American Lessons on Designing Reliable Impact Evaluations, from Studies of WIA and Its Predecessor Programs

Larry L. Orr, Stephen H. Bell, and Jacob A. Klerman[1]

October 30, 2009

This paper reviews the U.S. experience in evaluation of job training programs over the past 40 years, examines why it is so difficult to reliably estimate the impacts of training programs with nonexperimental methods, and discusses ways to make experimental evaluations more feasible and cost-effective. We focus exclusively on *impact* evaluations, studies which seek to measure the contribution of a training program to improving worker outcomes *above and beyond what the same workers would have achieved without the training* (known as the "counterfactual"). Other types of workforce-focused evaluations—such as process studies of program implementation, or participation analyses that examine program targeting—while important, are not considered here.

A major distinction in our discussion is between "experimental" impact evaluation methods and "nonexperimental" impact evaluation methods. The experimental method randomly assigns eligible applicants for a training program to two groups, a "treatment group" that is allowed to enter the program and a "control group" that is not allowed to enter the program. Only by chance will subsequent outcomes of the two samples differ, unless the training improves treatment group outcomes. The difference in average outcomes between the treatment and control groups, tested for statistical significance (to rule out chance as the explanation of the observed difference) is the measure of program impact.

Nonexperimental impact evaluation methods also measure outcomes for a sample of training program participants, but—not having done random assignment—have no similar control group to compare to; instead, preprogram earnings of participants or earnings of some set of non-participants (called a "comparison group") must be used as the counterfactual. The challenge is how to find a valid comparison group and then how to control for any remaining treatment group/comparison group background differences. The obvious approach is to select the comparison group from those who were eligible for the program, but chose not to enroll. However, given that they chose not to enroll, they must be different from those who chose to enroll.

The alternative is to chose a comparison group from among those not eligible to enroll (e.g., from a different time period or a different geographic area, or not meeting one of the enrollment conditions). Again, whatever the condition is that makes the comparison group ineligible to enroll will also make them different from those who did enroll. Of course, a non-experimental evaluation can and would control for observed

[1] Bell and Klerman are at Abt Associates Inc. Orr is an independent consultant.

differences between the treatment group and the comparison group, but nothing guarantees either that the only differences are in observed characteristics, or that the nature of the correction for those observed differences is correct. Thus, as we argue in detail below, those commissioning non-experimental evaluations will always be left with the nagging concern that the non-experimental methods chosen were not successful in producing accurate impact estimates.

A Brief Overview of U.S. Evaluations of Training Program Impacts

Serious evaluation of government employment and training programs began in the U.S. in the 1960s, with non-experimental impact analyses of programs funded by the Manpower Development and Training Act (MDTA). To estimate training impacts, analysts needed estimates of earnings with training and estimates of the counterfactual—what earnings would have been, for the same individuals, without training. Earnings with training were observed. The challenge was to estimate earnings without training. Some early MDTA studies took pre-program earnings for trainees as the benchmark. The impact of treatment could then be estimated as the change in earnings from before training to after training.[2] This approach clearly gave estimates of program impacts that were too large and the reason was clear. People generally enter job training programs when they are at a low point in their labor market trajectory—e.g., when they are unemployed. As a result their earnings tended to rise, even quite substantially, even without training's assistance. The pre-post change measure credited this natural rebound to the employment and training intervention, giving the appearance of a program impact where there was none.

As it became clear that pre-program earnings were not a good counterfactual, MDTA analysts turned to comparison group strategies, in which training participants' counterfactual earnings were estimated using a sample of similar workers in a comparison group who did not enroll in training. As noted above, the measure of program impact was the difference in average outcomes between participant and comparison group members, usually adjusted for measured differences in background characteristics between the two populations.

In the 1970s, the U. S. Department of Labor (DOL) sponsored a number of comparison group-based evaluations to measure the impacts of their training programs and demonstrations from that decade. Launched with high expectations, these efforts ended in disappointment. In many cases, the results were unclear or inconsistent; in others, the results were overshadowed by controversy, often acrimonious, about the ability of the methods used to produce accurate results. The first of these efforts was a series of evaluations of DOL's major job training program for disadvantaged workers, the Comprehensive Employment and Training Act (CETA) program. The second was a set of over 400 demonstrations of employment and training programs for youth under the Youth Employment Demonstration Program Act (YEDPA). Most of these demonstrations involved nonexperimental evaluations.

