Institute for Research on Poverty
Discussion Paper no. 1225-01

Explaining Variation in the Effects of Welfare-to-Work Programs

David Greenberg
Department of Economics
University of Maryland, Baltimore County
E-mail: dhgreenb@umbc7.umbcedu

Robert Meyer
Harris Graduate School of Public Policy Studies
University of Chicago
E-mail: meyer@cicero.spc.uchicago.edu

Charles Michalopoulos
Manpower Demonstration Research Corporation
E-mail: charles_michalopoulos@mdrc.org

Michael Wiseman
National Opinion Research Center
E-mail: MichaelLWiseman@cs.com

March 2001

Abstract

Evaluations of government-funded employment and training programs often combine results from similar operations in multiple sites. Findings inevitably vary. It is common to relate site-to-site variations in outcomes to variations in program design, participant characteristics, and the local environment. Frequently such connections are constructed in a narrative synthesis of multisite results. This paper uses data from the evaluation of California's Greater Avenues for Independence (GAIN) program and the National Evaluation of Welfare-to-Work Strategies (NEWWS) to question the legitimacy of such syntheses. The discussion is carried out using a simple multilevel evaluation model that incorporates models of both individual outcomes within sites and variation in program effects across sites. Our results indicate that tempting generalizations about GAIN and NEWWS effects are statistically unjustified, but that significant progress might be made in identifying the determinants of program effects in future demonstrations with some changes in evaluation strategy.

Explaining Variation in the Effects of Welfare-to-Work Programs

1. THE QUESTION

The number of evaluations of government-funded employment and training programs grows without sign of abatement (Barnow and King, 1999; Friedlander, Greenberg, and Robins, 1997; LaLonde, 1995; Blank, 1994; Greenberg and Wiseman, 1992; Gueron and Pauly, 1991). Virtually all these evaluated programs include large numbers of public assistance recipients. Indeed, many are specifically targeted at this population and are commonly called welfare-to-work programs.

Because many evaluations of government-funded employment and training programs attempt to obtain a broadly representative set of local conditions and an adequate sample size, they often simultaneously examine similar programs at several different sites. For example, the New York Child Assistance Program (CAP), which evaluated the consequences for recipients of public assistance of a combination of incentives and social services, was conducted experimentally in three counties (Hamilton et al., 1992); the Rockefeller Foundation's Minority Female Single Parent Demonstrations, which provided occupational and skill training, were carried out in four cities (Gordon and Burghardt, 1990); the National Job Training Partnership Act (JTPA) Evaluation, which evaluated the nation's major training program for the disadvantaged, was conducted in 16 different sites (Orr et al., 1996); the National Evaluation of Welfare-to-Work Strategies (NEWWS, formerly known as the evaluation of the Job Opportunities and Basic Skills [JOBS] program), involves 11 programs in seven sites (Hamilton and Brock, 1994; Freedman et al., 2000); and the Greater Avenues for Independence (GAIN) evaluation, which is a direct precursor of NEWWS, covered six counties in California (Riccio, Friedlander, and Freedman, 1994). Two evaluations of welfare-to-work programs are especially notable in terms of number of sites: the Food Stamp Employment and Training Program Evaluation involved around 13,000 Food Stamp recipients in 53 separate sites in 23 states (Puma et al., 1990), while the evaluation of the

AFDC Homemaker-Home Health Aide Demonstration was based on about 9,500 AFDC recipients in 70 sites in seven states (Bell and Orr, 1994).

Findings from multisite evaluations of employment and training programs inevitably vary across sites. To mention only a few of numerous possible examples, CAP was found to be more successful in one of the counties in which it was tested than in the other two; GAIN appears to have "worked" much better in Riverside County than in Los Angeles County; the Minority Female Single Parent intervention seemed to be effective in only one of four test sites, San Jose; and positive effects on earnings were found in some Food Stamp Employment and Training Program Evaluation and National JTPA Evaluation sites and negative effects in others.

It is natural to attempt to determine what it is that causes program effects to differ from place to place, for such differences seem to have the capacity to provide information about training program production functions by allowing examination of how the effects vary with cross-site variations in program design, participant characteristics, and the environment in which the program is implemented. For example, policy makers have often attributed Riverside's success in the GAIN program to the fact that, relative to the other GAIN sites and to other welfare-to-work programs operating at the time, it put special emphasis on placing participants into jobs as quickly as possible (Greenberg, Mandell, and Onstott, 2000). Similarly, San Jose's success in the Minority Female Single Parent intervention has been credited to the fact that, unlike the other three sites, it immediately provided job-specific skill training to all participants and integrated basic literacy and mathematics skill into job-specific training. Such policy lessons might well be correct, but they can also be unreliable. Hence, as stressed in this paper, great care should be exercised in actually making policy on the basis of observed cross-site variation in estimates of program effects.

The most common approach to explaining observed variation in cross-site program effects is to provide a description of each of the ways in which sites that differ in terms of program effects appear to vary from one another, an approach we term "narrative synthesis." Narrative synthesis, however, runs the risk of overinterpreting the data; there may be fewer sites than dimensions in which they differ. The GAIN evaluation provides a useful illustration. In attempting to “explain” differences among the six GAIN demonstration counties, the evaluators examined 17 separate explanatory variables, as well as interactions among these variables (Riccio, Friedlander, and Freedman, 1994, chapter 8).

Formal statistical procedures have only rarely been used to draw conclusions from observed cross-site variation in estimates of employment and training program effects, and in those few instances when they have been, it has not been possible to reach any useful conclusions. The statistical approach typically involves using regression models, which we term “macro equations,” to attempt to find the correlates of cross-site variations in program effects. The only two previous such attempts with which we are familiar occurred in the JTPA study (Orr et al., 1996, pp. 105–106, 123–124) and the Food Stamp Employment and Training Program evaluation (Puma et al., 1990, Table 7.10). Both yielded virtually no statistically significant coefficients[1] (see Greenberg, Meyer, and Wiseman, 1994, for a discussion).

We present macro equations that rely on cross-site variation in program effect estimates from the GAIN and NEWWS evaluations. The resulting regression coefficients are typically reasonable in sign and magnitude and often suggestive and provocative, but they are rarely statistically significant or robust to alternative regression specifications. Our paper explores the reasons for this and examines the circumstances under which evaluations of government-funded training programs might yield statistically significant coefficients. The major lesson is that a cross-site comparison of program estimates is difficult, and potentially hazardous, if based on relatively few sites (say, less than 20). Hence, policy inferences drawn from such a comparison may be misleading.

In the next section, we set out a simple model of what evaluations are about and how data from multiple sites can be used appropriately to examine training program production functions. In section 3, we use the model to examine the circumstances in which productive relationships are most likely to be uncovered and estimated with acceptable statistical precision. In both sections, we illustrate the issues by using results from the GAIN and NEWWS evaluations. Section 4 contains our conclusions. Although both GAIN and NEWWS are evaluations of welfare-to-work programs, our analysis is applicable to the evaluation of any social intervention aimed at individuals and introduced in several locations.

2. AN INTRODUCTION TO RESEARCH SYNTHESIS: A SIMPLE MULTILEVEL EVALUATION MODEL

Multilevel Analysis

We begin by presenting a simple multilevel statistical model that permits formal evaluation of the determinants of program net effects. The model is referred to as *multilevel* (or *hierarchical*) because it is based on both individual-level and site-level data from multiple sites. Multilevel models have been used extensively in the education literature and elsewhere (for descriptions of the methods, see Bryk and Raudenbush, 1992; Goldstein, 1995; Kreft and de Leeuw, 1998; or Snijders and Bosker, 1999). As will be demonstrated below, the model is a rudimentary extension of the evaluation framework commonly used to study employment and training programs.

We set the stage by assuming that an employment and training program innovation is introduced for some class of people in several different sites—for example, in several different local welfare offices—that are sufficiently separated to assure no significant spillover of program effects from one location to the next. Information is collected on the characteristics of the clients, the economic environment (for example, the local unemployment rate, the skills required for local jobs, etc.), the innovation, and the outcomes of interest (for example, earnings and welfare receipts). We assume that uniform data collection methods are used across sites. However, the methods we discuss can be applied, with some modification, to data obtained from multiple independent studies in which some variation in outcome and input measures and methods occurs.[2]

As is customary in the literature on multilevel modeling, we describe how an intervention produces effects with two sets of equations. The first is a set of *micro* models, one for each outcome within each site, based solely on individual-level data. The micro models provide information on the effects of a program in a particular site overall, or for particular subgroups of people. The second set consists of *macro* models, one for each outcome of interest, but including all the sites. The objective of the macro models is to understand what types of factors are related to differences in the effectiveness of programs at different sites. For example, the macro model might investigate whether a program's effects are affected by the state of the local economy, the demographic makeup of the caseload, and the type of intervention being tested.[3] In the remainder of this section, we assume that the macro and micro equations are properly specified and that the explanatory variables are accurately measured.

