The study presents benefit-cost ratios for 14 disability cohorts served by the Vocational Rehabilitation (VR) Program. The earnings impacts are estimated in a quasiexperimental framework using an internal comparison group. The earnings data are from a unique national panel constructed by linking client data of the Rehabilitative Services Administration with earnings histories from the Social Security Administration. These earnings data accommodate a series of statistical tests that allow us to identify and control for the presence of selection bias when estimating treatment impacts. The results indicate that the VR program is cost-effective in general, although not universally so across specific disabilities.

EVALUATING THE VOCATIONAL REHABILITATION PROGRAM USING LONGITUDINAL DATA

Evidence for a Quasiexperimental Research Design

DAVID H. DEAN
ROBERT C. DOLAN
ROBERT M. SCHMIDT
University of Richmond

This study estimates the earnings gains for persons with work disabilities served by the public sector Vocational Rehabilitation (VR) Program. The earnings estimates are combined with actual program cost to obtain stratified benefit-cost ratios for 14 cohorts. The significance of this study is twofold. First, the findings are arguably the most authoritative national estimates of VR treatment effects to date. This is due in large part to the 1980 RSA-SSA DataLink, a unique national panel data set constructed by linking the VR client data of the Rehabilitative Services Administration (RSA) with earnings histories from the Social Security Administration (SSA). Second, the estimation procedure employed here touches on several issues surrounding the use of quasiexperimental research methods in evaluation. In this respect, the study contributes to the continuing debate on the role of quasiexperimental research designs in the evaluation of public training programs.

EVALUATION REVIEW, Vol. 23 No. 2, April 1999 162-189

The first section of the article discusses the significance of the study in greater detail. The second section provides an overview of the DataLink. The estimation procedure is outlined in the third section. This discussion examines the appropriateness of an "internal" comparison group for VR and estimates treatment impacts using a "fixed effects" model. The fourth section presents benefit-cost ratios for the VR program, and concluding remarks appear in the fifth section.

SIGNIFICANCE OF THE STUDY

The VR program is a state-federal partnership providing a wide range of employment-related services to persons with physical, mental, or emotional impairments.[1] Formal evaluation of VR by economists began with Ronald Conley (1969). That study marks the beginning of what we regard as first-generation analysis (Bellante 1972; Worrall 1978; Nowak 1983; Lewis et al. 1992). A defining deficiency in this literature is a simplistic calculation of earnings impacts necessitated by data limitations. Conventional VR data offer a client earnings profile that contains a maximum of two weekly earnings observations—at acceptance and after completion of the program. Accordingly, the net impact of VR services is calculated as the difference between earnings at acceptance and closure. For reasons that we have discussed at length elsewhere, this calculation is seriously flawed as a measure of earnings gains.[2]

The common deficiencies in the earnings profiles of trainees are addressed by the 1980 RSA-SSA DataLink. This earnings cross-match merges VR client data with annual SSA earnings between 1972 and 1988 for all cases that came to closure in 1980. Although the implementation and analysis of a DataLink are not unprecedented in VR evaluation, several features of this study distinguish it from its predecessors.

The first formal SSA DataLink to VR administrative files was analyzed under federal contract by Berkeley Planning Associates (BPA) (1989).[3] This DataLink matched a sparse set of client data for VR cases closed in the fiscal year of 1975 (July 1974 to June 1975), with SSA earnings between 1972 and 1983. However, given that the average service duration in VR is roughly 18 months, the 1975 DataLink provided preprogram earnings for only 1 year prior to referral for a large share of cases. This fact has two implications. First, earnings 1 year prior to treatment will likely encompass "preprogram dip." Second, as we shall demonstrate in this analysis, a longer period of preprogram earnings accommodates useful tests of the appropriateness of an

internal comparison group. The relevance of the 1975 DataLink analysis is further diminished by the Rehabilitation Act of 1974. This act changed VR's service mandate from serving "those most likely to succeed" to serving "those with the most severe disabilities." Given the typical service duration, the vast majority of the closed cases from the 1975 DataLink do not represent the more severely disabled caseload that VR has served for the past two decades.[4]

Although similar in concept to the 1975 DataLink, the advantage of the 1980 DataLink lies in its historical depth. The 1980 DataLink provides up to eight years of pre- and postprogram earnings for persons served by VR and a richer client profile. Although the General Accounting Office (GAO) (1993) conducted the first analysis of the 1980 DataLink, the GAO effort is poorly grounded in key respects. In general, the GAO methodology is remarkably shallow given the depth of the earnings profile provided by the 1980 DataLink. In short, that analysis neither fully exploits the virtues of the longitudinal earnings nor adequately addresses the common problem of selection bias.[5]

The 1980 DataLink accommodates what we regard as the second generation in the evaluation of VR. Although the availability of a substantive longitudinal earnings profile eases a longstanding data constraint, issues of evaluation design remain. Indeed, measuring the efficacy of public training programs is a topic of considerable current debate among economists.[6] Broadly defined, this discussion boils down to alternative views of the merits of experimental versus quasiexperimental methodologies in program evaluation.[7] The major concern is whether a quasiexperimental framework adequately controls for the aforementioned selection bias, a problem inherent in the structure of training initiatives.

The sources of selection bias are well understood. Participants in training programs typically make a series of nonrandom choices—to seek training services, to participate in a prescribed training regimen, and, ultimately, to complete the program. These sources of self-selection, combined with possible administrative screens (explicit and implicit), strongly suggest that, in the absence of random assignment, members of the treatment group will differ systematically from individuals who do not participate in a training program. Moreover, although systematic, these differences may stem from unobservable qualities (i.e., motivation) within the treatment group. On these points, the advantages of a pure experimental research design are clear. Random assignment assures that any measured earnings differences between the treatment and control groups represent an unbiased estimate of programmatic effects.

Although experimental evaluations can be conducted with relative ease for numerous public programs (Bloom et al. 1997; Boruch 1997), public sector VR is less conducive to this methodology for a variety of reasons. First,

there are legal issues involved in that the RSA regulations preclude the use of controlled experiments for persons otherwise eligible for services. There are also several practical problems that surface when implementing an experiment in VR. For example, VR is perhaps unique among public programs in that the services provided to individuals vary widely in substance and duration. A "typical" service duration can range from as little as six weeks to as long as several years. Accordingly, an experimental evaluation of VR will require a relatively long timeframe to implement. As the timeframe of an experimental evaluation expands, the common problems of recidivism, attrition, and contamination bias are exacerbated.[8]

As a practical response to these problems, a substantial literature has explored the reliability of evaluations using a quasiexperimental framework.[9] This general method relies on identifying a valid comparison group against which effects on the treatment group can be measured. This literature examines alternative correction methods for selection bias, a major concern in a quasiexperimental setting. Notable contributions to this literature are Heckman (1979); Bassi (1983, 1984); Heckman and Robb (1985); Ashenfelter and Card (1985); Dickinson, Johnson, and West (1986); and Heckman and Hotz (1989).