[2] See Bell et al. (1995) for an in-depth history of U.S. training program evaluations and their impact estimation methodologies, from the MDTA era through the mid 1990s.

More than half a dozen CETA evaluations produced widely divergent estimates of the impact of the program on participants' earnings, even though all the studies were based on essentially the same data (Barnow 1987).. These differences in results were apparently due to differences in the assumptions underlying nonexperimental methods. And since those assumptions could not be tested or verified with data, there was no way to know which estimates were most reliable.[3] Moreover, when researchers applied the same set of non-experimental methods to data drawn from a social experiment, where the experimental estimate provided an unbiased benchmark, the results were again widely dispersed and generally did not replicate the experimental findings (LaLonde, 1986; Maynard and Fraker, 1987; Heckman and Smith, 1995). This experience led an expert panel convened to advise DOL on the evaluation of the Job Training Partnership Act (JTPA; the program that succeeded CETA) to recommend strongly that JTPA be evaluated with experimental methods (Stromsdorfer et al., 1985)..

Similarly, when evaluations of the youth employment demonstrations (YEDPA) of the late 1970s were reviewed by a National Academy of Sciences committee, the committee concluded that:

> "Despite the magnitude of the resources ostensibly devoted to the objectives of research and demonstration, there is little reliable information on the effectiveness of the programs in solving youth employment problems...It is evident that if random assignment had been consistently used, much more could have been learned." (Betsey, Hollister, and Pappageorgiou, 1985)

These recommendations led to the National JTPA Study, in which over 20,000 job training applicants in sixteen local programs across the country, including both adults and youths, were randomly assigned either to go into the program or into a control group that was excluded from the program. The study had two major conclusions (Orr et al., 1995): (1) that the adult program components were cost-effective, and (2) that the youth programs had no discernable positive effects, and for some youths (those with arrest records) might have had a negative effect. When the study findings were released, Congress cut the youth program by 90 percent but maintained funding for the adult program.

Since the JTPA study, DOL has successfully used randomized designs for many of its other program evaluations and demonstration projects. For example, Job Corps, a residential training program for youth, was evaluated with an innovative design in which a national probability sample of sites was drawn and a small number of program applicants were randomly assigned to control status in each site (Schochet et al., 2008). DOL also followed up on the negative findings for youth in the JTPA evaluation by testing two approaches that had shown promise in previous evaluations—that of the Center for Employment Training (Miller, et al., 2005) and the Quantum Opportunities Program (Schirm et al., 2006)—in an attempt to find more effective ways to serve

[3] See Heckman and Hotz (1998) for a (much later) attempt to address this lack of ability to test implicit assumptions.

disadvantaged youth. Because the studies had randomized designs, there was not disputing the findings when they showed both programs to be ineffective (i.e., to have no impacts).

Reliance on experimental designs has continued at DOL up to the present. For example, a recent randomized study of Project GATE (Growing America Through Entrepreneurship), measured the impact of providing microenterprise start-up services on participant employment and earnings (Benus et al., 2008). DOL's evaluation of Individual Training Accounts randomized consumers between three different voucher/counseling approaches (McConnell et al., 2006) to get unbiased measures of the *differential* effectiveness of the three strategies. A similar approach is being taken in the Workforce Investment Act (WIA) impact evaluation, which will use random assignment to determine which consumers participate in which WIA program components (Mathematica Policy Research, Inc., 2009). Another randomized study just underway at DOL, the Young Parents Demonstration, will have a true control group that receives no special services.[4]

The Current Consensus

Frustration with the failure of non-experimental methods to yield unequivocal estimates of program effects in cases such as CETA and YEDPA led to a consensus among evaluation specialists within the U.S. federal government that, where feasible, random assignment is the method of choice for evaluating public programs. Bell (2003) has argued that random assignment is almost always possible in federal workforce evaluations, even for mainline labor market interventions like local economic development assistance and Unemployment Insurance benefits. This consensus among the technical experts has in turn led policymakers to accept experimental designs not only as scientifically accurate, but also as a way to avoid the methodological debates that often accompany the presentation of non-experimental results, detracting from their credibility and deflecting the policy discussion from substance to method.

Experimental methods are also appealing to policymakers for their simplicity. In contrast to the complicated statistical complexity of many non-experimental methods, the experimental method is relatively simple and intuitively understandable. Even non-technical policymakers can appreciate the logic of a contrast between two groups, one exposed to the program and the other not exposed to the program, but differing otherwise only by chance. This makes experimental studies more accessible and credible to lay persons in the policy process.