The Micro Model

The micro model is the evaluation model that has been used in numerous studies of employment and training programs to estimate their effects on various outcomes such as earnings or welfare status at the site level (see the descriptions in Greenberg and Wiseman, 1992, and Friedlander, Greenberg, and Robins, 1997). It is given by[4]

$$Y_{ij} = \beta_j X_{ij} + \theta_j P_{ij} + e_{ij} \tag{1}$$

where i and j index individuals and sites, respectively. Here Y_{ij} is the outcome measure, X_{ij} is a vector of regressors with associated coefficients β_j, P_{ij} is a binary variable identifying program participation, and e_{ij} is the error. The micro model relies on a comparison of individuals at site j who participated in the employment and training program being evaluated (the treatment group, $P_{ij} = 1$) with similar persons who did not (the comparison group, $P_{ij} = 0$). Although the assignment of individuals between the treatment and comparison groups is often done randomly, this is not essential for estimating equation 1.[5] The key parameter in equation 1 is θ_j, which provides an estimate of the size of the program's effect in site j on the outcome of interest.

If individuals at each site are assigned at random to the treatment and control groups and there are no regressors in the equation, then the estimate of the effect of a program, θ_j, is simply the average difference in outcome between the treatment and control groups in each site.[6] When there is random assignment with regressors (X_{ij}), equation 1 provides a regression-adjusted effect of the program that is an unbiased estimate of the same parameter, θ_j. The overall effect of the program across all sites is given by the global mean, $\bar{\theta} = \sum n_j \theta_j / \sum n_j$, where n_j is the total combined number of program participants and controls in site *j*.

An example of the program effect estimates obtained by estimating equation 1 is shown in Table 1. The table shows regression-adjusted mean earnings levels for members of program groups and control groups in 13 programs assessed in two evaluations of welfare-to-work programs that are based on random assignment. The California GAIN program was studied by the Manpower Demonstration Research Corporation (MDRC) in six sites (see Riccio, Friedlander, and Freedman, 1994). In the National Evaluation of Welfare-to-Work Strategies (NEWWS), 11 programs that operated in seven different locations are being studied. Findings from seven of these programs were published well before results for the other four programs became available. Table 1 shows the results for these seven programs (see Hamilton et al., 1997, for the programs in Atlanta, Grand Rapids, and Riverside; and Scrivener et al., 1998, for the program in Portland). All of the programs were designed to increase the earnings of welfare recipients using the provided services. As discussed later, they differed from one another in the approaches for accomplishing this.[7]

In each case, ordinary least squares (OLS) regressions were used to calculate the effect of the program, but the coefficients on the covariates are excluded from Table 1 for simplicity. The effects of the programs differed substantially from one another. The most successful programs—Riverside GAIN[8] and Portland—increased earnings by more than $1,000 per year. The least successful programs—Riverside HCD and Tulare—had virtually no effect on earnings. As previously indicated, the overall effect across programs, $\bar{\theta}$, is simply the average of the effects of the individual programs, weighted by the sample size in each program. In this case, the average is an increase in earnings of about $600 per year.

TABLE 1
Effects of Selected Welfare-to-Work Programs on Annual Earnings of Single-Parent Families in 2nd Year after Random Assignment

	Sample	Program	Control	Program Effect		Standard
Site	Size	Group	Group	(Difference)		Error
GAIN						
Alameda	1,205	2132	1624	508	*	328
Butte	1,228	2998	2442	556		383
Los Angeles	4,396	1699	1589	110		173
Riverside	5,508	3416	2233	1183	***	183
San Diego	8,219	3503	2794	709	***	169
Tulare	2,234	2536	2531	5		250
GAIN AVERAGE		2940	2319	620	***	90
Chi-square statistic for homogeneity				24.59		
p-value				0.0002		
NEWWS						
Atlanta LFA	3,833	2828	2075	753	***	173
Atlanta HCD	3,881	2471	2075	396	**	179
Grand Rapids LFA	3,012	2858	2383	475	**	195
Grand Rapids HCD	2,997	2833	2383	450	**	196
Riverside LFA	6,726	2979	2418	561	***	130
Riverside HCD	3,192	1889	1849	39		189
Portland	5,547	4374	3183	1192	***	149
NEWWS AVERAGE		3010	2403	607	***	63
Chi-square statistic for homogeneity				27.68		
p-value				0.0001		
OVERALL AVERAGE		2979	2367	613	***	53
Chi-square statistic for homogeneity				52.17		
p-value				0.0000		

Sources: Table D.2 - D.7 of Riccio, Friedlander, and Freedman (1994); Table F.1 of Scrivener et al. (1998); and Tables E.1, E.2, E.3, F.1, F.2, and F.7 of Hamilton et al. (1997).

Notes: Statistical significance levels are indicated as ***=1 percent; **=5 percent; and *=10 percent.

Rounding may cause slight discrepancies in calculating sums and differences.

Standard errors for individual programs imputed based on significance levels reported in MDRC reports and assuming identical error structures across the sites.

The Macro Model

With respect to the literature on employment and training program effects, the new wrinkle in our analysis is the macro model.[9] In the context of a model with no subgroup effects, a simple macro equation is given by

$$\theta_j = \gamma F_j + w_j , \tag{2}$$

where θ_j is the estimate of program effects from the micro model, which is presumed to vary across sites; F_j represents a vector of program characteristics, community characteristics, and economic conditions (including a constant term) assumed to influence the size of the effect of a program in a given site; w_j is an error term; and γ represents a parameter vector that measures the influence of different site characteristics on program effects. In words, the macro equation tries to explain variation in the effects of employment and training services from site to site (θ_j) with a variety of factors (F_j) such as the types of services provided, the types of clients served, and site economic conditions. It is from estimates of the parameters of the macro equation, specifically γ, that insight into how a program generates effects may be gained.

The macro model can be estimated in one of two equivalent ways. First, it can be substituted into the micro equation (thereby eliminating the program effect estimate parameters θ_j) and estimated jointly with the remaining micro parameters and variance components. The combined model is a standard random effects model. Hsiao (1986, pp. 151–153), Amemiya (1978), and Bryk and Raudenbush (1992) discuss alternative methods of estimation. Second, the macro model can be estimated after the micro model program effect parameters ($\hat{\theta}_j$) have been estimated. In this case, equation 2 needs to be rewritten to accommodate the fact that $\hat{\theta}_j$ is estimated with error. Thus,

$$\hat{\theta}_j = \gamma F_j + w_j + \varepsilon_j , \tag{2′}$$

where ε_j is the error in estimating θ_j. Bryk and Raudenbush (1992) and Hedges (1992) discuss estimation methods for this approach.

It is our contention that the second step in macro modeling of intervention effects, that of estimating the macro parameters, γ, should be a primary goal in evaluating employment and training programs.[10] Only then will it be possible to determine the combination of program components that work best for particular types of individuals under various environmental conditions. Such knowledge is essential to improving the effectiveness of employment and training programs, but can only be obtained if macro parameters can be estimated with reasonable precision.

This depends upon two factors: (1) whether there is any genuine variation in program effects across sites and (2) whether this variation, if it exists, is related to identifiable variation in program characteristics, client characteristics, or local environmental conditions. Although rarely done, it makes sense to address the first question before attempting to model the determinants of variation in site effects. In principle, the null hypothesis that program effects are identical in all sites can be tested by a simple chi-square test (Rosenthal and Rubin, 1982; Hedges, 1984; Bryk and Raudenbush, 1992; Greenberg, Meyer, and Wiseman, 1994). It is often extremely helpful to learn that program effects *do not* vary significantly by site—that is, that all the apparent cross-site variation is due simply to random noise—and hence, attempts at explaining variation across sites are neither necessary nor appropriate. Indeed, if such attempts are made in a narrative synthesis, there is a risk of “explaining” apparent cross-site variation in effects in intriguing ways when, in fact, it cannot confidently be attributed to anything more than noise.

Testing Whether Program Effects in NEWWS and GAIN Are Identical across Sites

Table 1, which shows the estimated effects of the GAIN and NEWWS programs, also contains a chi-square test of homogeneity, that is, a test of the null hypothesis that the effects are identical across the sites.[11] This test is appropriate because the sum of a series of squared standard normal random variables has a chi-square distribution:

$$\sum_{j=1}^{J}\left(\frac{(\theta_j-\overline{\theta})}{s_j}\right)^2 \sim \chi^2(J-1)$$

It is based on a result implied by the Central Limit Theorem that the parameters estimated in equation 1 are asymptotically normally distributed.

For the six GAIN sites, the test statistic is 24.59, which allows us to reject the null hypothesis of homogeneity at a significance level below 1 percent. This is not surprising in light of the very large effect of the Riverside program and the near-zero effects of the program in Tulare and Los Angeles, as well as the large samples in each site that permit relatively precise estimates of effects. For the seven NEWWS sites, the test statistic is 27.68, again allowing us to reject the null hypothesis of homogeneity with great confidence. This result is also not a surprise, given the large effect in Portland, the near-zero effect of the Riverside HCD program, and the large samples in each site. Finally, we can also emphatically reject the null hypothesis that all 13 program effects are identical. These findings are important because they imply that there are systematic differences among program effects in both sets of evaluation sites that can potentially be explained by macro models. Estimates from such models are reported later.