Nevertheless, professional confidence in quasiexperiments hit low ebb in the mid-1980s. This low water mark is defined by strong empirical rebukes to quasiexperimental results by LaLonde (1986) and Fraker and Maynard (1987). In a careful response, Heckman, Hotz, and Dabos (1987) emphasize the importance of appropriate statistical procedures that test the quality of candidate comparison groups using longitudinal earnings.[10] They argue that such tests, when passed, signficantly improve the reliability of estimates based on comparison group methodology. Recently, Bell et al. (1995) provide additional empirical support for the value of using an internally drawn comparison group from program applicants. In this article, we follow the methodological tack set by Bassi (1983, 1984) and Heckman, Hotz, and Dabos (1987). In terms of the debate, we offer our results for the VR program as further evidence of the value of quasiexperimental evaluations using an internal comparison group.[11]

RSA-SSA DATALINK

The DataLink allows for evaluation of the long-term earnings impacts of VR. Our data set contains 28,986 records for clients closed from the VR program during the fiscal year of 1980. These records reflect a 10% random sam-

TABLE 1: Selected Variables From the RSA-SSA 1980 DataLink (28,986 Cases)

	Mean	SD	*Minimum*	*Maximum*
Demographic variables				
Age at referral to VR	32.79	10.42	18.00	57.00
Gender binary (% male)	55	50	0	1
Race binary (% Caucasian)	77	42	0	1
Highest grade completed (retarded = 0)	10.79	4.25	0	20
Welfare binary (% receiving at referral)	17	38	0	1
Disability descriptors (binary variables)				
Visual (%)	5	22	0	1
Hearing/speech (%)	5	22	0	1
Musculoskeletal (%)	28	45	0	1
Internal (%)	17	38	0	1
Mental illness (%)	26	44	0	1
Substance abuse (%)	9	29	0	1
Mental retardation (%)	9	29	0	1
Severely disabled (%)	54	50	0	1
Service variables				
Total case service expenditure	$880	$1,770	0	$44,920
Duration of services (months)	20.01	14.50	1	80.00
SSA DataLink earnings				
Earnings in year before referral (1980$)	$3,395	$4,596	0	$23,770
Earnings in year after closure (1980$)	$3,549	$4,840	0	$26,923

NOTE: RSA = Rehabilitative Services Administration; SSA = Social Security Administration; VR = Vocational Rehabilitation Program.

ple of the national VR caseload.[12] In addition to longitudinal earnings profiles, the data set contains selected client-specific attributes routinely collected by the state VR agencies. These data include client demographics, disability descriptors, and service expenditure. A statistical summary of these selected variables appears in Table 1.

Table 1 provides a useful overview of the program and its constituency. The mean age of a VR client at closure is 33 years; the standard deviation indicates that roughly 68% of the VR population is between the age of 22 and 43. The gender and racial compositions—55% male and 77% white— suggests that whites and males have slightly greater representation in the VR caseload than does the general labor force. The mean level of educational attainment, excluding persons with mental retardation, is 10.79 years. The welfare binary reveals that 17% of VR clients are receiving some form of welfare benefits (e.g., Aid to Families with Dependent Children, Supplemen-

tal Security Income) when referred to the program. The seven disability descriptors listed in Table 1 give a clear impression of the variety of impairments with which VR deals.[13] People with musculoskeletal disabilities compose the largest cohort (28%), followed closely by mental illness (26%) and then internal impairments such as cardiac or respiratory ailments (17%). Across all categories, 54% of the caseload is severely disabled as defined by RSA guidelines. In terms of service dimensions, the program spends a relatively modest $880 per accepted client over a period of roughly 20 months. However, note that substantial variation exists in terms of service provision. The most expensive case in our sample approached a remarkable $45,000, whereas one individual spent more than 6 years (80 months) in the program.

These VR program data are linked via social security numbers to obtain earnings histories for the period from 1972 to 1988, measured in constant 1980 dollars. The data set is unique in that it offers substantial pre- and post-program earnings histories on a client-specific basis. The earnings data are annual observations for each of 17 calendar years. For analytical purposes, we focus on pre- and posttreatment calendar years.[14] For illustrative purposes, Table 1 presents the mean SSA-reported earnings for the first pre- and posttreatment calendar years.[15] Given that these descriptive statistics include both the treatment and comparison groups, the only inference that should be drawn from these earnings data is that VR deals with a relatively low strata of the income distribution.

The 1980 DataLink earnings profiles represent a significant enhancement over the traditional VR earnings data for purposes of evaluation. Recall, RSA records only provide a snapshot of weekly earnings at referral and closure. However, despite the obvious virtues of the DataLink earnings profiles, there are noteworthy caveats related to using SSA-reported earnings. (See the final report by BPA [1989] for a more extensive survey of these issues as they pertained to the 1975 DataLink.)

Perhaps the most obvious shortcoming of using SSA earnings data is that not all occupations are in what is called *covered* employment.[16] For our data set, we failed to find an SSA earnings match for less than 4% of the cases. The treatment of these nonmatching cases is problematic. Do they in fact have zero earnings for the entire period of the DataLink? Or, rather, do they have earnings that simply are not reported on this file?[17] Given this ambiguity and the resulting lack of an earnings outcome measure, these cases are dropped from any further analysis.

Another potential problem is that the DataLink will not capture the full earnings for any individual whose earnings exceed the maximum of the Federal Insurance Contribution Act tax base. These "truncated" earnings will bias the earnings impacts of VR services downward (upward) if they occur in

the appropriate postclosure (prereferral) outcome (baseline) period. Given that SSA dramatically increased the earnings ceilings in reforms legislated in 1977, truncated earnings will be less of a problem in postclosure periods.[18] For 1,101 clients (2.7%), truncation occurs for one of their reported calendar years of earnings. Should this occur in a baseline or outcome year, a bias is introduced in any earnings impact analysis. These cases and any case with postclosure earnings exceeding the SSA ceiling in three or more years are dropped from the analysis.

ESTIMATION PROCEDURE

Controversy in the literature regarding the value of randomized experiments versus comparison groups in the assessment of manpower programs has left some analysts uneasy. The use of comparison groups—clients who have not been selected for treatment or who have been selected but have received few, if any, services—in a quasiexperimental design has been criticized for selection bias. A main theme of this article is the explication of a methodology for estimating unbiased VR impacts using a comparison group. Toward this end, our estimation procedure follows two general tacks established by Heckman, Hotz, and Dabos (1987): (a) different sources of bias dictate different types of statistical procedures, and (b) statistical tests exist to identify the form of bias.

Our evaluation adopts an *internal* comparison group. This term connotes a subgroup that has some exposure to the program but does not receive substantial treatment. In the vernacular of VR, such individuals are classified as Status 30 closures. By definition, this status implies persons who apply and are accepted to VR but never begin a prescribed service regimen.[19] The appeal of this group is that the potential problems of selection bias are attenuated because members of this group have passed through the same self-selection and programmatic screens as the treatment group. This brand of internal comparison group, which we shall refer to as *dropouts*, received recent empirical support in an evaluation by Bell et al. (1995).[20] Having defined the comparison group as dropouts, the treatment group thus is composed of the residual portion of persons accepted for services.

The longitudinal employment data available in the DataLink provide an ideal venue for putting an internal comparison group to the test in the context of the VR program. We base our approach on Bassi (1983, 1984)[21] and view the analysis as a sequential process of testing for increasingly serious forms of selection bias. In a perfectly executed randomized experiment, the only

difference that would exist between the control and treatment groups would be the intervention. In such an instance, simple difference-in-means tests on postprogram earnings would suffice. With longitudinal data, the time path of these earning differences also could be assessed.

In an imperfect world, however, a series of tests should be performed to discern the presence of any biases that might exist between treatment and control and/or comparison groups. The most elementary test would be on observed characteristics (e.g., race, gender, age, and education). Failure of any of these tests for differences moves the analysis into the world of multiple regression. Consider an earnings function of the form:

$$Y_{it} = \alpha + X_{it}\gamma + P_{it}\beta_t + \phi_i + \tau_t + \varepsilon_{it}, \quad (1)$$

where Y_{it} is the earnings of individual i in period t, X_{it} is a vector of observable characteristics affecting earnings, and P_i is a VR program participation binary. The longitudinal nature of the data set dictates decomposition of the error component into three separate terms: an individual-specific unobservable term (ϕ_i) constant across time, a time-specific error term (τ_t) constant across individuals, and a "white-noise" error term (ε_{it}) specific to the individual i at time t. If the three components of the error term are uncorrelated with the explanatory variables, then this model can be estimated on postintervention earning levels. If, as in the GAO study, the model was estimated on a single period, then ϕ_i, τ_t, and all t subscripts would be irrelevant and ordinary least squares could be employed. However, ϕ_i and τ_t must be modeled when longitudinal postprogram data are used to estimate the time path of the earning differences. Bassi (1983) employs a random-effects modeling (REM) that assumes that ϕ_i is distributed normally across individuals with zero mean and constant variance, that τ_t is distributed normally across time periods with zero mean and constant variance, and that it satisfies the classical assumptions.