For these reasons, not only has the number of social experiments funded and conducted in the United States increased enormously over the last three decades,[5] but on a number of occasions, random assignment evaluations have been mandated by Congress.

[4] Personal correspondence with Young Parents Demonstration study leader Karin Martinson, October 28, 2009.

[5] Greenberg and Shroder (2004) summarize over 200 completed social experiments; many more have been finished (and others initiated) in the five years since.

For example, the landmark welfare reform act passed in 1996 directed the Secretary of Health and Human Services to evaluate the programs funded under the act and, "to the maximum extent feasible, use random assignment as an evaluation methodology."[6] Similarly, the Education Sciences Reform Act (ESRA) of 2002, which established the Institute of Education Sciences (IES), defined "scientifically valid education evaluation" as evaluation that "employs experimental designs using random assignment, when feasible, and other research methodologies that allow for the strongest possible causal inferences when random assignment is not feasible…"[7] Congress has mandated random assignment evaluations of a number of specific programs in health, labor, housing, welfare, and education.

Challenges to the Consensus

One might, of course, ask whether nonexperimental evaluation methods have become more reliable in the 25 years since the publication of the National Academy of Sciences panel conclusions quoted above. There has, in fact, been a great deal of work on nonexperimental estimators during that period and there is some evidence that they have gotten more reliable. Using the same dataset that LaLonde (1986) employed in his classic analysis of nonexperimental evaluations of CETA, Dehejia and Wahba (1999) showed that the propensity score matching approach proposed by Rosenbaum and Ruben (1983) could replicate the experimental estimates with remarkable fidelity. And a recent meta-analysis by Greenberg et al. (2006) showed that, on average, twenty nonexperimental impact analyses of six job training programs yielded estimates that were quite similar to those obtained by nine randomized experiments.

On closer examination, however, these studies are less encouraging than they might seem on first examination. A re-analysis of the Dehejia-Wahba study by Smith and Todd (2005) found that the results were strongly sensitive to sample selection and specification of matching variables. In particular, although it was possible to find a non-experimental approach that yielded estimates similar to the (known) experimental results, equally plausible approaches—in fact only slight variations in the non-experimental methods—yielded results different from, and sometimes very different from, the experimental results. This is similar to the range of estimates from apparently reasonable non-experimental methods which was noted by the National Academy of Sciences and others a quarter century ago.

In Greenberg et al.'s meta-analysis , the nonexperimental studies reviewed evaluated different programs than the experimental studies examined.[8] The finding of no difference, on average, between experimental estimates for one set of programs and nonexperimental estimates for another set of programs does not address the key question—whether non-experimental methods estimate the true impacts *for a given program*. Furthermore, Greenberg et al.'s study seems to confound period with method:

[6]Public Law 104-193, Sec. 413(b)(2).
[7] Public Law 107-279, Sec. 102 (19)(D).
[8] In the one case where both a nonexperimental and an experimental evaluation of the same program were included, Job Corps, the latter was conducted 18 years after the former.

all but one of the non-experimental estimates are from before 1988 and all but two of the experimental estimates are from after 1988.

Tests of Nonexperimental Estimates Against Experimental Benchmarks

A number of studies do compare experimental and nonexperimental impact estimates of job training impact for the same program. Those studies consistently find that nonexperimental estimates fail to replicate the experimental findings when taken one program at a time. Pirog et al. (2009), for example, examined 18 articles that explicitly compared propensity score matching (PSM), difference in differences (DD), or regression discontinuity design (RDD) estimates with estimates for the same program drawn from randomized experiments. Their summary assessment was:

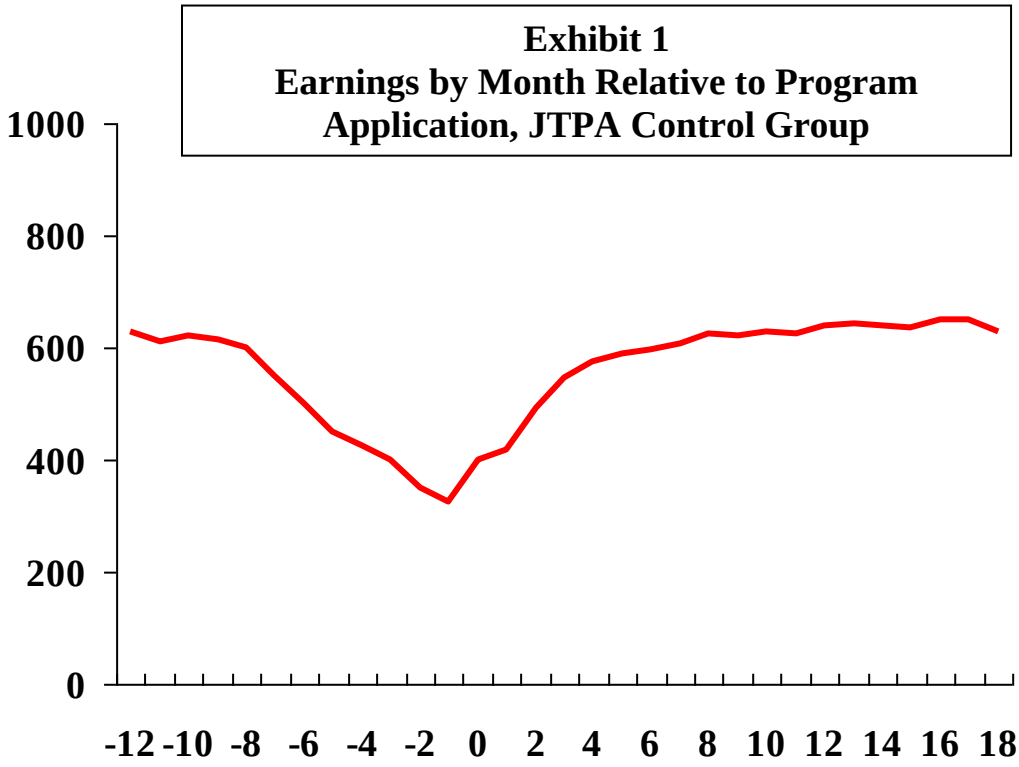
> "…all [econometric corrections] are sensitive to the sampling frame and analytic model used…these corrections do not uniformly and consistently reproduce the experimental results; therefore, they cannot be relied upon to provide a satisfactory substitute for RA experiments." (p. 171)

Of particular relevance here is one of these studies, Glazerman et al. (2003), which examined 17 "within-study" comparisons of experimental and nonexperimental estimates of the impacts of training programs – i.e., studies that used both a randomized control group and a nonexperimental comparison group to estimate impacts for the same program. On the basis of their review, Glazerman et al. concluded that nonexperimental methods often produce estimates that differ from experimental findings by policy-relevant margins. The other paper that looks predominantly at nonexperimental validation studies for employment and training programs is Bloom et al. (2005). The bottom line of that assessment is that

> "…with respect to what methods could replace random assignment, we conclude that there are probably none that work **well** enough in a single replication, because the magnitude of [program group versus comparison group] mismatch bias for any given nonexperimental evaluation can be large (p. 224)."

Why It's Not Working (the Nonexperimental Approach)

The inconsistent performance of nonexperimental methods in evaluations of job training programs is not surprising. Job training programs are characterized by a selection process that is very difficult to replicate in choosing a nonexperimental comparison group. As noted earlier, the most common case is that individuals apply to training programs when they have lost their job. This means that, at the point of application, their earnings are atypically low. Even without any intervention, many of these individuals would become employed again and their earnings would rise. Exhibit 1 shows the path of monthly earnings from the National JTPA Study (Orr et al., 1995) over a 30-month period beginning 12 months before application to the program (month 0). As can be seen, average earnings of program applicants bottomed out in the month prior to application and then rose steadily for the next 18 months to a level roughly double the



pre-program level. This is *without any assistance from the JTPA program*; the graph charts the progress of the *control group* sample. This exhibit illustrates the famous "pre-program dip" first noted by Ashenfelter (1978), and the natural recovery from the dip.[9]

It is *the net addition* to this upward trajectory caused by the program that an experiment measures, using as its benchmark a control group that experiences the same pre-program dip as the training group and then exhibits the recovery from that dip that the training group *would have experienced* in the absence of training. To yield a valid estimate of program impact, a nonexperimental method must be able to replicate—either through selection of the comparison group or through statistical adjustments—both the pre-program dip and the subsequent natural recovery of earnings. Many of the methods frequently used in nonexperimental evaluations are not well-suited to this task.