The JTPA evaluation provides an instructive contrast to these findings. Although the total sample of 15,981 observations is large, it is split among 16 sites and four target groups: adult men, adult women, male youths, and female youths. Thus, only around 250 observations are available, on average, to estimate a micro equation for each group in each site. In fact, some of the micro equations are based on fewer than 100 observations.[12] Thus, although the variation in the site estimates of program effects is enormous—the range is +$5,310 to −$2,637 for adult men, +$2,628 to −$2,033 for adult women, +$9,473 to −$5,836 for male youths, and +$3,372 to −$3,821 for female youths—it is not surprising that few of the individual estimates of program effects are statistically significant (specifically, only six of 62 are significant at the 10 percent level, the number chance alone should produce) and that tests conducted by the evaluators indicate that the null hypothesis of homogeneity among the sites cannot be rejected (Orr et al., 1996, Tables 4.5 and 4.16). Hence, as the evaluators recognize, it is also not surprising that the macro equations produced virtually no significant coefficients. There is simply no systematic variation to explain.

Determinants of Program Effects in GAIN and NEWWS

As mentioned in Section 1, it is common for policy analysts to confront cross-site variation in the effects of some program by attempting to relate observed variation in calculated effects to reported differences in program features. This narrative synthesis approach amounts to informal estimation of the parameters γ.

For example, after first determining that the differences in program effects among the six GAIN sites are statistically significant, Riccio, Friedlander, and Freedman (1994) then conducted a narrative synthesis to try to explain why. They examined a number of factors. Riccio and colleagues thought that local economic conditions, as measured by county unemployment rates and growth in jobs, might influence the effects of employment and training services on earnings, though it was not clear to them whether better economic conditions would strengthen them by making it easier for program participants to find work or weaken them by making it easier for control group members to find work. Greater utilization of program services such as job search and education and training were expected to result in greater effects on earnings. Programs that emphasized quick employment also were expected to have greater effects on earnings. Finally, the characteristics of those assigned to the programs would be expected to have an effect.

Table 2 shows a variety of measures of economic conditions, program characteristics, and sample composition for the GAIN and NEWWS programs that are similar to the factors considered by Riccio, Friedlander, and Freedman. Most of these variables are self-explanatory. Note, however, that because persons randomly assigned to control groups sometimes obtain services similar to those provided individuals assigned to program groups, program effects on receipt of job search and education and training are measured as the difference between the proportion of program participants and the proportion of controls receiving these services. For similar reasons, program costs are measured net of the cost of employment and training services received by controls. The term “applicants” refers to persons who were assigned to a GAIN or NEWWS program as they entered the welfare system, while “long-term

TABLE 2
Selected Characteristics of Welfare-to-Work Programs in GAIN and NEWWS Evaluations

Site	County Unemployment Rate	Percent of Staff Who Emphasized Quick Employment	Effect on Participation in Job Search	Effect on Participation in Education	Net Cost	Percent Black	Percent Hispanic	Percent Applicants	Percent Long-Term Recipients
GAIN									
Alameda	4.9	21.4	28.0	43.5	$6,437	68.6	7.5	0.0	100.0
Butte	8.7	22.6	n/a	n/a	$3,340	3.5	5.6	60.3	28.2
Los Angeles	6.8	45.0	9.8	41.7	$6,657	45.3	31.9	0.0	100.0
Riverside	9.5	90.6	36.6	54.8	$1,837	15.5	27.6	31.0	39.2
San Diego	5.6	48.8	26.2	42.3	$2,199	22.5	25.3	28.0	41.2
Tulare	15.3	43.9	22.5	56.6	$1,673	3.6	39.2	13.6	57.9
NEWWS									
Atlanta LFA	6.4	82.0	29.1	5.7	$2,277	94.9	0.8	0.3	66.0
Atlanta HCD	6.4	50.0	11.4	24.2	$3,428	94.9	0.8	0.3	66.0
Grand Rapids LFA	5.3	74.0	27.1	-3.0	$1,108	39.3	8	0.1	59.2
Grand Rapids HCD	5.3	74.0	12.8	14.6	$2,872	39.3	8	0.1	59.2
Riverside LFA	11.9	96.0	31.8	-0.8	$1,263	16.7	30.2	1.0	53.8
Riverside HCD	11.9	100.0	21.1	35.2	$2,930	16.7	30.2	1.0	53.8
Portland	6.6	54.0	32.2	9.7	$2,017	20.2	3.9	1.2	64.4

Sources: Tables 1.1, 1.2, and 2.5 and Figure 2.3 of Riccio, Friedlander, and Freedman (1994); Tables 1.1, 1.2, 3.4, and 4.4 of Scrivener et al. (1998); and Tables 1.1, 2.1, 5.5, 6.5, 7.4, and 8.4 and Figure 3.3 of Hamilton et al. (1997).

recipients" are individuals who had received welfare for two years or more prior to being assigned to a program.

Looking factor-by-factor across the GAIN sites, Riccio, Friedlander, and Freedman argued in their narrative synthesis that it was unlikely that differences in program emphasis on job search explained the variation in program effects on earnings because San Diego and Tulare had similar emphasis on job search but very different effects on earnings. Likewise, economic conditions, as measured by the county unemployment rate, were unlikely to explain differences in effects because the program in the worst economy, Tulare, produced virtually no effect, but the program in the second worst economy, Riverside, produced the largest effect. After eliminating several of the factors listed in Table 2 from consideration, Riccio, Friedlander, and Freedman ultimately concluded that a number of the remaining factors probably contributed to the cross-site variation in program effects. For examples, one factor stood out as being especially important in explaining why Riverside produced the largest effect on earnings, namely that nearly all staff in Riverside emphasized quick employment as a goal of the program. In no other GAIN site did more than half the staff emphasize quick employment.

Macro Equation Estimates for GAIN and NEWWS

Conclusions from this narrative synthesis may or may not be accurate. To investigate this issue, we estimated the macro regression equations reported in Table 3.[13] The first column in the table pertains to the GAIN sites. Ideally, we would simultaneously examine all the factors listed in Table 2, and others as well. However, with program effects measured in only six counties, there is no way to do so. Moreover, because one of the variables included in this regression (program effect on participation in job search) was not available for Butte, the regression is based on only five sites. Thus, the regression is limited to three macro explanatory variables, the maximum possible when there are only five sites. None of the coefficients on the three selected macro variables approaches conventional levels of statistical significance, such as 5 or 10 percent. Nonetheless, the three characteristics account for nearly all of the variation in effects across the GAIN sites, more than any other combination of three variables

TABLE 3
Macro Parameters of Effect on Earnings in GAIN and NEWWS

Variable	GAIN Alone	NEWWS Alone	GAIN and NEWWS Combined Sparse Model	GAIN and NEWWS Combined Full Model
Intercept	-262.12	857.83	165.72	-1271.38
Unemployment Rate	-63.89	10.81	-54.39	187.06
	(30.02)	(55.46)	(35.93)	(101.89)
	[0.280]	[0.858]	[0.169]	[0.208]
Percent of Staff Who Emphasized	11.14	-14.31	-0.59	5.37
Quick Employment	(5.92)	(8.36)	(5.20)	(5.68)
	[0.311]	[0.185]	[0.912]	[0.444]
Program Effect on Participation in	33.46	28.50	36.31	-12.68
Job Search	(12.43)	(13.11)	(12.35)	(23.85)
	[0.227]	[0.118]	[0.019]	[0.648]
Program Effect on Participation in				-34.96
Education and Training				(16.84)
				[0.173]
Percent Black				-15.11
				(5.49)
				[0.111]
Percent Hispanic				-68.64
				(25.19)
				[0.112]
Percent Applicants				201.57
				(86.42)
				[0.145]
Percent Long-Term Recipients				22.27
				(23.93)
				[0.450]
Net Program Cost				0.32
				(0.18)
				[0.216]
Adjusted R^2	0.998	0.596	0.511	0.917
Number of Sites	5	7	12	12

Notes: (standard errors are in parentheses); [p-values are in brackets] selected from those listed in Table 2. Moreover, the point estimates of their effects tell a reasonable story. For example, consistent with Riccio, Friedlander, and Freedman’s analysis, the coefficient on the percentage of staff that emphasized quick employment is positive, as is the coefficient on program effect on job search. One could conclude that the large effect of GAIN in Riverside is due both to that program’s emphasis on quick employment and to its greater use of job search. The macro regression for GAIN also suggests that the state of the local economy, as represented by the county unemployment rate, is also an important explanatory factor. According to the estimates, a 1 percentage point increase in the unemployment rate would reduce program effects on earnings by $63 per year. The implication is that the difference in unemployment rates in Tulare and either Alameda or San Diego should have made GAIN’s effects on annual earnings in Tulare about $630 less, and in fact that is what is observed. However, Los Angeles also had much lower unemployment than Tulare, but GAIN registered very small effects in both sites. The macro regression implies that the effects in Los Angeles were low because the GAIN program there made little use of job search. Each percentage point increase in use of job search is related to an increase in the program effect on annual earnings of $33. Thus, the 13 percentage point difference between Tulare and Los Angeles in their respective program’s effect on jobs search would make the earnings effect in Los Angeles around $400 smaller than in Tulare, almost exactly offsetting the negative effects of the higher unemployment rate in Tulare.

