initial activity. For such persons basic education might be less effective than for others who choose this option for themselves.[30]

At the same time, it is important to recall that the local programs found most effective by the original GAIN and NEWWS studies included basic education in their mix of client activities. Thus it may be that a more extreme emphasis on mandatory immediate basic education may be particularly detrimental but that in moderation and on a voluntary basis this activity might be effective.[31]

[30] See Hamilton and Brock, 1994.

[31] Our findings apply only to short-run effects. Effectiveness of job search, basic education, and vocational training during a period up to 5 years after random assignment are examined by Hamilton et al. (2001) for NEWWS; and by Freedman et al. (1996) for GAIN. Hotz, Imbens, and Klerman (2000) examine an additional 4 years of effects for GAIN (when the restriction was no longer in effect that prevented controls from being served by the program).

Job Search Activities Alone Do Not Assure Success, and the Effect of Vocational Training Is Unclear

Given the central role that job search has had in many past successful programs, it is noteworthy that its coefficient in our model is nearly zero and is not statistically significant. However, this finding does not necessarily mean that job search is unimportant for program success, or that some programs could operate just as effectively without it. Instead it might be the case that the particular kinds of messages and assistance that get conveyed to clients within job search activities may determine whether those activities are effective. For example, job search assistance may be an important vehicle for operationalizing a quick employment message for clients; but holding constant this message, job search assistance may have little or no impact.

It should also be noted that vocational training did not have a statistically significant influence on program effectiveness, and although its regression coefficient was positive it was much smaller than that for basic education. Thus, more or less use of this activity did not seem to influence program effectiveness appreciably. Like the finding for job search, it is possible that the finding for vocational training reflects that the specific employment-related activity used by a program matters less than the way it is used. And perhaps what most distinguishes job search and vocational training from basic education is that the former are directly employment-related, whereas the latter is not.

It's Hard to Raise Earnings when Jobs Are Scarce

The regression coefficient for the unemployment rate in Table 2 is highly statistically significant ($p = 0.004$) and implies that a one percentage point increase in the unemployment rate reduces program impacts by $94, other things equal. The partially standardized regression coefficient indicates that an increase of one standard deviation in the unemployment rate (3.1 percentage points) will reduce program impacts by $291. This sizable estimated decline was robust to sensitivity tests.

Thus it appears that other things being equal, the performance of welfare-to-work programs declines when unemployment rates rise and jobs for clients become harder to find. This result has particular relevance for setting performance standards in the current depressed economic climate.

Constellations of Program Characteristics Can Really Count

Perhaps the most useful way to apply the findings in Table 2 is to use them to project the likely effectiveness (impacts) of welfare-to-work programs with different constellations of approaches to serving clients. Consider the following examples:

Approach #1: Employment focus with close direction of program staff and clients

- Staff encourage clients to get jobs quickly,
- Staff support this strategy through personal client attention,
- Staff monitor client progress closely,
- Staff share this vision with each other, and
- Staff share this vision with their supervisors.

Approach #2: Laissez-faire management of staff and clients

- Client-to-staff ratios are very high,
- Clients do not receive personal attention,
- Client progress is not monitored closely,
- Staff do not share a common vision with each other, and
- Staff do not share a common vision with their supervisors.

The preceding examples involve program features that managers can influence. Hence, projecting the likely impacts of each example can help illustrate how managers might improve program performance.

For example, if each factor listed for approach #1 were one standard deviation from its mean value and all other factors in the model were at their mean value, the findings in Table 2 suggest that approach #1 would increase client earnings by $986 more than would the average program in the present sample. Hence, Approach #1 would increase client earnings by $1,865 or 38 percent of the mean counterfactual. (Note that this finding does not account for potential nonlinearities such as interaction effects or threshold effects, which we plan to explore in future work.)

If each factor listed for approach #2 were one standard deviation from its mean value and all other factors in the model were at their mean value, the findings in Table 2 suggest that approach #2 would reduce impacts by $534 compared to the average program in the present sample. Hence, approach #2 would only increase client earnings by $345 or 7 percent.

Now compare the projected effectiveness of the two examples. Approach #1 would increase client earnings by $1,865 or 38 percent whereas approach #2 would only increase client earnings by $345 or 7 percent. These projections suggest that differences in program implementation can produce important differences in program effectiveness even with the same types of clients, the same mix of program activities, and the same economic conditions.

The Types of People Served Have Some—but Limited—Consequences for Program Performance

Table 3 indicates how program impacts vary with each client characteristic in our model, holding constant all other client and program characteristics. These findings are estimates of regression coefficients for equation 1.

Because client characteristics are defined as distinct categories and represented in the model by 0 or 1 indicator variables, it is only necessary to report the regression coefficient for each category plus its p-value and standard error. (Partially standardized regression coefficients do not add useful information.) These coefficients represent the regression-adjusted difference in mean program impacts for a sample member with the specified characteristic and a sample member in the omitted category for that characteristic, other things being equal.