For example, immediate pre-program earnings (in, say, months -8 to -1) cannot be used as the basis of matching program participants to a comparison group. Such an approach will almost certainly result in a comparison group with lower normal earnings than the participants, whose earnings are temporarily depressed. Comparison group earnings will stay down in the outcome period while participant earnings naturally rise even if the intervention has no effect. This will impart an upward bias on the participant-

[9] For more recent analyses of the National JTPA Study data with respect to this issue, see Heckman and Smith (1999).

minus-comparison-group impact estimates.

Nor can participant/comparison group differences be removed through time-invariant covariates (e.g., education, demographics, etc.) in impact regressions or by methods that model time-invariant error terms. The mismatch between participants and comparison group members concerns the *dynamics* of earnings patterns over time. This essentially rules out both the use of propensity score matching on baseline characteristics and fixed effects estimators.

We want to be clear that our position is *not* that nonexperimental methods are never successful. Our position is simply that one cannot count on their success *a priori* and – in the absence of a randomized evaluation of the same program – cannot reliably tell *ex post* whether they have been successful. From over 40 years of experience with these methods, the American evaluation community has come to the conclusion that, if we are to base policy on evaluation results, the stakes are too high to accept this kind of risk and uncertainty. Until the evaluation community is convinced that some nonexperimental method can produce consistently reliable estimates of program impact in a given policy area, policymakers in that area will remain skeptical of all nonexperimental estimates. To date, whenever such estimators have been tested against an experimental benchmark they have been found wanting.

However, our critique suggests neccessary critieria for a more reliable approach to designing non-experimental methods to estimate training impacts: Statistically control for (e.g., via regression, or better, propensity score matching) detailed patterns of pre-training employment and earnings when comparing participant and comparison group postprogram outcomes to obtain impact measures. The control variables used should include variables that measure the time pattern of earnings prior to job loss (this would have to be measured well before job loss) and the timing of job loss (i.e., binary employment indicators, perhaps by quarter). Recent work by Hollenbeck (2009) and Heinrich, Mueser, and Troske (2008) satisfies these necessary criteria.

Nevertheless, we suspect that these necessary criteria are not sufficient; i.e., that even these improved propensity score methods controlling for rich measures of recent employment and earnings will not replicate "gold standard" experimental results. These improved methods are simply not that different from the earlier approaches (e.g. Heckman, Ichimura, and Todd, 1977; Bloom, Michalopoulos, and Hill, 2005) that have failed replication. More precisely, we can sometimes find non-experimental methods that pass a replication test, but this is not enough. To be useful, we need an algorithm—a rule specified before looking at the data—that identifies which estimate will be used; and it is that estimate that needs to pass replication, i.e., to provide an unbiased result just as does an experiment.

It is possible that the new results imply such an algorithm and that it would replicate the experimental results. But this has not been tested, and we are skeptical. We therefore urge the European Commission (EC) not to proceed with a purely non-experimental approach until such an algorithm is proposed and shown to replicate

multiple experimental results. Experiments take many years and they are expensive. Nevertheless, the alternative—making policy based on flawed non-experimental methods—is much worse. The United States has gone down that path, spending billions of dollars on training programs which were later shown to have small or even negative impacts (e.g., JTPA; see Orr et al., 1995). Proceeding with unproven non-experimental evaluation methods as a guide to policy is setting up the EC to repeat America's mistakes.

Making Experiments More Feasible and Affordable

As a final point, we would note that recent advances in experimental methods in the U.S. are making random assignment studies more feasible and affordable. Feasibility has been enhanced by a number of methodological developments, including for example:

- Spreading the control group over many sites, so that very few individuals have to be turned away from program participation by the random assignment "lottery" in any location—a method used in the National Job Corps Study (Schochet et al 2001);
- Allowing program operators to increase the odds of assignment to the treatment group for preferred applicants (proposed for the Upward Bound evaluation; Olsen et al. 2007);
- Conducting "bump up" experiments in which *more* of the intervention is applied to the treatment group than in a normal program, rather than applying *less* than the customary amount to the control group (proposed for evaluating the impact of Unemployment Insurance benefits; Bell, 2003).

Beyond these methodological advances, advances in data collection strategies can substantially lower costs and increase data quality. Early evaluations of training programs used survey data. However, survey data have several major disadvantages: (i) high cost, leading to relatively small sample sizes; (ii) non-response bias due to imperfect survey tracking and refusals; (iii) large measurement error for contemporaneous outcomes (Duncan and Hill, 1983; Bound and Krueger, 1991; Bound, et al., 1994), (iv) limited retrospective histories due to the weakness of recall.