Should the point estimates in the first column of Table 3 be taken seriously? They are statistically insignificant at conventional levels. Moreover, they come from a highly unusual regression with only three right-hand variables and only five observations. Hence, they could well be biased because of omitted variables or because the five sites on which they are based are not representative of a randomly drawn sample of sites. While it is easy to see potential drawbacks to running the regression reported in the first column of Table 3, it is important to recognize that doing this is quite analogous to conducting a narrative synthesis based on a comparison of only a few sites. In both instances, the sites may not be representative and the restricted degrees of freedom limit the number of explanatory variables that can be examined at a time. Consequently, both approaches have the potential to produce misleading policy conclusions. Moreover, narrative synthesis has the added disadvantage of not permitting formal tests of statistical significance.

One way to examine how much confidence one should place on the point estimates in the first column of Table 3 is to test how robust they are to changes in the sample of sites on which they are based and to the inclusion of additional explanatory variables in the regression model. The regression findings reported in the last three columns of Table 3 provide such tests.

The regression reported in the second column of Table 3 uses the same three variables as the regression appearing in the first column, but is based on the seven available NEWWS sites, rather than the five GAIN sites. Only the coefficient on program effect on participation in job search continues to have the same sign and to be of the same order of magnitude. Moreover, the second macro regression explains much less of the cross-site variation. Thus, our findings for GAIN are not very robust to a change in the sample of sites.

The third column in Table 3 shows findings for the three explanatory variables when the 12 available GAIN and NEWWS sites are combined. The coefficient on the unemployment rate variable once again becomes negative and the coefficient on the job search variable becomes statistically significant at conventional levels (p = 0.02) and is of roughly the same magnitude as in the first two of the macro regressions.

So far, we have limited ourselves to only three explanatory variables. However, there are certainly others that may be important. By combining the GAIN and NEWWS sites, we are able to estimate a macro regression that includes a much fuller set of explanatory variables. This regression is shown in the last column in Table 3. Taken at face value and ignoring statistical significance, the findings imply that a program’s emphasis on quick employment does not matter very much and its achievement in increasing participants’ exposure to job search and education and training actually reduces its effect on earnings. According to these results, it is not the design of Portland NEWWS and Riverside GAIN programs—for example, their emphasis on quick employment and the use of job search—that resulted in such large earnings gains. Rather, if anything matters, it is the fact that both sites had mostly white, non-Hispanic participants and, in the case of the Riverside GAIN program, a fair number of welfare applicants. Again, however, it is difficult to place much confidence in these findings. None of the coefficients in the fourth column are statistically significant at conventional levels. Moreover, as a comparison with the third column indicates, they are not very robust—for example, the coefficients on the unemployment rate and job search variables both change sign as additional explanatory variables are added to the macro regression.

3. THE LACK OF PRECISION OF THE MACRO PARAMETERS

Given the sensitivity of the macro regression coefficient estimates in Table 3 to changes in the sites and variables included in the macro regressions and the lack of precision with which most of them are estimated—that is, the fact that the standard errors are large relative to the estimated coefficients—it is evident that great caution should be exercised in drawing policy inferences from them. Even greater prudence is needed in basing policy on informal narrative synthesis, however, because evidence on reliability comparable to that provided by the coefficient standard errors in Table 3 is rarely available. In this section, we examine the reasons for the lack of precision of the estimates of the macro regression coefficients appearing in Table 3 and what might be done about it.

Precision of Micro Parameters

Because the precision of the macro estimates depends in part on the precision with which the effects at each site are estimated, we examine this topic first. The precision of an estimated effect for site *j* ($\hat{\theta}_j$) is given by[14]

$$\sigma_j^2 \equiv var(\hat{\theta}_j | \theta_j) = \frac{\sigma_{e_j}^2 / s_{P_j}^2}{(n_j - K_1)(1 - \overline{R}_j^2)}, \qquad (3)$$

where $\sigma^2_{e_j}$ is the variance of the individual error e_{ij} in site j, n_j is the total number of program participants and controls in site j, K_1 is the number of regressors in the micro model (excluding the constant term), $s^2_{P_j}$ is the variance of the participation indicator in site j,[15] and $\overline{R}^2_j$ is the variance explained by a regression of the participation indicator on the other variables in the micro model (the vector *X*), as measured by the corrected R^2 statistic. Thus $\overline{R}^2_j$ measures the degree of multicollinearity between the participation indicator and *X*. It should be close to zero if individuals are assigned randomly to the participant and control groups.

Although the formula for the variance of the estimated program effect looks complicated, it represents a number of concepts that are probably familiar. The first term in the numerator indicates the dispersion of the outcome that is being analyzed. The smaller the dispersion, the greater will be the precision of a program effect estimate. For example, if a group of people have very similar earnings levels, then a small change in earnings for people who enter a program will be much easier to detect than if the group has many people with very low earnings and many people with high earnings.

The second term in the numerator implies that the precision of the estimated effect will be greater the more equal the distribution of sample members between participants and nonparticipants (or program and control group members). For any given sample size, having, say, only 10 percent of the sample assigned to a program makes any estimate of the resulting program effect less precise because it makes the estimated outcome for program participants less precise. The precision of the estimated impact is maximized when the sample is divided half and half between those in the program and those who are not.

Turning to the denominator, the first term (n_j - K_1) implies that large samples, such as those at the GAIN and NEWWS evaluation sites, yield more precise estimates of a program's effects. The second term in the denominator ($1-\overline{R}^2_j$) implies that the better one can predict who in the sample will be in the program, the harder it will be to estimate the effects of a program. For example, if all Hispanics in the sample are assigned to the program group and all non-Hispanics are assigned to the control group, then it is impossible to estimate an effect of the program because it is impossible to tell whether differences between the two groups stem from the program, or from underlying differences in the earnings levels of Hispanics and non-Hispanics. As suggested above, because the GAIN and NEWWS evaluations are based on random assignment, the X_i should not help predict whether someone was in a program group or a control group. Hence, $\overline{R}_j^2$ should approximate zero and the last term in the denominator should be close to 1.

Precision of Macro Parameters

The formula for the precision of a macro parameter is substantially more complicated than the formula for the precision of an estimated program effect (Bryk and Raudenbush, 1992). To simplify the discussion, we assume:

- the number of individuals in the sample at each site is identical (equal to *n*);
- the variance of the individual error is identical in all sites (equal to σ_e^2);
- the fraction of individuals assigned to the control group at each site is identical and, hence, the variance of the participation indicator is identical in all sites (equal to s_P^2);
- individuals are assigned randomly to the participant and control groups and, thus, $\overline{R}_j^2 = 0$.

Under these assumptions, the precision of $\hat{\theta}_j$ is identical in all sites (see equation 3), and the formula for the precision of the macro parameter $\hat{\gamma}_f$, the parameter associated with variable *f* in the vector *F*, is given by

$$\sigma_f^2 \equiv var(\hat{\gamma}_f \mid \gamma_f) = \frac{\dfrac{\sigma_W^2}{(J - K_2)} + \dfrac{\sigma_e^2 / s_P^2}{(J - K_2)(n - K_1)}}{s_{F_f}^2 (1 - \overline{R}_f^2)}, \qquad (4)$$

where *J* is the total number of sites, K_2 is the number of regressors in the macro model (excluding the constant), σ_W^2 is the variance of the error in the macro equation, $s_{F_f}^2$ is the variance of the regressor corresponding to the macro parameter γ_f, and $\overline{R}_f^2$ is the variance explained by a regression of the regressor *f* on the other variables in the macro model, as measured by the corrected R^2 statistic. This statistic captures the level of multicollinearity among the variables in the macro model.[16]

The numerator in equation 4 reflects the dispersion of the effects of programs across sites, and is divided into two parts. The term to the left of the plus sign reflects the proportion of the "true" program effects that cannot be explained by the observed program, client, and community characteristics. As this proportion increases, the estimate of the macro parameter becomes less precise. The term to the right of the plus sign reflects dispersion in estimated program effects that stems from how precisely they were estimated. If, as discussed earlier, sample sizes in each site are small, the estimates of program effects in each site will be relatively imprecise, and it will be more difficult to reliably attribute differences in them across sites to program and community characteristics. Both terms in the numerator are divided by the factor J-K_2. This takes account of the obvious point that macro parameters cannot be estimated unless there are fewer explanatory characteristics (K_2) than sites (J) and that they will be more precisely estimated the greater the amount by which the number of sites exceeds the number of explanatory characteristics.

The denominator in equation 4 also has two terms, both of which are similar to the factors that affect the precision of the micro parameters shown in equation 3. The first term ($s^2_{F_f}$) represents the variance of *f*, the macro variable of interest. It shows something that is commonly known about regressions: the more an explanatory characteristic varies among observations, the more precisely its effect can be estimated. If, for example, all the sites had the same unemployment rate, the variance in unemployment rates across sites would be zero, and it would then be impossible to determine how the level of unemployment influences program effects. Indeed, under these circumstances, differences in unemployment rates would not be responsible for any differences in the effectiveness of programs across sites. The second term in the denominator ($1-\overline{R}^2_f$) captures the magnitude of the collinearity between the macro variable *f* and all the other explanatory variables. As equation 4 implies, the effect of *f* will be more precisely estimated the lower its correlation with the other variables in the macro equation. The reason for this is that *f* would incorporate more information that is not captured by the other variables.