For example, the regression coefficient of $653 for clients with a high school diploma or GED at random assignment (in the first row of the table) implies that the impact for sample members with this credential is $653 greater than the impact for clients without it, other things being equal. This finding is highly statistically significant ($p = 0.001$). It implies that, on average, if programs differed only in terms of the proportion of their clients having a high school credential, those serving a higher proportion of such people would have larger impacts than those serving a smaller proportion.

Table 3. The relationships between client characteristics and program impacts.

	Regression Coefficient	Statistical Significance (*p*-value)	Standard Error
Was a high school graduate or had a GED	$ 653***	0.001	$ 187
Was a recent welfare applicant	– 145	0.532	232
Had received welfare for past 12 months	444*	0.085	258
Had a child under six years old	34	0.841	171
Had one or no children (omitted category)			
Had two children	301	0.160	214
Had three or more children	591***	0.003	199
Was less than 25 years old	206	0.557	351
Was 25 to 34 years old	105	0.707	281
Was 35 to 44 years old	305	0.376	345
Was 45 or older (omitted category)			
Was White, non-Hispanic (omitted category)			
Was Black, non-Hispanic	– 178	0.369	199
Was Hispanic	– 213	0.527	337
Was Native American	– 696	0.115	442
Was Asian	353	0.560	606
Was some other race/ethnicity	726	0.487	1,044
Had zero earnings in the past year (omitted category)			
Had earned $1 to $2499	– 186	0.222	152
Had earned $2500 to $7499	72	0.787	267
Had earned $7500 or more	22	0.965	501

Regression coefficients represent the conditional difference in mean impacts on follow-up earnings (in 1996 dollars) for the category specified and the omitted category listed or implied. These coefficients are estimated simultaneously with those reported in Table 2. The grand mean impact is $879 or 18 percent of the counterfactual. Statistical significance is indicated by * for the 0.10-level, ** for the 0.05-level and *** for the 0.01-level.

It is straightforward to extend this type of interpretation to characteristics with more than two categories. Consider the number of children that clients had at the time of random assignment. The regression coefficient for "had three or more children" (which is highly statistically significant) indicates that program impacts for clients in this category are $591 greater than for those who are similar in all other ways except that they "had one or no children" (the omitted category for this characteristic).

The only other statistically significant coefficient in the table is for women who received welfare during all 12 months before random assignment and thus were more welfare dependent than average. This coefficient indicates that program impacts for women with this characteristic were $444 larger than for those who were less welfare dependent, other things equal.

Taken together, the findings in Table 3 reflect a mixed picture of how client characteristics affect program impacts. Impacts are not consistently larger or smaller for clients that are likely to be easier or harder to employ. Thus, while some characteristics matter to program effectiveness, others do not.

Another way to understand the relative importance of the types of people served is to consider how much of the cross-office variance in the unconditional impacts is explained by the cross-office variation in the characteristics of clients. (This is determined, in essence, by measuring how much lower the variation in conditional impacts is relative to the variation in unconditional effects.) Client characteristics explain about 16 percent of the variation in program impacts across offices. By contrast, much more variance is explained by the set of implementation-related factors (that is, program strategies and economic context). When these are added to the model, the variance explained jumps to 80 percent. In sum, a program's effectiveness is only modestly determined by the nature of its clientele; what is done for and with them matters much more.

CONCLUDING THOUGHTS

This paper illustrates what can be achieved by a quantitative synthesis of original data from random assignment experiments that provide valid and reliable estimates of program impacts for many sites, plus valid and reliable measures of how these sites implemented their programs. Thus, this study is an example of research using multiple levels of information, with high-quality data from a number of sites, which Lynn, Heinrich, and Hill (2001) argue can provide the most useful insights for public sector governance.

Our findings, which were substantial in magnitude, statistically significant, and robust to variations in sample composition and structure, demonstrate that holding other factors in the model constant:

- Management choices for how welfare-to-work programs are implemented matter a great deal to their success. In particular: a strong employment message is a powerful medium for stimulating clients to find jobs, a clear staff focus on personal client attention can markedly increase their success, and large client caseloads can undercut program effectiveness.
- Increased reliance on mandatory basic education reduces short-run program effectiveness. Thus, programs that directly emphasize employment are more successful in the short run.
- The local economic environment is a major determinant of program success; programs are much less effective when jobs are scarce.
- Welfare-to-work programs can be effective for many different types of clients, although some client characteristics may make a difference. However, it is not clear that targeting on clients who are especially job-ready (or not) influences program effectiveness.
- Overall, the way that a program is implemented has much more influence on its effectiveness (impacts) than does the types of clients it serves.