This "levels" model imposes two strong assumptions on the estimation: (a) the X vector adequately captures any differences between the treatment and comparison groups (i.e., there are no latent differences between those ending up in treatment or comparison); and (b) with the exception of a treatment effect, the earnings equation is the same between the treatment and comparison groups (i.e., the functional form and X coefficients are the same). Suppose, however, that the error terms are not distributed randomly; rather, there is some correlation between the error term and the right-hand side explanatory variables. In the circumstances of VR, such correlation likely will exist with respect to the program participation binary due to self-selection and programmatic screens throughout the VR application and acceptance process. If so, estimation on earning levels will generate earnings

impacts that are biased and inconsistent. More specifically, if selection into VR occurs due to an individual's unobservable traits, the levels estimates will be flawed.

Testing the validity of these assumptions is straightforward given preprogram earnings data. The first assumption can be tested through a Hausman test, which tests the significance of a treatment binary on preprogram earnings. The second assumption can be tested using a Chow test for a difference in any coefficient, including the intercept, in the preprogram earnings equation. The levels model apparently is inappropriate for a comparison group of dropouts, as both tests failed for earnings in the second and third years before VR.[22]

The failure of these tests indicates that unobservable differences (e.g., motivation, health) exist between the treatment group and our comparison group. If these characteristics are assumed to be constant (fixed) over time, then bias can be eliminated through a differencing of a post-VR and pre-VR treatment period.

$$Y_{it} - Y_{is} = P_{it}\,\beta_t + (X_{it} - X_{is})\gamma + (\tau_t - \tau_s) + (\varepsilon_{it} - \varepsilon_{is}), \tag{2}$$

where the *s* subscript refers to a period prior to referral for VR services and *t* denotes the postprogram period. Time-invariant, individual-specific effects are removed in this differencing. Included among these are certain *X* variables (e.g., race and education) and, critical to obtaining unbiased estimates, the unobservable traits resulting in selection bias. Of course, several things that have changed between periods *s* and *t* remain in the model, most notably, the VR intervention for the treatment group and the disability.

Equation 2 represents a fixed-effects modeling (FEM) of ϕ_i and τ_t. This FEM specification will provide unbiased and consistent VR earnings impacts under a less restrictive set of assumptions. Unlike the levels model, FEM does not require the first assumption; however, the second assumption remains. This key assumption can be validated by testing for differences in the preprogram earnings structures of the treatment versus comparison groups. Note that one of the salient virtues of the 1980 DataLink is that it provides preprogram earnings profiles of sufficient length (i.e., at least three years) to perform the necessary tests. Recall, a substantial portion of the sample from the 1975 DataLink did not have the necessary three years of preprogram earnings to conduct the appropriate tests for selection bias. BPA (1989) recognized this as a major constraint in testing for the quality of the comparison group in their evaluation of VR. Although this data limitation is removed in the 1980 DataLink, the GAO 1993 analysis did not include these tests for selection bias, tests that we regard as methodologically essential.

Specifically, checking the appropriateness of the FEM involves the Hausman and Chow tests again, this time on preprogram earnings in difference form. Theoretically, Equation 2 should ensure that the first assumption holds; passing the Hausman test verifies that this is so. The Chow test is performed on differences in preprogram earnings equation to assess the validity of the second assumption—that is, that the influence of the *X* variables (e.g., education and experience) on earnings is the same for both the treatment and comparison groups. If either test is significant, then selection bias problems remain with FEM estimation.

The choice of an appropriate preprogram year (the *s* subscript in Equation 2) is an important one. The year immediately preceding referral often reflects a transitory reduction in expected earnings ("preprogram dip"). Bassi (1984) suggests the second preprogram year as being near enough to be relevant but less likely to include substantial dip. Thus, our preprogram tests using Equation 2 set the *s* and *t* subscripts as the third and second preprogram years, respectively. Significance of either test indicates structural differences in the earnings equations between treatment and comparison that cannot be controlled for adequately by FEM. Intuitively, if there are structural differences well before treatment, then program analysts have no uncontaminated reference point against which to measure earnings gains. We also present results of these tests using the second and first preprogram years. A distinction in this period is interesting in that it indicates an unexplained difference in preprogram dip. Nevertheless, such a result would not provide strong enough evidence for a permanent difference between groups to reject the use of FEM.[23]

We have tested the viability of the FEM rendering for this comparison group vis-à-vis the treatment group using the Hausman and Chow tests previously described. All estimation in this article has been stratified by gender and seven disability classifications. The test results for these 14 stratifications are presented in Table 2. With respect to the Hausman test, the binary denoting members of the treatment group is insignificant for each cohort, with the sole exception of men with mental illness. In short, the Hausman tests on preprogram earnings changes reveal almost no unobserved differences between the treatment and comparison groups.[24] The Chow tests for the third versus second preprogram years are universally insignificant. Of minor note, one Chow test for second versus first preprogram year is significant, that for men with hearing and/or speech impairments. Together, the Hausman and Chow results suggest that the dropouts qualify statistically as an acceptable comparison group within FEM.[25]

Equipped with a valid comparison group, we now can estimate VR earnings impacts in the FEM framework of Equation 2. Fully specifying the *X*

TABLE 2: Chow and Hausman Tests for Treatment Versus Comparison Group, Stratified by Gender and DRG

	Visual	*Hearing/ Speech*	*Musculo-skeletal*	*Internal*	*Mental Illness*	*Substance Abuse*	*Mentally Retarded*
Women: Hausman test coefficients on treatment binary							
Third to second preprogram year	–338.17	–218.75	236.80	–108.94	40.67	–264.55	109.75
Second to first preprogram year	–609.50	–262.01	–195.42	–300.58	–114.69	110.09	–23.35
Men: Hausman test coefficients							
Third to second preprogram year	–606.16	–735.24	–282.40	17.35	299.62	[a] 173.16	–231.77
Second to first preprogram year	–10.54	–448.67	–112.71	–511.30	–219.08	–59.90	51.05
Women: Chow test *p* values							
Third to second preprogram year	0.842	0.243	0.098	0.170	0.694	0.776	0.220
Second to first preprogram year	0.211	0.326	0.670	0.326	0.127	0.331	0.770
Men: Chow test *p* values							
Third to second preprogram year	0.295	0.167	0.176	0.410	0.175	0.357	0.454
Second to first preprogram year	0.436	0.020 [b]	0.875	0.165	0.513	0.874	0.687

NOTE: The model is presented as Equation 2′ in the text. All statistical tests are based on a coefficient covariance matrix which is robust to heteroscedasticity (White, 1980). DRG = disability-related grouping.
a. The *t* test demonstrates a statistically significant difference at the 5% level.
b. The Wald χ^2 test demonstrates a statistically significant difference at the 5% level.