Variance of Macro Parameters in GAIN and NEWWS

Table 4 presents the calculations that go into the variance formula (equation 4), as well as the variances themselves, for the three variables that are common to the four macro regressions presented in Table 3. Estimates of the standard errors of the coefficients in Table 3, which are simply the square roots of the variances, also appear in Table 4. These calculations are all based on the site and program values presented in Table 2. Thus, Table 4 indicates why the macro parameters in Table 3 are not more precisely estimated. More generally, it highlights the characteristics of a collection of evaluation sites that make it more or less difficult to identify the determinants of differences in program effects.

As indicated earlier, equation 4 is based on several simplifying assumptions. The calculations presented in Table 4 are based on the same assumptions. Specifically, it is assumed that the sample was equally distributed between program and control groups, that the sample was equally distributed across the sites, and that the variance in unobserved individual factors ($\sigma^2_{e_j}$) was similar across the sites. The third of these assumptions is not inconsistent with the observed significance levels reported for the GAIN and NEWWS programs, but the first two are not valid for either GAIN or NEWWS. Consequently, the estimates of standard errors that appear in Table 4 differ from those in Table 3, although they are quite similar.

The two top panels in Table 4 show the calculations that determine the two terms in the numerator of the macro variance equation 4. The first term in the numerator depends, in part, on the proportion of the variance among the program effects estimates that is not explained by program, client, and community characteristics (σ^2_W). As indicated by the adjusted R^2 in Table 3, almost all the variance in the cross-site effects in GAIN is captured by the three macro variables. By itself, as implied by the first

TABLE 4
Calculation of Estimated Variances of Macro Parameters

	GAIN Alone	NEWWS Alone	GAIN and NEWWS Combined Sparse Model	GAIN and NEWWS Combined Full Model
1st term in numerator				
Variance of effects (σ^2_w)	484	42,704	81,951	13,860
Number of sites (J)	5	7	12	12
Number of macro parameters (K_2)	4	4	4	10
J-K_2	1	3	8	2
$\sigma^2_w/ (J\text{-}K_2)$	484	14,235	10,244	6,930
2nd term in numerator				
Variance of error in micro equation (σ^2_e)	28,619,171	28,619,171	28,619,171	28,619,171
Variance of program participation (σ^2_p)	0.25	0.25	0.25	0.25
Sample size (n)	3,798	4,170	3,998	3,998
$[\sigma^2_e/\sigma^2_p\ ]/[(J\text{-}K_2)(n)]$	30,141	9,151	3,579	14,317
1st term in denominator variance of covariate ($s^2_{F_f}$)				
County unemployment rate	14.3	8.6	10.4	10.4
% staff emphasizing quick employment	630.6	383.2	692.0	692.0
Effect on job search participation	95.4	75.9	76.3	76.3
2nd term in denominator adjusted R^2 ($\overline{R}^2_f$)				
County unemployment rate	-0.834	0.307	-0.082	0.842
% staff emphasizing quick employment	-0.407	0.313	-0.015	0.258
Effect on job search participation	-0.515	-0.370	0.117	0.279
Variance of macro coefficient (σ^2_f)				
County unemployment rate	1167	3941	1227	12895
% staff emphasizing quick employment	35	94	20	41
Effect on job search participation	212	225	205	386
Standard error of coefficient (σ_f)				
County unemployment rate	34	63	35	114
% staff emphasizing quick employment	6	10	4	6
Effect on job search participation	15	15	14	20

row of the first panel in Table 4, this means that macro parameters estimated for the GAIN sites alone are more precise than those estimated either for the NEWWS sites alone or for the combined sites. The second term in the numerator depends, in part, on the variance in the error in the micro regressions estimated for each site ($\sigma^2_{e_j}$), which, as indicated by the first row of the second panel, is assumed to be equal across the sites. It is also influenced by how the sample was distributed between the program and control groups. As previously indicated, it is assumed that half the sample was assigned to each group in each of the sites. Thus, the variance of program participation (σ^2_p) equals .25.

Everything else equal, macro parameters based on the GAIN sites alone will be less precisely estimated than those based on the NEWWS sites alone or on the combined sites because there are only five GAIN sites available for estimating the macro equations, but seven NEWWS sites and 12 combined sites can be used. When the macro equation is limited to only three explanatory variables, $J–K_2$ is only 1 for GAIN, but 3 for NEWWS and 8 for GAIN and NEWWS combined. Since $J–K_2$ is divided into both terms in the numerator, this means that all else equal, the variance in the estimated macro parameters are three times larger for GAIN than for NEWWS and eight times larger for GAIN than for GAIN and NEWWS combined. Thus, the ability to understand the influence of community, client, and program characteristics on program effect sizes can be markedly increased by a small increase in the number of sites. Indeed, the one statistically significant macro coefficient in Table 3 was estimated after the GAIN and NEWWS sites were combined.

As a comparison of the last two columns of either Table 3 or Table 4 suggests, increasing the number of macro parameters can also substantially increase the imprecision of the estimated parameters when this increase is large relative to the number of sites. When all the variables shown in Table 2 are included in the macro regression, the number of parameters (which includes the constant term) increases from four to ten. Because there were 12 usable sites in GAIN and NEWWS combined, $J–K_2$ is only one-fourth as large with the full set of parameters and, therefore, the second term of the numerator of the variance formula is four times as large.

Two factors affect the denominator of equation 4: the variance of the particular macro variable being examined and the ability of the other variables to explain that variation (in other words, the collinearity among the macro variables). The two middle panels of Table 4 show the values of these two factors for the each of the three variables being examined.

The third panel in Table 4 indicates that these three macro characteristics have a somewhat larger variance across the five GAIN sites than across the seven NEWWS sites. This may reflect the fact that the GAIN sites include both large urban counties and smaller counties, while the NEWWS sites are all in large urban areas. Most sites in both GAIN and NEWWS had modest unemployment rates, but the highest unemployment in GAIN (15.3 percent) is substantially larger than the highest unemployment rate in NEWWS (11.9 percent). Likewise, the variation in staff emphasis on quick employment is much greater in GAIN than NEWWS. As indicated in Table 2, nearly all staff in Riverside GAIN emphasized quick employment, while only about one in five in Alameda and Butte did so. In contrast, in each of the NEWWS sites half or more of staff emphasized quick employment. Finally, there is apparently somewhat greater variation in program effects on job search across the GAIN sites than across the NEWWS sites.[17] By itself, the greater variance of the three macro variables across the GAIN sites means that it would be easier to detect the effects of these characteristics on program effects in GAIN than in NEWWS.

Collinearity among the macro variables is measured by the adjusted R^2's of regressions of each variable on all the others that are used in each macro equation. The values of these adjusted R^2's appear in the fourth panel of Table 4. In GAIN, there is little relationship between the three variables listed in Table 4. As a result, the adjusted R^2's are actually negative. In contrast, in NEWWS, the other two characteristics "explain" about 30 percent of the variation in unemployment rates and in staff emphasis on quick employment. This means that, all else equal, it is more difficult to detect the independent effect of these two macro factors in NEWWS than in GAIN.

Of course, other things are not equal. Based on equation 4, the fifth panel in Table 4 presents computations of the variance of the macro parameter estimates and, thereby, incorporates the effects of all the factors discussed above. Other things equal, the smaller number of sites in GAIN than in NEWWS would cause the variance of the estimated macro parameters for GAIN alone to be larger than for NEWWS alone. However, the greater independence, variance, and explanatory power of the macro characteristics in GAIN work in the other direction and increase the precision of the GAIN parameter estimates compared to those for NEWWS. The independence, variance, and explanatory power effect is apparently stronger than the sample size effect, for the variance of the effects of the unemployment rate and staff emphasis on quick employment are smaller for GAIN than for NEWWS. Hence, as shown in the bottom panel of Table 4, the standard errors of the coefficients are also smaller. For the third variable, program effects on job search, the ability of the two other macro variables to explain variation across sites is about as poor in GAIN as in NEWWS. As a result, the variance and standard error of the estimated effect of job search are similar for the two sets of sites.

As anticipated, a comparison of the second and third columns in Table 4 implies that the precision of the estimated macro parameters is much greater when the GAIN and NEWWS sites are combined than when looking at the NEWWS sites separately. This difference stems largely from the greater number of available sites when they are combined. However, a comparison of the first and third columns suggests that the variances and standard errors of the three macro parameters are similar for the GAIN sites alone and the combined sites, even though the latter estimates are based on a larger number of sites. There are two explanations for this surprising result. First, the three macro variables explain much more of the variation in effects for GAIN alone than they do when the sites are combined, and, second, there is virtually no collinearity among the three macro variables for the GAIN sites alone.