These findings are based on a solid foundation of valid and reliable impact estimates from random assignment experiments for 59 local program offices. Nevertheless, they also rely on a non-experimental model of the natural variation in these impacts. Therefore, these findings are only as valid as the model upon which they are based. To maximize their validity, we have carefully specified our model to ensure that it represents all four general categories of factors likely to influence program effectiveness—implementation, activities, environment, and client characteristics. And within each category we have attempted to include specific factors that are judged by researchers and practitioners to be most relevant for program success. Furthermore, we have subjected our findings to a series of stringent sensitivity tests. Thus, we have confidence in the results presented but acknowledge that certainty about them is not possible.

In closing, we emphasize that our research was possible only because of the careful, comprehensive, and consistent data collection efforts of the experiments that we pooled and the broad range of circumstances that they represent. Thus, as social scientists and policy researchers develop their research agendas and as government agencies and foundations make their future research funding plans, we urge them to emphasize a long-run strategy for accumulating program knowledge based on:

(1) random assignment experiments that make it possible to obtain valid and reliable estimates of program effectiveness,
(2) multi-site experiments that reflect the existing range of natural variation in program effectiveness,
(3) careful specification of the likely determinants of program effectiveness based on social science theory, past empirical research, and experiential knowledge of practitioners,
(4) equally careful and consistent measurement of these hypothesized determinants across studies and sites, and
(5) adequate support for and attention to quantitative syntheses of this information.

In this way we believe that the most progress possible can be made toward unpacking the "black box" of social programs and thereby acquiring the information needed to improve them.

The authors thank Larry Lynn for his collaboration on the conception and initiation of the project upon which this article is based; Judy Gueron for her feedback and support throughout the project; colleagues at MDRC for their helping us to understand the programs studied and data used for the project; two anonymous reviewers for helpful comments; and The Pew Charitable Trusts for funding the project.

HOWARD S. BLOOM is Chief Social Scientist at MDRC.

CAROLYN J. HILL is Assistant Professor at the Georgetown Public Policy Institute, Georgetown University.

JAMES A. RICCIO is a Senior Fellow at MDRC.

REFERENCES

Bane, M.J. (1989). Welfare reform and mandatory versus voluntary work: policy issue or management problem? Journal of Policy Analysis and Management, 8(2), 285-289.

Behn, R. (1991). Leadership counts: lessons for public managers from the Massachusetts welfare, training and employment program. Cambridge, MA: Harvard University Press.

Bloom, H.S., Hill, C.J, & Riccio, J. (2001). Modeling the performance of welfare-to-work programs: the effects of program management and services, economic environment, and client characteristics. New York: MDRC.

Brodkin, E.Z. (1997). Inside the welfare contract: Discretion and accountability in state welfare administration. Social Service Review, 71(1), 1-33.

Economic Report of the President. (2000). Washington, DC: U.S. Government Printing Office.

Freedman, S., Friedlander, D., Hamilton, G., Rock, J., Mitchell, M., Nudelman, J., Schweder, A., & Storto, L. (2002). Two-year impacts for eleven programs. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of the Assistant Secretary for Planning and Evaluation and US Department of Education, Office of the Under Secretary, Office of Vocational and Adult Education.

Freedman, S., Friedlander, D., Lin, W., & Schweder, A. (1996). The GAIN evaluation: five-year impacts on employment, earnings, and AFDC receipt. New York: MDRC.

Greene, W.H. (1993). Econometric analysis. Upper Saddle River, NJ: Prentice Hall.

Greenberg, D., Meyer, R., & Wiseman, M. (1994). Multi-site employment and training evaluations: a tale of three studies. Industrial and Labor Relations Review, 47(4), 679-691.

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

Hagen, J.L., & Lurie, I. (1994). Implementing JOBS: progress and promise. Albany, NY: Nelson A. Rockefeller Institute of Government.

Hamilton, G. (2002). Moving people from welfare to work: lessons from the National Evaluation of Welfare-to-Work Strategies. Washington DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of the Assistant Secretary for Planning and Evaluation and U.S. Department of Education, Office of the Under Secretary, Office of Vocational and Adult Education.

Hamilton, G., & Brock, T. (1994). The JOBS evaluation: early lessons from seven sites. Washington DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of the Assistant Secretary for Planning and Evaluation and U.S. Department of Education, Office of the Under Secretary, Office of Vocational and Adult Education.

Hamilton, G., Freedman, S., Gennetian, L., Michalopoulos, C., Walter, J., Adams-Ciardullo, D., Gassman-Pines, A., McGroder, S., Zaslow, M., Brooks, J., & Ahluwalia, S. (2001). National evaluation of welfare-to-work strategies: how effective are different welfare-to-work approaches? Five-year adult and child impacts for eleven programs. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of the Assistant Secretary for Planning and Evaluation and U.S. Department of Education, Office of the Under Secretary, Office of Vocational and Adult Education.