With the spread of computer technology in the administration of (near) universal public programs (e.g., social insurance programs), the role of surveys and thereby the cost of data collection for evaluations can decline sharply, while simultaneously increasing coverage, data quality, and earnings history. In most cases, intermediate and long-term follow-up can be left entirely to administrative data, such as Unemployment Insurance quarterly wage data or Social Security Administration annual earnings records. Surveys need only be used for short-term follow-up to determine usage of "similar" training services outside the program being studied and to capture richer descriptors of the employment obtained by sample members.

Existing direct comparisons suggest that findings from survey and administrative data are often qualitatively similar. However, administrative data clearly under-report earnings, apparently omitting earnings from the informal sector. (Kornfeld and Bloom 1999; Wallace and Haveman, 2007). There is also some evidence of differential non-

response between treatment and control groups in surveys (Schochet, Burghardt, and McConnell, 2008). In light of these mixed indicators, reliance on administrative sources of earnings data is certainly appealing for reasons of economy. It is on the economy and efficiency front that DOL now looks to improve its use of experiments.[10] That random assignment studies provide the "gold standard" of scientific reliability has for now been firmly established, as the main lesson of past and ongoing job training evaluations in the U.S.

[10] Discussions with DOL Employment and Training Administration evaluation staff, October 29, 2009.

References

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings", *Review of Economics and Statistics*, Vol. 60, Issue 1, pp. 47-57.

Barnow, Burt S. 1987. "The Impact of CETA Programs on Earnings: A Review of the Literature." *Journal of Human Resources 22* (Spring): 157-93.

Bell, Stephen H. 2003. *Review of Alternative Methodologies for Employment and Training Program Evaluation*. U.S. Department of Labor Employment and Training Administration, http://wdr.doleta.gov/research/keyword.cfm?fuseaction=dsp_resultDetails&pub_id=2369&bas_option=Keywords&start=1&usrt=4&stype=basic&sv=1&criteria=Methodologies.

Bell, Stephen H., Larry L. Orr, John D. Blomquist, and Glen G. Cain. 1995. *Program Applicants as a Comparison Group in Evaluating Training Programs*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Benus, Jacob, Sheena McConnell, Jeane Bellotti, Theodore Shen, Kenneth Forston, and Daver Kahvecioglu. 2008. *Growing America Through Entrepreneurship: Findings from the Evaluation of Project GATE*. Columbia, MD: IMPAQ International. LLC.

Betsey, Charles L., Robinson G. Hollister, and Mary R. Papageorgiou. 1985. *Youth Employment and Training Programs: The YEDPA Years.* Committee on Youth Employment Programs, Commission on Behavioral and Social Sciences and Education, National Research Council. Washington, D.C.: National Academy Press.

Bloom, Howard S., Charles Michalopoulos, and Carolyn J. Hill. 2005. "Using Experiments to Assess Nonexperimental Comparison-Group Methods for Measuring Program Effects," in *Learning More From Social Experiments: Evolving Analytic Approaches*, Howard S. Bloom ed. New York: Russell Sage Foundation, pp. 173-235.

Bound, John and Alan B. Krueger. 1991. "The Extent of Measurement Error in Longitudinal Data: Do Two Wrongs Make a Right?" *Journal of Labor Economics*. 9:1-24.

Bound, John, Charles Brown, Greg J. Duncan, and Willard L. Rodgers. 1994. "Evidence on the Validity of Cross-sectional and Longitudinal Labor Market Data?" *Journal of Labor Economics.* 12(3):345-368.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs", *Journal of the*

American Statistical Association, Volume 94, Number 448 (December), pp. 1053-1062.

Duncan, Greg J. and Daniel H. Hill. 1985. "An Investigation of the Extent and Consequences of Measurement Error in Labor-economic Survey Data." *Journal of Labor Economics*. 3(4):508-532.

Glazerman, Steven, Dan M. Levy, and David Myers. 2003. "Nonexperimental versus experimental estimates of earnings impacts", *Annals of the American Academy of Political and Social Science*, 589, 63–93.

Greenberg, David, Charles Michalopoulos, and Philip Robins. 2006. "Do experimental and nonexperimental evaluations give different answers about the effectiveness of government-funded training programs?" MDRC.