variables in earnings Equation 1 is useful to understanding the empirical results of the difference Equation 2. Our earnings model represents a slight modification of that presented in Bassi (1984). Specifically,

$$Y_{it} = \alpha + \gamma_1 Exper_{it} + \gamma_2 Exper_{it}^2 + \gamma_3 Exper_{it}^3 + \gamma_4 Educ_i + \gamma_5 Educ_i^2 + \gamma_6 White_i + \gamma_7 Welfare_i + \gamma_8 StAvgY_{it} + P_{it}\beta_t + \phi_i + \tau_t + \varepsilon_{it}, \quad (1')$$

where

$Exper_{it}$ = potential work experience of individual *i* in year *t*, calculated as age in year *t*, less education, less 6;
$Educ_i$ = years of education of individual *i*;
$White_i$ = race binary (1 = White);
$Welfare_i$ = welfare binary (1 = receiving some form of public assistance at referral);
$StAvgY_{it}$ = per capita income in the state of individual *i* in year *t*;

and the remaining variables were defined above. Equation 2 represents the result of differencing Equation 1′ using two separate time subscripts: *t*, to represent a post-VR year, and *s*, to represent the second year prior to referral. Conspicuously absent from the resulting equation will be two common "human capital" variables (race and education) and all other terms without a time subscript. Their absence is suggestive of the intuition behind the term *fixed effects* model. Despite the general importance of these attributes on earnings levels, recognize that they are *fixed* (i.e., they do not change between any two time periods) and that their influence on post-VR earnings already is accounted for in pre-VR earnings.[26]

Algebraic manipulation yields the appropriate functional form for FEM Equation 2.[27] Where k_{it} represents the difference between post-VR year *t* and pre-VR year *s*, the estimation equation is:[28]

$$Y_{it} - Y_{is} = \alpha + P_{it}\beta_t + \gamma_1(k_{it}) + \gamma_2(2\,k_{it}Exper_{is} + k_{it}^2) + \gamma_3(3\,k_{it}Exper_{is}^2 + 3k_{it}^2 Exper_{is} + k_{it}^3) + \gamma_8(StAvgY_{it} - StAvgY_{is}) + (\varepsilon_{it} - \varepsilon_{is}). \quad (2')$$

Our data set consists of a panel of eight post-VR years for each individual. Consequently, there will be eight separate treatment coefficients tracing out a time path of the dollar value of the net impact of VR treatment for each of 14 gender-disability cohorts. Although the remaining variables represent some complicated transformations of their original specification, their coefficients retain the same interpretation as in Equation 1′). That is γ_1, γ_2, and γ_3 continue to measure the cubic influence of potential work experience on the change in earnings, ceteris paribus. Similarly, γ_8 estimates the impact of the state's economic climate on earnings, ceteris paribus. The influences of the remaining,

time-invariant characteristics cannot be estimated in this framework. More important, a possible defect in this and most VR data sets is the absence of any measure of functioning. Within the FEM, a lack of any measure of change in functioning is unfortunate because the disabling condition may be changing at different rates over time for members of the treatment and comparison groups.

The FEM estimates of VR earnings impacts are presented for women and men in Tables 3 and 4, respectively. The first eight rows contain the panel estimates for each of eight postprogram years. First, verify that the treatment effects for women are generally positive and significant across all seven disability cohorts. There are two minor exceptions. For women with hearing and/or speech impairments, the earnings gains are not statistically significant in the seventh and eighth postprogram years. Also, women with substance abuse problems appear to enjoy substantial long-term earnings gains (postprogram years five through eight), but not in the early postprogram years. The estimated earnings impacts for men are slightly weaker. We find positive and significant earnings gains for six of the seven disability cohorts; the notable exception is men with hearing and/or speech impairments. But also note that the earnings gains for men with substance abuse problems and mental retardation are not sustained. Verify that for these two groups, the earnings gains last for only one and three years, respectively. In general, however, the panel estimation of the FEM reveals positive and significant earnings impacts for 13 of 14 gender and/or disability cohorts; moreover, these treatment effects generally are sustained for nine of these groups.

Certain aspects of the panel estimation procedure used here are noteworthy, indeed striking, when juxtaposed with the methods applied in the earlier DataLink studies by BPA and GAO. We submit that neither study used the longitudinal earnings data fully or appropriately. The GAO analysis of the 1980 DataLink only estimated treatment effects based on earnings in the fifth postprogram year (the choice of which was entirely arbitrary). This Spartan use of a potentially rich longitudinal data set is particularly disconcerting given that GAO (1993) publishes conclusions stating that the "long-term economic gains (for VR rehabilitants) were disappointing" (64). Of course, this conclusion contrasts sharply with our finding that VR earnings gains generally are sustained when appropriately measured against a comparison group.

The BPA (1989) analysis of the 1975 DataLink is somewhat of an improvement in that it makes full use of the longitudinal data. However, it does not use these data in a panel estimation. Rather, an individual's treatment impact is a single earnings variable calculated as the sum of the difference in earnings between the 1st year prior to referral and each postclosure year. This specification makes the implicit assumption that potentially influ-

(text continues on p. 179)

TABLE 3: Fixed-Effect Treatment Impacts for 8 Post-Program Years and VR Treatment Cost Stratified by Disability-Related Grouping and Gender: Panel A—Women

DRG	*Visual*	*Hearing/ Speech*	*Musculoskeletal*	*Internal*	*Mental Illness*	*Substance Abuse*	*Mentally Retarded*
Years after							
First	1087.89 [b]	1574.10 [b]	959.01 [b]	641.38 [b]	1,453.69 [b]	255.32	587.20 [b]
	(2.96)	(4.08)	(5.99)	(3.60)	(12.27)	(0.76)	(4.34)
Second	1,122.14 [b]	1,136.39 [b]	837.97 [b]	494.74 [b]	1,127.36 [b]	129.14	364.51 [b]
	(3.17)	(3.08)	(5.53)	(2.96)	(10.02)	(0.43)	(2.81)
Third	1,093.86 [b]	786.59 [a]	846.22 [b]	488.59 [b]	866.23 [b]	43.32	217.68 [a]
	(3.17)	(2.19)	(5.65)	(3.01)	(7.81)	(0.15)	(1.71)
Fourth	1,324.86 [b]	778.36 [a]	898.97 [b]	482.75 [b]	675.41 [b]	365.90	194.94
	(3.90)	(2.16)	(5.97)	(3.02)	(6.00)	(1.27)	(1.52)
Fifth	1,571.83 [b]	615.11 [a]	1,001.96 [b]	595.15 [b]	540.09 [b]	788.67 [b]	267.56 [a]
	(4.55)	(1.68)	(6.43)	(3.63)	(4.59)	(2.57)	(1.97)
Sixth	1,838.36 [b]	652.96 [a]	1,179.59 [b]	593.55 [b]	520.85 [b]	1,057.68 [b]	307.95 [a]
	(5.16)	1.71)	(7.05)	(3.48)	(4.08)	(3.24)	(2.14)
Seventh	1,950.67 [b]	555.95	1,251.33 [b]	761.45 [b]	454.64 [b]	1,309.85 [b]	316.99 [a]
	(5.26)	(1.38)	(6.93)	(4.16)	(3.27)	(3.70)	(2.08)
Eighth	2,184.16 [b]	417.69	1,349.60 [b]	819.42 [b]	282.92 [a]	1,327.58 [b]	429.35 [b]
	(5.60)	(0.97)	(6.91)	(4.17)	(1.89)	(3.44)	(2.58)
Gamma 1	50.74	478.30 [b]	271.46 [b]	207.84 [b]	327.39 [b]	156.91 [a]	245.69 [b]
	(1.00)	(8.42)	(8.65)	(6.35)	(13.22)	(2.38)	(9.39)
Gamma 2	–3.72 [a]	–6.75 [b]	–6.25 [b]	–1.75 [a]	–1.01	–6.48 [b]	–10.09 [b]
	(2.49)	(4.04)	(6.82)	(2.14)	(1.33)	(2.92)	(11.82)

(continued)