The last comparison in Table 4 is between the two macro regressions that use the full set of 12 available sites, one with only three macro variables and the other with the full set of nine macro characteristics. As shown, the numerator of the macro variance is substantially smaller with the smaller set of characteristics. Moreover, as would be expected, collinearity increases as more variables enter macro regressions. For example, the adjusted R^2 for the county unemployment rate is essentially zero when only the percentage of staff emphasizing quick employment and program effect on participation in job search are used as regressors, but exceeds 0.8 when all eight of the other variables are used. All else equal, this means the variance of the estimated effect of the unemployment rate would be five times larger in the full regression than in the constrained regression (1 – 0.8 = 0.2 is one fifth as large as 1 – 0 = 1), and, in practice, the difference is even greater than fivefold because the number of macro characteristics is nearly as large as the number of sites in the full regression. As a result, the variance is about ten times greater in the full regression than in the constrained regression (12,895 vs. 1,227), so that the standard error of the estimated effect of unemployment is about three times greater in the full regression.

Increasing the Precision of Macro Parameters

Equation 4 indicates that one way in which the precision of estimates of the macro parameters can be increased is by increasing the number of sites on which they are based. This possibility is investigated in Table 5. The first column in each set of three columns contains the estimated macro regression coefficients reported in Table 3 for the 12 available GAIN and NEWWS sites. The third column in each set uses equation 4 to compute the standard errors that would result if the number of sites were doubled, tripled, or quadrupled, from 12 to 24, 36, or 48, but the other parameters in the equation remained unchanged (i.e., the number of regressors in the micro and macro models, the variance of the error in the macro equation, the sample size used to estimate the program effects on earnings, the variance of each of the macro variables included in the macro regression, and the level of multicollinearity among the variables in the macro model). The middle column in each set of three columns provides a power test; each of the estimates appearing in the column is the minimum value of the macro regression coefficient that would be needed to be considered statistically significant at the 5 percent level, given the corresponding standard error.

The key finding is that the magnitude of the required value declines relatively rapidly as the number of sites increases. In other words, doubling, tripling, or quadrupling the number of sites would

TABLE 5
Influence of the Number of Sites on the Minimum Detectable Effect of Selected Macro Variables

	SPARSE MODEL			FULL MODEL		
Variable	Estimated Effect	Minimum Detectable Effect[a]	Standard Error	Estimated Effect	Minimum Detectable Effect[a]	Standard Error
Unemployment Rate						
12 sites	-54.39	-76.94	33.37	187.06	459.98	106.91
24 sites		-44.02	21.10		86.66	40.41
36 sites		-33.98	16.68		60.95	29.65
48 sites		-28.67	14.23		49.65	24.53
Percent of Staff Who Emphasized Quick Employment						
12 sites	-0.59	10.36	4.49	5.37	29.08	6.76
24 sites		5.93	2.84		5.48	2.55
36 sites		4.58	2.25		3.85	1.87
48 sites		3.86	1.92		3.14	1.55
Program Effect on Participation in Job Search						
12 sites	36.31	30.58	13.26	-12.68	-82.95	19.28
24 sites		17.49	8.39		-15.63	7.29
36 sites		13.50	6.63		-10.99	5.35
48 sites		11.39	5.65		-8.95	4.42
Program Effect on Participation in Education and Training						
12 sites				-34.96	-68.61	15.95
24 sites					-12.93	6.03
36 sites					-9.09	4.42
48 sites					-7.41	3.66
Percent Black						
12 sites				-15.11	-26.18	6.09
24 sites					-4.93	2.30
36 sites					-3.47	1.69
48 sites					-2.83	1.40
Percent Hispanic						
12 sites				-68.64	-109.64	25.48
24 sites					-20.66	9.63
36 sites					-14.53	7.07
48 sites					-11.83	5.85
Percent Applicants						
12 sites				201.57	354.59	82.41
24 sites					66.81	31.15
36 sites					46.98	22.86
48 sites					38.27	18.91

table continues

TABLE 5, continued

	SPARSE MODEL			FULL MODEL		
Variable	Estimated Effect	Minimum Detectable Effect[a]	Standard Error	Estimated Effect	Minimum Detectable Effect[a]	Standard Error
Percent Long-Term Recipients						
12 sites				22.27	98.65	22.93
24 sites					18.59	8.67
36 sites					13.07	6.36
48 sites					10.65	5.26
Net Program Cost						
12 sites				0.32	0.97	0.23
24 sites					0.18	0.09
36 sites					0.13	0.06
48 sites					0.11	0.05

[a] Minimum effect size that would be statistically significant at the 5 percent level, given the number of sites.

greatly increase the probability of obtaining estimates of macro parameters that are statistically significant at conventional levels such as 5 percent.

Each of the coefficients that we estimated with 12 available sites—the figures in the first column of each set of columns—might, of course, increase or decrease in magnitude as the number of sites increases and, hence, their statistical precision increases. Indeed, the sign of any given coefficient could change. Nonetheless, the coefficient estimates provide a rough benchmark that can be compared to the values that would just reach the 5 percent level of significance, which are reported in the middle column in each set of columns. As can be seen, with 36 sites, or even 24, most of the latter values are considerably smaller in absolute value than the coefficients that were actually estimated. Thus, although this comparison is merely suggestive at best, it implies that if 20 or 30 sites were available, instead of only 12, it would probably be possible to obtain several statistically significant estimates of macro regression coefficients.

4. CONCLUSIONS

The most important lesson from the research described in this article is that great care should be exercised in drawing conclusions as to the reasons why some employment and training programs are apparently more successful than others. One approach to imposing discipline on drawing such conclusions is to estimate macro regression equations that test whether apparent relationships between program effect sizes and program design, client characteristics, and local economic conditions are statistically significant.[18]

We did this and found that we were unable to draw conclusions in which we had confidence. For example, while we found some evidence that welfare-to-work programs can increase their effect on earnings by increasing the extent to which those enrolled utilize the job search services the programs provide, this result was not robust to the addition of other explanatory variables to the macro regressions.

We examined a number of other potentially important explanatory variables, including measures of client characteristics, site economic conditions, and program design, in our attempt to learn more about why government-funded training programs are more successful in some sites than in others, but we were unable to obtain coefficients that are statistically significant at conventional levels. Learning more about why some programs have larger effects than others requires that steps such as the following be taken to obtain more reliable coefficient estimates:

- Add more sites. The findings in the preceding section suggest that increasing the number of evaluation sites is crucial, but that the size of the necessary increase may be fairly modest. Perhaps as few as 20 or 30 sites in total would suffice. There appear to be sufficient candidates for inclusion. The U.S. currently has around 1,600 local welfare offices, and training programs funded by the national Workforce Investment Act are administered by over 600 local agencies. Moreover, as the number of evaluation sites increases, it may be possible to decrease the sample size at each site, thereby partially offsetting the cost of adding sites.[19]
- Allow for subgroup effects. The analysis in this paper is based on a measure of overall program effect in each site. However, a program may affect different types of persons differently. Consequently, it may be desirable to estimate separate program effects for subgroups of program participants defined on the basis of, for example, demography, time on welfare when assigned to a training program, prior educational achievement, or previous work experience. It is possible to do this as long as membership in the subgroups is not affected by the program, as would be the case, for example, if subgroups were defined on the basis of labor force status after the program began. When effects do vary by group, allowing for separate effect size measures reduces the residual variance in the micro equation and improves the precision of estimation of macro equation parameters.
- Refine the program measures. Available measures of participation in program activities are far from perfect. For example, individuals are counted as participants in job search and education and training even if they took part in these activities for as little as one day. Thus, intensity of participation is

not measured.[20] As a result, the program-related regressors in the macro equation contain errors, and the related coefficient estimates are generally biased downward. Greater attention needs to given to obtaining measures that usefully and accurately describe the nature of the services provided in different evaluation sites. Doing this may require developing better descriptors for training programs and closer observation of what participants actually do.

- Reduce treatment variation. Equation 4 implies that the precision of the macro estimates can be increased by constraining variation, if possible, in program effects associated with factors unaccounted for in the macro equation (this would diminish σ_w^2). This, for example, might involve restricting the extent to which individual evaluation sites vary in the manner in which they implement the policy interventions to be evaluated.
- Constrain adaptive response. Equation 4 also implies that the precision of the macro estimates can be increased by assigning variation in interventions randomly to sites whenever possible and by minimizing adaptive responses by sites to the treatment to which they are assigned. Both of these steps would reduce multicollinearity (i.e., they would decrease $\overline{R}_f^2$).

The last two suggestions obviously require some imposition of conditions on the state and local agencies that administer the programs to be evaluated. In the numerous evaluations of government-funded training programs conducted in the 1980s and 1990s, however, state and local welfare agencies have typically exercised great discretion both in determining what sorts of programs to implement and with respect to whether to participate in evaluations of these programs (U.S. Department of Health and Human Services, 1997). Often, they even determined the rigor with which they were evaluated (for example, whether random assignment was used).

All this leads to recognition that moving beyond single-site evaluations in a responsible way presents a considerable political, as well as technical, challenge. Some way must be found to create the sense of common purpose that might lead to multiplication of sites and greater rigor in implementation.