Greenberg, David, and Mark Shroder. 2004. *The Digest of Social Experiments, Third Edition.* Washington D.C.: The Urban Institute Press.

Heckman, James, and Joseph Hotz. 1989. "Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training." *Journal of the American Statistical Association*, 84, 862-880.

Heckman, James J., H. Ichimura, and P. Todd. 1997. "Matching As an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program" *Review of Economic Studies*, Vol. 64(4).

Heckman, James, and Jeffrey Smith. 1999. "The Pre-Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies." *Economic Journal*. 109(457): 313-348.

Heinrich, Carolyn J., Peter R. Mueser, and Kenneth Troske. 2008. "Workforce Investment Act Non-Experimental Net Impact Evaluation." IMPAQ International.

Hollenbeck, Kevin M. 2009. "Workforce Investment Act (WIA) Net Impact Estimates and Rates of Return" W.E. Upjohn Institute for Employment Research.

Kornfeld, Robert, and Howard S. Bloom. 1999. "Measuring program impacts on earnings and employment: Do unemployment insurance wage reports from employers agree with surveys of individuals?" *Journal of Labor Economics*, 17(1).

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review 76* (September): 604-20.

Mathematica Policy Research, Inc. 2009. "National Evaluation of the Workforce Investment Act", http://www.mathematica-mpr.com/Labor/wia.asp.

Maynard, Rebecca, and Thomas Fraker. 1987. "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs." *Journal of Human Resources 22* (Spring): 194-227

McConnell, Sheena, Elizabeth Stuart, Kenneth Fortson, Paul Decker, Irma Perez-Johnson, Barbara Harris, and Jeffrey Salzman. 2006. *Managing Customers' Training Choices: Findings from the Individual Training Account Experiment*. Washington D.C.: Mathematica Policy Research, Inc.

Miller, Cynthia, Johannes M. Bos, Kristin E. Porter, Fannie M. Tseng, and Yasuyo Abe. 2005. *The Challenge of Repeating Success in a Changing World: Final Report on the Center for Employment Training Replication Sites*. MDRC.

Olsen, Robert, Stephen Bell, and Jeremy Luallen. 2007. "A Novel Design for Improving External Validity in Random Assignment Experiments." Paper presented to the Annual Conference of the Association for Public Policy Analysis and Management, Abt Associates, Inc., November 2007.

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Fred Doolittle, Winston Lin, and George Cave. 1995. *Does Training for the Disadvantaged Work? Evidence from the National JTPA Study*. Washington, D.C.: The Urban Institute Press.

Pirog, Maureen A., Anne L. Buffardi, Colleen K. Chrisinger, Pradeep Singh, and John Briney. 2009. "Are the Alternatives to Random Assignment Nearly as Good? Statistical Corrections to Nonrandomized Evaluations", *Journal of Policy Analysis and Management*, Vol. 28, No. 1, pp. 169-172.

Rosenbaum Paul, and Donald B. Rubin. 1983 "The central role of propensity score in observational studies for causal effects", *Biometrika*, **70:**41-55.

Schirm, Alan., Elizabeth Stuart, and A McKie. 2006. *The Quantum Opportunity Program Demonstration: Final Impacts.* Washington, DC: Mathematica Policy Research, Inc.

Schochet, Peter Z., John Burghardt, and Steven Glazerman. 2001. *National Job Corps Study: The Impacts of Job Corps on Participants' Employment and Related Outcomes.* Princeton, NJ: Mathematica Policy Research.

Schochet, Peter Z., John Burghardt, and Sheena McConnell. 2008. "Does Job Corps Work? Impact Findings from the National Job Corps Study." *American Economic Review*, 98(5): 1864–86.

Smith, J. A., & Todd, P. E. 2005. "Does matching overcome LaLonde's critique of nonexperimental estimators?", *Journal of Econometrics*, 125, 305–353.

Stromsdorfer, Ernst, Howard Bloom, Robert Boruch, Michael Borus, Judith Gueron, A. Gustman, Peter Rossi, Fritz Scheuren, M. Smith, and F. Stafford. 1985. *Recommendations of the Job Training Longitudinal Survey Research Advisory Panel.* Washington, D.C.: Employment and Training Administration, U.S. Department of Labor.

Wallace, Geoffrey, and Robert Haveman. 2007. "The implications of differences between employer and worker employment/earnings reports for policy evaluation." *Journal of Policy Analysis and Management* 26(4), 737-753.