TABLE 3: Continued

DRG	*Visual*	*Hearing/ Speech*	*Musculoskeletal*	*Internal*	*Mental Illness*	*Substance Abuse*	*Mentally Retarded*
Gamma 3	0.00	0.01	0.02	–0.05 [b]	–0.04 [b]	0.03	0.12 [b]
	(0.17)	(0.55)	(1.22)	(4.55)	(3.45)	(0.96)	(8.25)
StAvgChg	0.18 [a]	–0.06	0.00	0.14 [b]	0.03	–0.10	0.06
	(1.94)	(0.63)	(0.08)	(2.71)	(0.80)	(1.00)	(1.18)
Constant	–176.89	–2,330.58 [b]	–1,241.26 [b]	–532.38	[a]–1,916.52	[b] 443.01	–171.89
	(0.36)	(4.37)	(5.05)	(2.10)	(10.48)	(0.90)	(0.77)
R-squared	0.03	0.05	0.03	0.04	0.02	0.02	0.04
Standard error	4,838.73	5,415.14	5,624.87	4,670.96	5,083.39	5,418.69	3,070.16
Number of observations	6,408	6,112	24,024	22,632	31,632	4,848	9,488

NOTE: The dependent variable is the change in annual earnings in one of eight post-VR years from the 2nd pre-VR year. The model is presented and discussed in the text as Equation 2′. *t* values, noted in parentheses, have been White-corrected for heteroscedasticity. VR = Vocational Rehabilitation Program; DRG = disability-related grouping; StAvgChg = change in state average income (see Equation 2′).
a. Significant at the 5% level.
b. Significant at the 1% level.

TABLE 4: Fixed-Effect Treatment Impacts for 8 Postprogram Years and VR Treatment Cost Stratified by Disability-Related Grouping and Gender: Panel B—Men

DRG	*Visual*	*Hearing/ Speech*	*Musculoskeletal*	*Internal*	*Mental Illness*	*Substance Abuse*	*Mentally Retarded*
Years after							
First	728.56 [a]	–32.55	461.04 [b]	1,286.81 [b]	1,055.02 [b]	777.18 [b]	862.25 [b]
	(1.82)	(0.06)	(3.02)	(5.65)	(7.04)	(3.09)	(6.00)
Second	721.71 [a]	–568.60	323.27 [a]	1,079.92 [b]	885.52 [b]	303.46	462.22 [b]
	(1.91)	(1.05)	(2.23)	(5.07)	(6.28)	(1.28)	(3.45)
Third	599.38	–779.74	405.69 [b]	1,082.35 [b]	647.88 [b]	–103.83	199.03
	(1.64)	(1.46)	(2.85)	(5.22)	(4.70)	(0.46)	(1.50)
Fourth	909.39 [b]	–759.68	760.12 [b]	1,383.65 [b]	755.29 [b]	–136.09	120.84
	(2.55)	(1.45)	(5.27)	(6.69)	(5.44)	(0.62)	(0.90)
Fifth	1,262.42 [b]	–561.48	1,132.07 [b]	1,578.42 [b]	869.41 [b]	–94.34	–6.50
	(3.43)	(1.06)	(7.51)	(7.59)	(6.03)	(0.42)	(0.05)
Sixth	1,535.73 [b]	–446.96	1,443.24 [b]	1,661.96 [b]	892.34 [b]	–416.02	–100.03
	(3.98)	(0.83)	(8.99)	(7.70)	(5.81)	(1.77)	(0.67)
Seventh	1,652.47 [b]	–236.81	1,655.09 [b]	1,766.07 [b]	870.43 [b]	–452.21	–168.95
	(4.12)	(0.42)	(9.67)	(7.61)	(5.28)	(1.78)	(1.05)
Eighth	1,827.99 [b]	–482.83	1,872.22 [b]	1,954.86 [b]	859.94 [b]	–641.66	–197.63
	(4.26)	(0.83)	(10.12)	(7.76)	(4.81)	(2.37)	(1.13)
Gamma 1	445.89 [b]	953.62 [b]	438.43 [b]	580.11 [b]	393.42 [b]	578.85 [b]	543.63 [v]
	(6.43)	(13.47)	(15.03)	(13.63)	(13.44)	(12.58)	(20.01)
Gamma 2	–26.03 [b]	–37.39 [b]	–23.19 [b]	–22.08 [b]	–16.68 [b]	–14.76 [b]	–19.25 [b]
	(11.94)	(16.22)	(25.33)	(17.61)	(17.05)	(10.97)	(21.85)

(continued)

TABLE 4: Continued

DRG	*Visual*	*Hearing/ Speech*	*Musculoskeletal*	*Internal*	*Mental Illness*	*Substance Abuse*	*Mentally Retarded*
Gamma 3	0.28 [b]	0.40 [b]	0.21 [b]	0.16 [b]	0.17 [b]	0.12 [b]	0.23 [b]
	(8.76)	(11.09)	(14.89)	(8.46)	(9.77)	(6.14)	(14.88)
StAvgChg	0.39 [b]	0.25 [a]	0.39 [b]	0.20 [b]	0.46 [b]	0.48 [b]	0.15 [b]
	(3.28)	(1.99)	(8.05)	(2.68)	(9.92)	(6.05)	(3.89)
Constant	–416.55	–1,306.49					
	–447.80 [a]	–2,067.14 [b]	–893.10 [b]	–1,357.80 [b]	–1,134.13 [b]		
	(0.72)	(1.82)	(2.06)	(6.34)	(4.04)	(3.73)	(5.26)
R-squared	0.06	0.13	0.06	0.11	0.04	0.04	0.09
Standard error	6,783.63	6,773.35	7,382.18	6,586.17	5,868.83	6,254.21	3,526.84
Number of observations	5,976	5,160	41,328	17,344	28,504	16,160	12,272

NOTE: The dependent variable is the change in annual earnings in 1 of 8 post-VR years from the second pre-VR year. The model is presented and discussed in the text as Equation 2′. *t* values, noted in parentheses, have been White-corrected for heteroscedasticity. VR = Vocational Rehabilitation Program; DRG = disability-related grouping; StAvgChg = change in state average income (see Equation 2′).
a. Significant at the 5% level.

ential factors (e.g., client work experience, local economic environment) do not change over time. The results of our panel estimation confirm the flaw in this assumption. Recall, our model specification includes two variables to control for time-varying factors. And indeed, note in Tables 3 and 4 that the work experience and state average income variables are statistically significant.

BENEFIT-COST RATIOS FOR THE VR PROGRAM

Table 5 presents a broad accounting of program cost. The average cost of VR-purchased services is presented for each gender-disability cohort. This figure represents the cost of services purchased from vendors that are spent on a client-specific basis. An allocation of counseling and/or placement ($876) and overhead ($219) costs must be added to the purchased service costs to obtain an average total cost estimate by gender-disability cohort.[29]

Total benefits for each gender-disability cohort are obtained by taking the present value of each period's earnings. These earnings are the estimated treatment coefficients as reported for women and men in Tables 3 and 4, respectively. Insignificant treatment coefficients are set to zero in this calculation, although this is arguably a conservative approach because insignificant coefficients still represent the best mean estimates.[30] A 4% discount rate is applied.[31] The results of the present value calculation appear in Table 6 (present value earnings). This table also reproduces the total service costs derived in Table 5. Together, these numbers constitute the benefit-cost ratios shown.

First note that the benefit-cost ratios for VR services exceed unity for 10 of 14 gender-disability classifications. Viewed broadly, one can conclude that the program is somewhat more cost-effective for women than for men. Except for mental retardation, each of the benefit-cost ratios for women exceeds unity. These ratios range from a minimum of 1.79 for women with substance abuse problems to a maximum of 4.98 for women with visual disabilities. In comparison, the services provided to men are cost-effective for only four of the seven disability cohorts. However, also verify that the four benefit-cost ratios for men that exceed unity are of substantial magnitude. The benefit-cost ratios for men with mental illness, visual, or musculoskeletal disabilities all approach three; for internal impairments the ratio is 4.6.

One noteworthy consistency in the results is the program's ineffectiveness with regard to persons with mental retardation. For both men and women, mental retardation proves to be a relatively expensive cohort to serve for which

b. Significant at the 1% level.

TABLE 5: Summary of VR Service Costs by Gender-Disability Cohort

Cost Component	*Visual*	*Hearing/ Speech*	*Musculoskeletal*	*Internal*	*Mental Illness*	*Substance Abuse*	*Mentally Retarded*
Panel A: Women							
Purchased services ($)	914	873	990	915	842	829	1,294
Counseling/placement ($)	876	876	876	876	876	876	876
Overhead ($)	219	219	219	219	219	219	219
Total costs ($)	2,009	1,968	2,085	2,010	1,937	1,924	2,389
Panel B: Men							
Purchased services ($)	1,417	862	1,160	1,014	875	558	1,257
Counseling/placement ($)	876	876	876	876	876	876	876
Overhead ($)	219	219	219	219	219	219	219
Total costs ($)	2,512	1,957	2,255	2,109	1,970	1,653	2,352

NOTE: VR = Vocational Rehabilitation Program.

TABLE 6: Benefits, Costs, and Benefit-Cost Ratios by Gender-Disability Cohort

	Visual	*Hearing/ Speech*	*Musculoskeletal*	*Internal*	*Mental Illness*	*Substance Abuse*	*Mentally Retarded*
Panel A: Women							
Present value earnings ($)	10,012	4,950	6,911	4,057	5,195	3,451	2,280
Total service cost ($)	2,009	1,968	2,085	2,010	1,937	1,924	2,389
Number of cases	758	732	2,694	2,679	3,538	545	1,085
Benefit-cost ratio	4.98	2.52	3.31	2.02	2.68	1.79	0.95
Panel B: Men							
Present value earnings ($)	6,988	0	6,449	9,762	5,764	747	1,256
Total service cost ($)	2,512	1,957	2,255	2,109	1,970	1,653	2,352
Number of cases	682	615	4,467	1,900	3,144	1,864	1,403
Benefit-cost ratio	2.78	0.00	2.86	4.63	2.93	0.45	0.53

the earnings impacts are relatively modest.[32] In contrast, a seeming inconsistency surfaces with regard to persons with hearing and/or speech impairments. Although the service costs are virtually identical for men and women, the benefit-cost ratios differ markedly due to differences by gender in the estimated earnings impacts. Indeed, note that the present value of earnings for men is 0, as compared to almost $5,000 for women. Recognize, however, that this result follows partly from our procedure for calculating the present value of earnings.[33]

For methodological reasons that are fairly well-known, we have adopted gender-disability stratifications throughout this analysis. Moreover, there are perhaps policy implications to be wrought from a stratified framework. Still, the broader question is inevitably asked—How does VR perform overall? An aggregate benefit-cost ratio is constructed easily from the information in Table 6. Using the number of cases in each cohort as the appropriate weights, we obtain an aggregate ratio of 2.61 for 12,031 women and 2.43 for 14,075 men. In sum, our findings indicated that the VR program returns roughly $2.50 for each dollar spent. Recognize, of course, that this aggregate number does not accurately represent the program's performance with respect to specific disabilities.

CONCLUSIONS

The benefit-cost ratios presented here reflect the most authoritative estimates of VR earnings impacts to date. This claim is based largely on the availability of a unique data set containing rich longitudinal earnings profiles for VR applicants. These earnings data accommodate a series of statistical procedures that allow us to identify and control for the presence of selection bias when estimating the treatment impacts of VR services. The earnings estimates indicate that the VR program is cost-effective in general, although not universally so across specific disabilities. Notably, these findings run counter to the GAO (1993) conclusion, which found little evidence of long-term treatment effects from VR. Based on an eight-year postprogram profile, our results indicate that the long-term earnings gains can be substantial.

In broader perspective, we add these findings to what will likely be a continuing methodological debate. Can evaluators confidently rely on treatment estimates obtained in a quasiexperimental research setting? Although each training program will surely have its institutional nuances, the foregoing analysis suggests that, at least in the case of VR, it is possible to use program dropouts as an appropriate comparison group to obtain unbiased estimates of

the earnings gains. Although controlled experiments are conceptually superior, their cost and logistical constraints are well-known. Thus, as a practical matter, we believe the potential of quasiexperimental methods should be reconsidered or, rather, honed. Our reading of the literature indicates that the most common failing in the quasiexperimental genre results from using external comparison groups. In contrast, we have shown that using the appropriate statistical tests, one can identify a viable internal comparison group with which to conduct an evaluation that addresses the key concern of selection bias. Although it is unlikely that every training program will offer a promising internal comparison group, the appropriateness of a quasiexperimental research design should be assessed on a case by case basis. Here, we have provided a template for how to proceed using a panel data set for a training program serving persons with disabilities.

NOTES

1. Clients receive varying combinations of counseling, physical or mental restoration, job training, formal education, transportation, income support, and job placement. Services are provided both in-house by state Vocational Rehabilitation (VR) Program counselors (e.g., guidance and placement) and purchased from private sector providers (e.g., medical procedures, education). The program is financed under a federal/state matching formula of roughly 80%/20%, respectively.

2. First, earnings reported at acceptance are unlikely to reflect the true preprogram earnings path of a client due to "preprogram dip." If so, these earnings do not represent how the client would fare in the absence of treatment and therefore are a poor benchmark for assessing net training effects. Furthermore, VR may represent an extreme case of preprogram dip. It is common for VR clients to report zero earnings in the week prior to application to the program (85% in 1980). A second problem exists in the earnings reported for rehabilitated clients. This datum only reflects earnings after 60 days of employment. Earnings for such a short employment spell may misrepresent the true postprogram earnings path. Indeed, given the high recidivism rate in VR, it would seem more appropriate to assume some decay in the postprogram earnings streams. A third data problem follows from the fact that a significant fraction of VR clients do not complete the program successfully. Thus, no closure earnings are available for this cohort, although some of these clients ultimately get jobs. These potential treatment effects can only be captured by a longer span of postprogram earnings. For a fuller discussion of these issues, see Dean and Dolan 1991 (574-76).

3. Technically, the Social Security Administration (SSA) initiated the concept of an earnings cross-match using 1971 closed VR cases. The data from this prototype effort was not released by SSA and thus was never subjected to external analysis. For an overview of this precursor, see Greenblum 1975.

4. Indeed, the Berkeley Planning Associates (BPA) (1989, A1-A2) clearly acknowledge that the revised service mandate of the Rehabilitation Act of 1974 only had been implemented partially even as late as 1976.

5. A fuller account of the specific shortcomings of the GAO (1993) analysis is provided in conjunction with the discussion of the Estimation Procedure section of this article.

6. For a concise survey and interpretation of the history of training program evaluation efforts, see chapter 1 of Bell et al. (1995).

7. The tenor of this debate is reflected faithfully in the recent companion pieces by Burtless (1995) and Heckman and Smith (1995).

8. For an extensive discussion of these issues as they arise in an experimental evaluation of VR, see Dean, Dolan, and Schmidt (1998). This paper reports on "Project NetWork" (1992-1998), a recent SSA demonstration that examined the efficacy of rehabilitative services in an experimental setting specifically for disability insurance (DI) beneficiaries.

9. There are numerous methodologies for controlling for the selection bias common to most manpower training evaluations. Hotz (1992), commenting on a GAO (1993) study of VR that relied on a quasiexperimental design, suggested three alternative estimation strategies. First mentioned is the "control function" estimation approach using the Heckman procedure (Heckman and Robb 1985) or the propensity score derived by Rosenbaum and Rubin (1983). Second is the class of longitudinal data estimators, including the fixed effects estimator used in this analysis, as well as a "random growth" model. The third approach involves "statistical matching" procedures such as the Mahalanobis "nearest neighbor" technique (see Dickinson, Johnson, and West 1986), or the matched sampling of Rubin (1979). Two more recent methodological advances are available that hold potential in evaluating VR: (a) the use of instrumental variables to identify local average treatment effects (Imbens and Angrist 1994) and (b) the bounded estimation procedures developed by Manski (1990).

10. Specifically, their response makes four points: (a) different sources of bias dictate different types of statistical procedures, (b) statistical tests exist to identify the form of bias, (c) appropriate statistical procedures provide similar results using quasiexperimental design as random assignment on the Fraker and Maynard (1987) data set, and (d) the poor performance of quasiexperimental designs in these studies follows from the use of inappropriate statistical procedures.

11. See Grossman and Tierney (1993) for a more recent rebuke of quasiexperimental methods to correct adequately for selection bias in training program evaluation. We note, however, that the findings of this study do not extend to the focus of our article (i.e., using an internal comparison group drawn from program enrollees).

12. The initial 10% sample actually yielded 41,775 cases, although 12,789 cases do not qualify for analysis due to three criteria: (a) 6,330 persons were too young (i.e., clients too young at referral to allow for two years of preprogram work experience), and 2,937 were too old (i.e., clients too old at closure to have the full 8 years of postprogram earnings prior to retirement); (b) 2,480 cases were lost due to missing SSA earnings or persons whose earnings exceeded the Federal Insurance Contribution Act ceiling; and (c) 1,042 cases were dropped due to missing demographic data needed for the estimated earnings equation.

13. These categories reflect substantial aggregation on our part. There are actually hundreds of medical classifications applied in diagnosis by VR.

14. The use of calendar-year earnings from SSA raises a nettlesome alignment problem—defining pre- and postprogram earnings around the in-program period. Lacking quarterly data that would ameliorate the problem, it is useful to allot these 17 years across a sequence of five time periods: (a) earnings in the calendar years prior to the year of referral, (b) earnings in the calendar year of referral, (c) in-program calendar-year(s) earnings, (d) earnings in the calendar year of closure, and (e) earnings in the calendar years following the year of closure. Recognize that items b and d are of suspect value—the accuracy of these data as pre- or postprogram earnings is likely tainted by the fact that some portion of the calendar-year earnings could really be in-program earnings. By definition, this also would be the case for Interval 3. The earnings

periods defined as items b through d above must be clearly identified so that they can be distinguished from purely pre- and postprogram earnings periods (i.e., items a and e, respectively).

15. Recognize, of course, that in the actual estimation process, we have multiple years in the pre- and postearnings profiles, although the length of the profile will vary by individual. This variation will be slight for the postprogram interval, seven versus eight years, depending on whether a person closes in the second or first half of fiscal year 1980. There is obviously greater variation in the length of the preprogram earning profile due to the fact that VR does not provide a fixed duration treatment regimen. Although almost 70% of the sample has preprogram earnings profiles of at least five years, a small percentage of our sample (613 cases) actually was receiving services in 1972. Obviously, the DataLink will not yield any pre-VR earnings history in these relatively few cases.

16. As Browning and Browning (1983) note, "Jobs specifically excluded from coverage include federal government jobs that provide their own pensions, such as the civil service system. State and local governments can select on a voluntary basis to have their employees covered by social security . . . In all, about one job in ten is not covered (in 1981)" (210).

17. The obvious example is earnings from the "underground" economy. Indeed, unreported earnings of this nature can be especially important to low-income individuals, many of whom may have work disabilities.

18. Indeed, the problem is particularly acute in 1972, the first year of reported calendar-year earnings for the 1980 DataLink. In this period, 1,836 cases (4.4% of all cases) have earnings exceeding the existing $9,000 earnings ceiling. For 1973, when the ceiling is $10,800, there are 1,343 (3.2%) truncated cases. As the earnings ceiling rises to $45,000 in 1988, the truncated earnings problem became steadily less prevalent. Indeed, in this last year, truncation is an issue for only 297 cases (0.7%).

19. Officially, the Rehabilitative Services Administration (RSA) identifies nine reasons for dropping out. Of the attrition in our sample (2,007 persons), 70% is accounted for by persons who either "refused services" (27%), "could not be located" after acceptance (23%), or "failed to cooperate" (20%). The less meaningful category, "all other reasons," accounts for another 16% of attrition. Although far less prevalent, clients also drop out due to institutionalization or death.

20. Technically, Bell et al. (1995) describe this cohort as "no shows." Unlike true no shows, dropouts in VR may receive a diagnostic evaluation and limited counseling services. They do not, however, receive any of the remedial, training, educational, or job placement services of VR. Furthermore, it is important to note that unlike selection bias, this contamination bias works against finding a positive treatment effect. For a fuller discussion of the conceptual basis for using dropouts in a VR setting, see Dean and Dolan (1991, 571-73).

21. Bassi (1983, 1984) restricts her focus to a single class of estimators that is straightforward and adequate in the present instance. Heckman and Robb (1985) and Heckman and Hotz (1989) provide more exhaustive treatments of tests and corrections for various forms of bias.

22. The results of these tests are available from the authors upon request. Note that the first pre-VR year is not used in these tests because it might show evidence of preprogram dip.

23. Of course, the transitory versus permanent components of preprogram dip for a cohort with disabilities depends on the nature and severity of the impairment. It is assumed that the stratification into seven disability categories addresses the differences that may arise across disability type.

24. Of course, as Bassi (1983, 543) notes, preprogram earnings differences may be large but insignificant due to relatively large standard errors for these slope coefficients. In our case, however, 8 of the 13 insignificant coefficients were negative, indicating that the comparison group had the advantage in preprogram earnings growth; 2 more were positive but less than $50; another was around $100; and the remaining 2 were around $200.

25. It is worth emphasizing that our conclusions regarding the statistical quality of the dropout comparison group, although consistent with Bell et al. (1995), refute the findings of BPA's 1975 DataLink analysis. We believe that there are at least two factors that account for the conflicting conclusion. First, the additional years of preprogram earnings on the 1980 DataLink allowed us to check for differences in the earnings structure beyond the period most likely to encompass preprogram dip (i.e., Year 2 versus Year 1). Examining earnings changes between Year 3 and Year 2 reduced the magnitude of selection bias. Second, the BPA finding suffered from an aggregation problem. When the analysis is stratified by gender-disability classifications, as is typical in most training evaluations, the bias is eliminated.

26. In at least two respects, the use of the fixed effects framework imposes a relatively strong assumption regarding the unchanging nature of explanatory variables. First, household characteristics that influence labor force participation can change (e.g., divorce, birth of a child). Second, the model assumes stability in terms of individuals' physical, mental, and emotional functioning. This second assumption is relatively more serious in our view because persons with disabilities may be more prone to experience declining health. Recognize that for this to be a problem, these phenomena have to be systematically different between the treatment and comparison groups. To some extent, the problem is diminished by stratifying by disability.

27. The derivation of the experience terms in Equation 2′ merits additional explanation. Experience in the postprogram period (t) is equal to experience in the preprogram earnings period (s) plus the number of years elapsed (k_{it}) (i.e., $\text{Exper}_{it} = \text{Exper}_{is} + k_{it}$). Equation 1′ for the postprogram year t is then:

$$Y_{it} = \alpha + \gamma_1(Exper_{is} + k_{it}) + \gamma_2 (Exper_{is} + k_{it})^2 + \gamma_3 (Exper_{is} + k_{it})^3 + \ \ldots \ + \varepsilon_{it}.$$

Expanding this expression and subtracting Equation 1′ for preprogram year s yields Equation 2′:

$$[Y_{it} - Y_{is}] = \gamma_1 k_{it} + \gamma_2(2\, k_{it} Exper_{is} + k_{it})^2 + \gamma_3(3\, k_{it} Exper_{is}^{\ 2} + 3\, k_{it}^{\ 2} Exper_{is} + k_{it}^{\ 3}) + \ \ldots + (\varepsilon_{it} - \varepsilon_{is}).$$

28. Three points might be noted concerning this formulation. First, the calendar year corresponding to the s subscript differs across clients because referral years differ across clients. Less obviously, the same holds true for the t subscript because these cases closed in fiscal year 1980, but SSA earnings are reported for calendar years. Thus, $t = 1$ represents 1980 for 1979 closures but 1981 for 1980 closures. Clearly then, k_{it} is both individual and time specific. Second, although α cancels out when differencing, we are not inclined to force Equation 2′ through the origin. Consequently, we include an intercept in our estimation equation. Third, although the term $(\tau_t - \tau_s)$ remains in Equation 2, we drop it because we believe that the change in per capita state income better captures period influences. In other words, we assume that without treatment, there is no time trend in earning changes after controlling for changes in individual experience and the economic climate.

29. The estimates for counseling and/or placement and administrative overhead are obtained from the Rehabilitation Services Administration (1982a, 1982b). Because fiscal year 1980 figures are unavailable, an average of 1979 and 1981 figures is used. Total costs for counseling and/or placement and administration for general agencies for both periods ($351,038,486 and $401,715,501, along with $83,499,724 and $105,017,690, respectively) are averaged. The resulting figures then are divided by the number of rehabilitated (277,136) and not rehabilitated (152,672) cases closed during the year to arrive at the average charges on a per client basis. This assumes a uniform distribution of both administrative and counseling time across clients with different disabilities and genders. These figures also include charges for counseling and

administrative time spent on clients not accepted for services and those still on the active caseloads (i.e., not closed) during 1980. A total of 1,095,139 cases were served during this period. It is impossible to distinguish counseling and administrative time spent on active cases, program-eligible closed cases and ineligible closed cases. Accordingly, the reported figures will overestimate the charges in these two areas.

30. This assumption has a major impact on the benefit stream for 3 of 14 cohorts—men with hearing and/or speech disabilities, substance abuse, and mental retardation.

31. This discount rate is the one suggested by the U.S. Department of Education for the BPA (1989) analysis of the 1975 DataLink.

32. It is important to interpret this result in relation to the initial earnings situation for persons with mental retardation. Perhaps unsurprisingly, this cohort has very low preprogram earnings; thus, even the modest postprogram gains reported here can be viewed as a notable relative improvement.

33. Recall that in the net present value calculation, insignificant earnings coefficients in any year are treated as 0. In other words, the estimates earnings impacts for men with hearing and/or speech impairment are statistically insignificant for each of the 8 years of the panel data. We suspect that this statistical outcome is very likely due to the fact that this male cohort turned out to have a relatively small comparison group.

REFERENCES

Ashenfelter, O., and D. Card. 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics* 67:648-60.

Bassi, L. 1983. The effect of CETA on the postprogram earnings of participants. *The Journal of Human Resources* 18:539-56.

Bassi, L. 1984. Estimating the effect of training with non random selection. *The Review of Economics and Statistics* 66:36-43.

Bell, H., L. Orr, J. Blomquist, and G. Cain. 1995. *Program applicants as a comparison group in evaluating training programs*. Kalamazoo, MI: Upjohn Institute for Employment Research.

Bellante, D. 1972. A multivariate analysis of a vocational rehabilitation program. *The Journal of Human Resources* 7:226-41.

Berkeley Planning Associates. 1989. *The economic benefits of the Vocational Rehabilitation Program*, contract no. 300-86-0115. Berkeley, CA: U.S. Department of Education.

Bloom, H., L. Orr, S. Bell, G. Cave, F. Doolittle, W. Lin, and J. Bos. 1997. The benefits and costs of JTPA Title II-A Programs. *Journal of Human Resources* 32:549-76.

Boruch, R. 1997. *Randomized experiments for planning and evaluation*. Thousand Oaks, CA: Sage.

Browning, E., and J. Browning. 1983. *Public finance and the price system*. New York: Macmillan.

Burtless, G. 1995. The case for randomized field trials in economic and policy research. *Journal of Economic Perspectives* 9:63-84.

Conley, R. 1969. A benefit-cost analysis of the vocational rehabilitation program. *Journal of Human Resources* 4:226-52.

Dean, D., and R. Dolan. 1991. Assessing the role of vocational rehabilitation in disability policy. *Journal of Policy Analysis and Management* 10:568-87.

Dean, D., R. Dolan, and R. Schmidt. 1998. Evaluating public-sector vocational rehabilitation in an experimental framework. Manuscript in preparation.
Dickinson, K., T. Johnson, and R. West. 1986. An analysis of the impact of CETA programs on participants' earnings. *The Journal of Human Resources* 20:64-91.
Fraker, T., and R. Maynard. 1987. Evaluating comparison group designs with employment-related programs. *The Journal of Human Resources* 22:194-227.
Greenblum, J. 1975. Evaluating vocational rehabilitation programs for the disabled: National long-term follow-up study. *Social Security Bulletin* 38:3-12.
Grossman, J., and J. Tierney. 1993. The fallibility of comparisons groups. *Evaluation Review* 17:556-71.
Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47:153-61.
Heckman, J., and V. J. Hotz. 1989. Choosing among alternative non-experimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association* 84:862-74.
Heckman, J., V. J. Hotz, and M. Dabos. 1987. Do we need experimental data to evaluate the impact of manpower training on earnings? *Evaluation Review* 11:395-427.
Heckman, J., and R. Robb. 1985. Alternative methods of evaluating the impact of interventions. In *Longitudinal analysis of labor market data*, edited by J. Heckman and B. Singer. Cambridge, UK: Cambridge University Press.
Heckman, J., and J. Smith. 1995. Assessing the case for social experiments. *Journal of Economic Perspectives* 9:85-110.
Hotz, V. J. 1992. Comments on the analysis plans for the GAO study of the vocational rehabilitation program. Unpublished memo to the U.S. General Accounting Office Program Evaluation and Methodology Division.
Imbens, G., and J. Angrist. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62:467-75.
LaLonde, R. 1986. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76:604-20.
Lewis, D., D. Johnson, T. Chen, and R. Erickson. 1992. The use and reporting of benefit-cost analyses by state vocational rehabilitation agencies. *Evaluation Review* 16:269-87.
Manski, C. 1990. Nonparametric bounds on treatment effects. *American Economic Association Papers & Proceedings* 80:320-23.
Nowak, L. 1983. A cost-effectiveness evaluation of the federal/state vocational rehabilitation program—using a comparison group. *The American Economist* 27:23-29.
Rehabilitation Services Administration. 1982a. Caseload statistics, State Vocational Rehabilitation Agencies, fiscal year 1980, information memorandum RSA-IM-82-11. Washington, DC: U.S. Dept. of Education.
Rehabilitation Services Administration. 1982b. State Vocational Rehabilitation Agency program data, fiscal year 1981, information memorandum. Washington, DC: U.S. Dept. of Education.
Rosenbaum, P., and D. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41-55.
Rubin, D. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* 74:318-28.
U.S. General Accounting Office. 1993. Vocational rehabilitation: Evidence of federal program effectiveness is mixed. Report to the Subcommittee on Select Education, House of Representatives. Washington, DC: Author.
White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817-38.

Worrall, J. 1978. A benefit-cost analysis of the vocational rehabilitation program. *Journal of Human Resources* 13:285-98.

David H. Dean is an associate professor in the Economics Department at the University of Richmond where he also is a codirector of the Bureau of Disability Economics Research.

Robert C. Dolan is a professor in the Economics Department at the University of Richmond where he also is a codirector of the Bureau of Disability Economics Research.

Robert M. Schmidt is an associate professor in the Economics Department at the University of Richmond where he also is a codirector of the Bureau of Disability Economics Research.