Endnotes

[1] As discussed later, a test made by the JTPA evaluators indicated that the cross-site variation in the effect estimates was not statistically significant. Thus, it is not surprising that attempts to find correlates of the cross-site variation were not successful. A similar test was not made in the Food Stamp evaluation.

[2] An existing methodology that has much in common with multilevel analysis, but puts particular emphasis on techniques for treating interstudy differences in methods and in outcome and input measures, is known as meta-analysis (see Hedges and Olkin, 1985; Hunter and Schmidt, 1990; Rosenthal, 1991; Cook et al., 1992; and Cooper and Hedges, 1994).

[3] An alternative presentation strategy would be to combine the micro and macro equations into a single equation that contains individual and site-level data. Indeed, it is often convenient to estimate multilevel models in this combined form, and it is common among economists to do so. The major advantage of the multilevel framework is that it allows us explicitly to contrast evaluation strategies designed to estimate site-specific treatment effects with strategies designed to estimate the determinants of those effects.

[4] Equation 1 and the equations that follow involve vector multiplication. For simplicity of notation, we have eliminated specific notational reference to the necessary vector transpositions.

[5] The primary advantage of the random assignment approach is that, in principle, it guarantees that treatment and control groups are drawn from the same population (Burtless and Orr, 1986). Random assignment is a key element in the methodology employed by the Manpower Demonstration Research Corporation in its widely cited studies of welfare-to-work programs (see Gueron and Pauly, 1991; Greenberg and Wiseman, 1992; and Friedlander, Greenberg, and Robins, 1997, for descriptions of these studies as well as other welfare-to-work studies) and has frequently been used by other research

organizations as well for evaluating employment and training programs. A critical assessment of the role of random assignment in social program evaluation is presented in Heckman and Smith (1995), while a vigorous defense of the technique can be found in Burtless (1995).

[6] More efficient estimates of program effects (θ_j) can be obtained by estimating equation 1 as is, rather than using the simple difference in site means. However, it is not essential to include *X* in the model if individuals are assigned randomly to the treatment and control groups since random assignment implies that *X* and P_{ij}, the participation indicator, are uncorrelated.

[7] For presentational convenience, we use the terms "programs" and "sites" interchangeably in the remainder of this article. Used in this way, the number of "sites" can exceed the number of locations. For example, three of the programs listed in Table 1 operated in California's Riverside County.

[8] Table 1 presents estimates of program effects on earnings for a single year, the second year after random assignment. Results in Hotz, Imbens, and Klerman (2000) indicate that when program effects on earnings are summed over the nine-year period after random assignment, the relative superiority of the GAIN program in Riverside over the GAIN programs in Alameda, Los Angeles, and San Diego shrink. It does not disappear, however.

[9] In Section 1, we mention the only two previous attempts to estimate a macro model in evaluating employment and training programs with which we are familiar: the JTPA study and the Food Stamp Employment and Training Program Evaluation. In addition, James Riccio, Howard Bloom, and Carolyn Hill (2000) are conducting a study using hierarchical modeling at the office level to investigate how differences in program administration influence differences in program effects. Like our study, their investigation focuses on the GAIN and NEWWS evaluations. However, their study is conducted at a considerably less aggregate level than ours. Also, Heinrich and Lynn (2000a and 2000b) have used multilevel analysis to explore issues that arise in administering the JTPA program. They focus on

explaining variation in earnings outcomes across individual program participants and across program sites. In this paper, in contrast, we focus on examining differences in program effects across sites. These effects are estimated at each site by comparing the postprogram earnings of program participants with those of control group members who did not participate in the evaluated programs. Heinrich and Lynn did not use a control group in conducting their analysis. The multilevel model outlined here has rarely been used in evaluations of employment and training programs, but it has been more often used elsewhere (see Cook et al., 1992, for several examples). Its use is particularly common in education evaluations, most notably in the work of Hedges (1982a, 1982b). Stigler (1986, cited in Hedges, 1992) reports discovery of structurally similar multilevel analyses of multiple research studies in nineteenth-century astronomy. Rubin (1992) refers to the macro equation as the "effect-size surface."

[10] See Rubin (1992) for a similar proposition stated within the framework of meta-analysis.

[11] Throughout this article, we rely on conventional levels of statistical significance, such as 1 percent or 5 percent, to test null hypotheses. However, one might argue that a more lenient test should be employed in determining whether program effects differ from one another. After all, the 95 percent confidence interval is not Holy Writ. If the effect of a program in one site appears larger than the effect in another, but the difference is not significantly different using conventional standards, this may be the only information available, and some further investigation of possible reasons for observed differences in mean effects may be justified. Our point is that the degree of uncertainty about such inference seems rarely to be appreciated. As a consequence, exploratory speculations about causality are easily transmogrified to "lessons" that serve as real bullets in the policy wars.

[12] An important implication of multilevel analysis (as well as meta-analysis) is that it is possible, in principle, to estimate the macro parameters (γ) with great precision even when it is not possible, due to inadequate sample sizes at each site, to estimate the individual program effect parameters (θ_j)

precisely. Doing so, however, requires that the macro equation be based on a sufficient number of observations (i.e., sites). In the case of the JTPA evaluation, 16 sites were apparently not sufficient.

[13] These regressions are estimated with OLS. However, because of the potential for hetroskadasticity, the program effect estimates were weighted by the inverse of the standard error of the program effect estimates, and the regressions appearing in Table 3 were reestimated by GLS. The coefficients from the unweighted OLS regressions and the weighted GLS regressions are virtually identical. To take account of the fact that the program effect at each site is not precisely estimated, each of the macro regressions in Table 3 was estimated 1,000 times. In making each estimate, a random error term was added to each program effect estimate based on the normal distribution implied by their standard errors. The macro parameter estimates that appear in Table 3 are the means of these 1,000 estimates, while their standard errors were computed as the square root of the sum of the variance of the 1,000 iterations plus the variance from a regression run without adding an error term.

[14] Equation 3 is written as it stands to highlight the consequences of multicollinearity, sample size, and so forth. To see that the equation is correct, note that the variance of $\hat{\theta}_j$ is given by $\sigma^2_{e_j}/ RSS_j$, where RSS_j is the residual sum of squares from an auxiliary regression of P_{ij} on X_{ij} for site j (Maddala, 1988, p. 101). The equation follows automatically from the definition of corrected R^2, $\overline{R}^2 = 1 - [RSS_j / (n_j - K_1)] / S^2_{P_j}$.

[15] Note that the variance of the participation indicator in the population is given by $s^2_{P_j} \equiv s_j(1 - s_j)n_j/(n_j - 1)$, where s_j is the fraction of individuals in the participant group, $(1 - s_j)$ is the fraction of individuals in the control group, and $n_j/(n_j - 1)$ reflects the standard (although in this case unnecessary, given the large sample sizes reported in Table 1) adjustment for the loss of a degree of freedom in computing a sample variance. We use the standard formula for a sample variance to be

consistent with the definition of $\overline{R}^2$ (see the previous endnote). If the participant and the control groups are the same size, then $s^2_{P_j} \approx 0.25$, which is the largest possible value.

[16]The corrected R^2 statistic is used to measure multicollinearity because it is not affected by sample size, as in the case of the standard R^2 statistic. Note that, unlike the uncorrected R^2 statistic, it is technically possible for $\overline{R}^2_f$ to be less than zero.

[17]This finding is somewhat surprising. The inclusion in NEWWS of the LFA programs—which required nearly all participants to look for work initially—and the HCD programs—which required nearly all participants to enroll in education or training initially—provided a broader mix of education and job search in NEWWS than in GAIN. Thus, we anticipated that there would be greater variance in program effect on job search in NEWWS. In fact, there was considerably greater variation in program effect on participation in education across the NEWWS sites than across the GAIN sites, but less in program effect on job search.

[18]Another approach is to use random assignment to test alternative program designs at the same site. For example, the NEWWS evaluation is rigorously comparing the work-first and human capital approaches at three sites (Riverside, Atlanta, and Grand Rapids) by randomly assigning households to one of three groups: a labor force attachment program, a human capital development program, or a control group. Though this technique is very useful and should be viewed as complementary to the estimation of macro parameters, it is limited to simple comparisons between two types of programs. For example, it does not allow one to examine whether differences in economic conditions matter or to estimate the effects of relatively small increases in participation in job search or training.

[19]For a discussion of the trade-off between the number of sites and the number of observations per site, see Greenberg, Meyer, and Wiseman (1993). Some evidence on the trade-off between the

number of sites and the number of observations is suggested by the national evaluation of JTPA. Combining the costs of the two evaluation firms (Abt and MDRC) involved plus operations payments made to the 16 evaluation sites produces a cost estimate of approximately $239,000 ($346,000 in year 2000 dollars) for one additional site (holding the total number of observations constant) and a cost of $384 ($556 in year 2000 dollars) for adding an observation at an existing site. These estimates are based on a regression that uses cost data provided by Larry Orr of Abt Associates and Fred Doolittle of MDRC and the 16 evaluation sites as observations. The site subsidy payments were regressed on a constant term, a variable measuring site observation counts, and a term representing the number of organizations at each site that were involved in random assignment. The adjusted R^2 for the regression is .94.