Heinrich, C.J. (2002). Outcomes-based performance management in the public sector: implications for government accountability and effectiveness. Public Administration Review, 62(6), 712-725.

Hotz, V.J., Imbens, G.W., & Klerman, J.A. (2000). The long-term gains from GAIN: a re-analysis of the impacts of the California GAIN program. Cambridge, MA: National Bureau of Economic Research.

Kemple, J., & Haimson, J. (1994). Florida's Project Independence: program implementation, participation patterns, and first-year impacts. New York: MDRC.

Lynn, L.E., Jr., Heinrich, C.J., & Hill, C.J. (2001). Improving governance: a new logic for empirical research. Washington, DC: Georgetown University Press.

Mead, L.M. (1983). Expectations and welfare work: WIN in New York City. Policy Studies Review, 2(4), 648-661.

Mead, L.M. (1986). Beyond entitlement: the social obligations of citizenship. New York: The Free Press.

Meyers, M.K., Glaser, B., & MacDonald, K. (1998). On the front lines of welfare delivery: Are workers implementing policy reforms? Journal of Policy Analysis and Management, 17(1), 1-22.

Michalopoulos, C., Schwartz, C., with Adams-Ciardullo, D. (2001). National evaluation of welfare-to-work strategies: what works best for whom? Impacts of 20 welfare-to-work programs by subgroup. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of the Assistant Secretary for Planning and Evaluation and U.S. Department of Education, Office of the Under Secretary, Office of Vocational and Adult Education, January.

Miller, G. (1992). Managerial dilemmas: the political economy of hierarchy. New York: Cambridge University Press.

Nathan, R. (1993). Turning promises into performance: the management challenge of implementing workfare. New York: Columbia University Press.

Raudenbush, S., & Bryk, A. (2002). Hierarchical linear models: applications and data analysis methods, 2nd ed. Thousand Oaks, CA: Sage Publications.

Raudenbush, S.W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. Psychological Methods, 5(2), 199–213.

Riccio, J., Bloom, H.S., & Hill, C.J. (2000). Management, organizational characteristics, and performance: the case of welfare-to-work programs. In Heinrich & Lynn, eds. Governance and performance: new perspectives (pp. 166-198). Washington, DC: Georgetown University Press.

Riccio, J., & Friedlander, D. (1992). GAIN: program strategies, participation patterns, and first-year impacts in six counties. New York: MDRC.

Riccio, J., Friedlander, D., & Freedman, S. (1994). GAIN: benefits, costs, and three-year impact of a welfare-to-work program. New York: MDRC.

Riccio, J., & Hasenfeld, Y. (1996). Enforcing a participation mandate in a welfare-to-work program. Social Service Review, 70 (4), 516-542.

Riccio, J., & Orenstein, A. (1996). Understanding best practices for operating welfare-to-work programs. Evaluation Review, 20 (1), 3-28.

Wiseman, M. (1987). How workfare really works. Public Interest, 89, 36-47.

APPENDIX

Table A.1. Client participation in employment-related activities.

	Basic Education	Job Search Assistance	Vocational Training
Mean percentage of program group members who participated in the activity	19	22	27
Mean percentage of control group members who participated in the activity	8	5	22
Mean *difference* in participation rates between program and control group members	11	17	5
Standard deviation across the 59 offices of the difference in participation rates	13	12	10
Range across the 59 offices of the difference in participation rates	–11 to 50	–13 to 47	–21 to 35

Source: MDRC surveys of randomly sampled program and control group members from each local office.

Table A.2. Client characteristics.[a]

At Random Assignment the Sample Member:	Percentage of Full Sample of Individuals	Cross-Office Range (%)
Was a high school graduate or had a GED	56	17 to 74
Was a welfare applicant	17	0 to 99
Had received welfare for past 12 months	44	0 to 96
Had a child under 6 years old	46	7 to 73
Had one or no children	42	30 to 56
Had two children	33	28 to 50
Had three or more children	25	11 to 39
Was younger than 25 years old	19	1 to 42
Was 25 to 34	49	23 to 57
Was 35 to 44	26	14 to 45
Was 45 or older	6	2 to 34
Was white, non-Hispanic	41	1 to 87
Was black, non-Hispanic	41	0 to 98
Was Hispanic	14	0 to 92
Was Native American	2	0 to 21
Was Asian	2	0 to 23
Was some other race/ethnicity	< 1	0 to 5
Had zero earnings in the past year	56	29 to 81
Had earned $1 to $2499	21	10 to 30
Had earned $2500 to $7499	14	6 to 26
Had earned $7500 or more	9	2 to 27
Sample size = 69,399		

[a] The sample in this analysis is restricted to females only.