[20] In addition, participation in program activities is measured over several years. By that time, even programs that emphasized education and training (e.g., Alameda GAIN) had a considerable effect on job search simply because people had graduated from education and training and were ready to seek work. Moreover, some programs may require participants to seek jobs first and provide them with education and training only if they fail to find employment, while other programs may provide education and training first and job search afterward. The participation measures do not necessarily reflect such differences in program philosophy.

References

Amemiya, Takeshi. 1978. "A Note on a Random Coefficients Model." *International Economic Review* 19: 793–796.

Barnow, Burt S., and Christopher T. King, eds. 1999. *Improving the Odds: Increasing the Effectiveness of Publicly Funded Training*. Washington, DC: Urban Institute Press.

Bell, Stephen H., and Larry L. Orr. 1994. "Is Subsidized Employment Cost Effective for Welfare Recipients? Experimental Evidence from Seven State Demonstrations." *Journal of Human Resources* 29: 42–61.

Blank, Rebecca. 1994. "The Employment Strategy: Public Policies to Increase Work and Earnings." In *Poverty and Public Policy: What Do We Know? What Should We Do?*, edited by S. H. Danziger, G. D. Sandefur, and D. H. Weinberg. Cambridge, MA: Harvard University Press.

Burtless, Gary, and Larry L. Orr. 1986. "Are Classical Experiments Needed for Manpower Policy?" *Journal of Human Resources* 21: 606–639.

Burtless, Gary. 1995. "The Case for Randomized Field Trails in Economic and Policy Research." *Journal of Economic Perspectives* 9: 63–84.

Bryk, Anthony S., and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models*. Newbury Park, CA: Sage Publications.

Cook, Thomas D., Harris Cooper, David S. Cordray, Heidi Hartmann, Larry V. Hedges, Richard J. Light, Thomas A. Louis, and Frederick Mosteller. 1992. *Meta-Analysis for Explanation: A Casebook*. New York: Russell Sage Foundation.

Cooper, Harris M., and Larry V. Hedges, eds. 1994. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

Freedman, Stephen, Daniel Friedlander, Gayle Hamilton, JoAnn Rock, Marisa Mitchell, Jodi Nudelman, Amanda Schweder, and Laura Storto. 2000. *Evaluating Alternative Welfare-to-Work Approaches: Two-Year Impacts for Eleven Programs*. Washington, DC: U.S. Department of Health and Human Services Administration for Children and Families and Office of the Assistant Secretary for Planning and Evaluation; and U.S. Department of Education Office of the Under Secretary and Office of Vocational and Adult Education.

Friedlander, Daniel, David H. Greenberg, and Philip K. Robins. 1997. "Evaluating Government Training Programs for the Economically Disadvantaged." *Journal of Economic Literature* 35: 1809–1855.

Goldstein, Harvey. 1995. *Multilevel Statistical Models*, 2nd edition. New York: John Wiley and Sons.

Gordon, Anne, and John Burghardt. 1990. *The Minority Female Single Parent Demonstration: Short-Term Economic Impacts*. New York: Rockefeller Foundation.

Greenberg, David, Marvin Mandell, and Mathew Onstott. 2000. "The Dissemination and Utilization of Welfare-to-Work Experiments in State Policy Making." *Journal of Policy Analysis and Management* 19: 367–382.

Greenberg, David, Robert Meyer, and Michael Wiseman. 1993. "Prying the Lid from the Black Box: Plotting Evaluation Strategy for Welfare Employment and Training Programs." Discussion Paper No. 999-93, Institute for Research on Poverty, University of Wisconsin–Madison.

Greenberg, David, Robert Meyer, and Michael Wiseman. 1994. "Multisite Employment and Training Program Evaluation: A Tale of Three Studies." *Industrial and Labor Relations Review* 47: 679–691.

Greenberg, David, and Michael Wiseman. 1992. "What Did the OBRA Demonstrations Do?" In *Evaluating Welfare and Training Programs*, edited by C. F. Manski and I. Garfinkel. Cambridge, MA: Harvard University Press. Pp. 25–75.

Gueron, Judith M., and Edward Pauly. 1991. *From Welfare to Work*. New York: Russell Sage Foundation.

Hamilton, Gayle, and Thomas Brock. 1994. *Early Lessons from Seven Sites*. Washington, DC: U.S. Department of Health and Human Services and U.S. Department of Education.

Hamilton, Gayle, Thomas Brock, Mary Farrell, Daniel Friedlander, and Kristen Harknett. 1997. *Evaluating Two Welfare-to-Work Program Approaches: Two-Year Findings on the Labor Force Attachment and Human Capital Development Programs in Three Sites*. Washington, DC: U.S. Department of Health and Human Services and U.S. Department of Education.

Hamilton, William L., Nancy R. Burstein, Elizabeth Davis, and Margaret Hargreaves. 1992. *The New York Child Assistance Program: Interim Report on Program Impacts*. Cambridge, MA: Abt Associates, Inc.

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

Hedges, L. V. 1982a. "Estimation of Effect Size from a Series of Independent Experiments." *Psychological Bulletin* 92: 490–499.

Hedges, L. V. 1982b. "Fitting Continuous Models to Effect Size Data." *Journal of Educational Statistics* 7: 245–270.

Hedges, L. V. 1984. "Advances in Statistical Methods for Meta-Analysis." In *Issues in Data Synthesis*, edited by W. H. Yeats and P. M. Wortman. San Francisco: Jossey-Bass. Pp. 25–42.

Hedges, L. V. 1992. "Meta-Analysis." *Journal of Educational Statistics* 17 (4): 279–296.

Hedges, L. V., and I. Olkin. 1985. *Statistical Methods for Meta-Analysis*. New York: Academic Press.

Heinrich, Carolyn J., and Laurence E. Lynn, Jr. 2000a. "Governance and Performance: The Influence of Program Structure and Management on Job Training Partnership Act (JTPA) Program Outcomes." In *Governance and Performance*, edited by C. J. Heinrich and L. E. Lynn, Jr. Washington, DC: Georgetown University Press.

Heinrich, Carolyn J., and Laurence E. Lynn, Jr. 2000b. "Means and Ends: A Comparative Study of Empirical Methods for Investigating Governance and Performance." Unpublished manuscript.

Hotz, Joseph V., Guido Imbens, and Jacob A. Klerman. 2000. "The Long-Term Gains from GAIN: A Re-Analysis of the Impacts of the California GAIN Program." Unpublished manuscript.

Hsiao, Cheng. 1986. *Analysis of Panel Data*. New York: Cambridge University Press.

Hunter, John E., and Frank L. Schmidt. 1990. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Newbury Park, CA: Sage Publications.

Kreft, Ita G. G., and Jan de Leeuw. 1998. *Introducing Multilevel Modeling*. Thousand Oaks, CA: Sage Publishing, Inc.

LaLonde, Robert J. 1995. "The Promise of Public Sector Sponsored Training Programs." *Journal of Economic Perspectives* 9: 149–168.

Maddala, G.S. 1988. *Introduction to Econometrics*. New York: Macmillan.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, et al. 1996. *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study*. Washington, DC: Urban Institute Press.

Puma, Michael J., Nancy R. Burstein, Katy Merrell, and Gary Silverstein. 1990. *Evaluation of the Food Stamp Employment and Training Program: Final Report*. Bethesda, MD: Abt Associates, Inc.

Riccio, James, Howard S. Bloom, and Carolyn J. Hill. 2000. "Management Organizational Characteristics, and Performance: The Case of Welfare-to-Work Programs." In *Governance and Performance*, edited by C. J. Heinrich and L. E. Lynn, Jr. Washington, DC: Georgetown University Press.

Riccio, James, Daniel Friedlander, and Stephen Freedman. 1994. *GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program*. New York: Manpower Development Research Corporation.

Rosenthal, Robert. 1991. *Meta-Analytic Procedures for Social Research*, revised edition. Newbury Park, CA: Sage Publications.

Rosenthal, Robert, and D. B. Rubin. 1982. "Comparing Effect Sizes of Independent Studies." *Psychological Bulletin* 92: 500–504.

Rubin, Donald B. 1992. "Meta-Analysis: Literature Synthesis or Effect-Size Surface Estimation?" *Journal of Educational Statistics* 17: 363–374.

Scrivener, Susan, Gayle Hamilton, Mary Farrell, Stephen Freedman, Daniel Friedlander, Marisa Mitchell, Jodi Nudelman, and Christine Schwartz. 1998. *Implementation, Participation Patterns, Costs and Two-Year Impacts of the Portland (Oregon) Welfare-to-Work Program*. Washington, DC: U.S. Department of Health and Human Services and U.S. Department of Education.

Snijders, Tom A. B., and Roel J. Bosker. 1999. *Multilevel Analysis*. London: Sage Publications, Ltd.

Stigler, S. M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press.

U.S. Department of Health and Human Services. 1997. *Setting the Baseline: A Report on State Welfare Waivers*. Washington, DC: U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation.