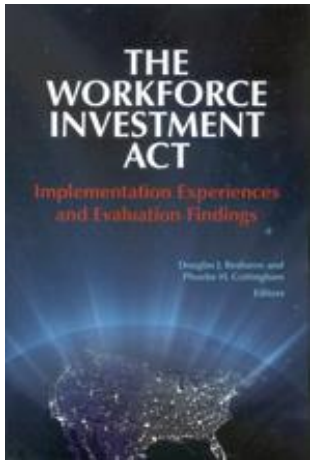
W.E. UPJOHN INSTITUTE
FOR EMPLOYMENT RESEARCH

Upjohn Institute Press

Improving Impact Evaluation in Europe



Jeffrey Smith
University of Michigan

Chapter 17 (pp. 473-494) in:
The Workforce Investment Act: Implementation Experiences and Evaluation Findings
Douglas J. Besharov and Phoebe H. Cottingham, eds.
Kalamazoo, MI: W.E. Upjohn Institute for Employment Research, 2011
DOI: 10.17848/9780880994026.ch17

The Workforce Investment Act

Implementation Experiences and Evaluation Findings

Douglas J. Besharov
Phoebe H. Cottingham
Editors

2011

W.E. Upjohn Institute for Employment Research
Kalamazoo, Michigan

Library of Congress Cataloging-in-Publication Data

The Workforce Investment Act : implementation experiences and evaluation findings / Douglas J. Besharov, Phoebe H. Cottingham, editors.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-88099-370-8 (pbk. : alk. paper)
ISBN-10: 0-88099-370-7 (pbk. : alk. paper)
ISBN-13: 978-0-88099-371-5 (hardcover : alk. paper)
ISBN-10: 0-88099-371-5 (hardcover : alk. paper)
1. Occupational training—Government policy—United States. 2. Occupational training—Government policy—United States—Evaluation. 3. Occupational training—Law and legislation—United States. 4. Employees—Training of—Law and legislation—United States. 5. Vocational guidance—Law and legislation—United States. 6. United States. Workforce Investment Act of 1998. I. Besharov, Douglas J. II. Cottingham, Phoebe H.

HD5715.2.W666 2011
370.1130973—dc22

2011010981

W.E. Upjohn Institute for Employment Research
300 S. Westnedge Avenue
Kalamazoo, Michigan 49007-4686

Cover design by Alcorn Publication Design.
Index prepared by Diane Worden.
Printed in the United States of America.
Printed on recycled paper.

17
Improving Impact Evaluation in Europe

Jeffrey Smith
University of Michigan

This chapter briefly addresses three themes related to the evaluation of active labor market programs (ALMPs), drawing on evidence from the North American experience and contrasting it with current practice in Europe.[1] I begin by making the (measured) case for greater use of random assignment methods in Europe, including both familiar and, I suspect, less familiar, arguments. Second, I make the case for greater (which in many European countries means "any") use of serious cost-benefit analysis as a component of the evaluation of ALMPs. Third, I discuss the organization of the evaluation "industry" in North America and offer some suggestions about lessons it provides for the organization of evaluation in Europe.

The conference came at an opportune time given the explosion in nonexperimental evaluation work related to ALMPs in Europe. The papers by Kluve (2006) and Card, Kluve, and Weber (2009) describe and meta-analyze this work; see also Bergemann and van den Berg (forthcoming). The European Social Fund surely deserves praise for venturing across the pond in search of ways to improve the quality and quantity of this evaluation work (broadly conceived to include performance management). At the same time, I think it well worth noting that the United States and Canada have much to learn from the countries at the top of the European evaluation league tables as well. Lessons worth learning include both the general value of rich, well-maintained, and relatively accessible (to qualified researchers and with appropriate privacy protections) administrative data and the value of specific data elements such as caseworker ratings of the employability of the unemployed and detailed, complete data on educational qualifications. Though this view may generate some controversy, I read the recent

nonexperimental evaluations of WIA by Heinrich et al. (2009) and Hollenbeck (2009) as indicating that existing U.S. administrative data systems do not quite have what it takes to provide compelling impact estimates. Perhaps this is not surprising given that the design of current U.S. administrative data systems did not include program evaluation as an objective. On another policy dimension, certain European countries have also done a good job of implementing, documenting, and studying regimes of sanctions for benefit recipients not sufficiently inspired by the "carrot" side of activation policies. Recent examples here include Arni, Lalive, and van Ours (2009), Boockmann, Thomsen, and Walter (2009), and Svarer (2007). The United States has sanctions in some programs, but to my knowledge, not much in the way of good data on them or—what follows immediately from the lack of good data—good studies. A related but different point concerns the sometime conflation in these sorts of discussions of U.S. policy with optimal policy. I make neither the claim that current U.S. policy is optimal in any meaningful sense for the current U.S. context or that all or even most of the good things about current U.S. evaluation policy can easily transfer to Europe. Nonetheless, I will argue for the view that some aspects of U.S. policy and practice suggest reforms worth considering in some (if not all) European countries.

The tremendous heterogeneity among European countries in the current state of research evaluating the performance of ALMPs and, more broadly, the heterogeneity in the relevant political and research institutions and in evaluation capacity also deserve note. Some European countries remain at the very beginning of the process of seriously evaluating their programs, while others have much to teach the North Americans. It nearly goes without saying that different aspects of the North American experience have relevance to different countries in Europe, depending on the current state of play in those countries.

Even on the topics directly covered in this chapter, much remains unsaid due to space limitations. In addition, I have not considered a variety of other topics closely related to the evaluation of ALMPs, such as recent developments in the literature regarding data and methods for nonexperimental evaluations (see, e.g., Dolton and Smith [2010]; Fredriksson and Johansson [2008]; Sianesi [2004]); performance management (see, e.g., Radin [2006]; Barnow and Smith [2004]; and Heckman, Heinrich, and Smith [2002]); statistical treatment rules (see,

e.g., Smith and Staghøj [2009] and the references therein); and the broader issue of the role of caseworkers as gatekeepers, monitors, and information providers (see, e.g., Lechner and Smith [2007] and Buurman and Dur [2008]). These omissions reflect not lack of interest or importance but rather division of labor over time and among authors.[2]

EXPERIMENTATION

As a quick perusal of the *Digest of the Social Experiments* (Greenberg and Shroder 2004) makes clear, the United States has conducted the vast majority (indeed, all but a handful) of social experiments, most of them related to active labor market programs, primary and secondary education, and the criminal justice system.[3] The situation has not really changed since the publication of that volume. In the United States, experiments have provided evidence of great value for both policy and for our understanding of social interventions more broadly in areas as diverse as health insurance, electricity pricing, responses to domestic violence, educational interventions related to teachers, schools, and curricula, and of course, ALMPs. Widely hailed in the social science community (see, e.g., Burtless and Orr [1986] and Burtless [1995]), the key advantage of social experiments is that their simple design makes them easy to explain and hard to argue with. This gives them a policy-influencing power not enjoyed by even the cleanest nonexperimental designs.

In addition to these direct benefits, experiments have the underappreciated benefit of providing high-quality data for other research purposes. In addition to the large literature that uses experimental impact estimates as a benchmark for the study of various combinations of nonexperimental estimators and data (see, e.g., LaLonde [1986], Fraker and Maynard [1987], Heckman and Hotz [1989], Friedlander and Robins [1995], Dehejia and Wahba [1999, 2002], and Smith and Todd [2005a,b]), experiments also have yielded a lot of substantive knowledge, particularly about low-income labor markets, and have provided a platform for methodological analyses of heterogeneous treatment effects that avoid the complications associated with first dealing with selection bias (see, e.g., Heckman, Smith, and Clements [1997], Bitler,

Gelbach, and Hoynes [2006], and Djebbari and Smith [2008]). Experimental data have even helped researchers to learn about structural models (in the sense that economists used that term), as in Todd and Wolpin (2006) and Lise, Seitz, and Smith (2004).

The literature documents a variety of limitations of experimental evaluations relative to nonexperimental evaluations. These limitations weigh against the advantages just discussed. At a most basic level, technological, political, and ethical concerns make it impossible to randomly assign some treatments of great interest, such as gender or family background. Except in unusual circumstances, such as the Progresa evaluation in Mexico, where random assignment took place at the level of relatively isolated villages, experimental evaluations capture only the partial equilibrium effects of policies (see Angelucci and di Giorgio [2009]). Depending on the placement of random assignment in the process of treatment receipt and on the availability of substitutes from other sources, both treatment group dropout and control group substitution often complicate the interpretation of the estimates from experimental evaluations of ALMPs (see the discussions in Heckman, Smith, and Taber [1998] and Heckman et al. [2000]).

The implementation of random assignment sometimes requires institutional changes that may compromise external validity. In the case of the National JTPA Study (NJS), the local sites in the experiment were concerned that the requirement of the design that they serve roughly the same number of participants while also filling a control group would mean digging deeper into the pool of potential participants. Depending on the nature of this pool and of the selection process, doing so could mean serving people with lower expected impacts. Some sites reacted to this by changing the nature of their selection process, e.g., reducing the number of visits to the center required to enroll, so as to reduce the extent of attrition during the process. Obviously, such changes compromise the external validity of the results. The scientific and political desirability of using volunteer sites also has implications for external validity. As documented in Doolittle and Traeger (1990), in the NJS, more than 200 of the (approximately) 600 local service delivery areas were contacted, and a substantial amount of money was spent on side payments in order to induce 16 sites to volunteer to participate, and even then at least one site left the experiment early. This issue often arises in evaluations of educational interventions conducted

by the Institute of Education Sciences (IES) at the U.S. Department of Education as well. A related but different point is that heterogeneity in the size and organization of local sites may limit the set of sites at which it makes budgetary sense to do random assignment. The presence of random assignment may also alter the behavior of potential participants in ways that less salient and intrusive nonexperimental methods might not. For example, it might induce additional selection on risk aversion, or it might deter complementary investments. Such changes, sometimes dubbed "randomization bias" in the literature, are distinct from Hawthorne effects, which result from the mere fact of observation, and pose yet another threat to external validity. Heckman and Smith (1995) and Section 5 of Heckman, LaLonde, and Smith (1999) summarize these concerns about experiments.

In addition to these real issues, policymakers and program administrators sometimes offer ethical objections to random assignment. In my experience, these objections nearly always represent a cover for simply not wanting to know the answer. Experiments often provide compelling evidence that treatments do not work at all or do not work well enough to pass a cost-benefit test. Educational researchers have dubbed the What Works Clearinghouse, a formal compendium of quality-rated evidence on the impacts of educational treatments funded by the IES and operated by Mathematica Policy Research, the "Nothing Works Clearinghouse."[4] This usage illustrates the very real empirical pattern that many, maybe most, programs fail when subjected to serious evaluation. Programs that deliver ineffective treatments, and thus do not benefit their participants, still benefit important constituencies, such as the workers and agencies or firms that provide the treatments. Indeed, one sometimes suspects that it is these constituencies, and not the population served, who represent the real reason for the program's existence in the first place. These constituencies have an interest in the production of low-quality (and sometimes deliberately manipulated), nonexperimental evaluations and misleading performance measures in place of compelling experimental (or even nonexperimental) evidence.

One way to confront these specious ethical arguments is to point out what they miss, namely the problematic ethical position of forcing taxpayers to fund programs without any serious evidence that they pass cost-benefit tests when such evidence could easily be produced. Such "speaking truth to power" provides the warm glow of righteous

satisfaction and carries some sway with stakeholders not completely in the service of their own narrow interests, but it does not always carry the day.

Variants of random assignment that do not require the complete denial of service to any potential clients constitute another response to the phony ethical arguments offered up against random assignment, as these arguments typically revolve around concerns about service denial. In contexts where some eligible individuals would not receive service anyway, advocates of serious evaluation can (and do) frame random assignment as an equitable way to allocate scarce resources. In contexts where resource constraints do not bind, variants of random assignment that do not assign anyone to a no-services control group can help to derail malicious objections.

The literature offers three variants of random assignment that (more or less) avoid a no-treatment control group. One rather obvious variant consists of random assignment with multiple treatment arms but no control arm. For example, in the WIA context one might randomly assign some clients to only core and intensive services, while excluding them from training services. Another variant consists of a randomized encouragement design, as in Hirano et al. (2000). Here eligible individuals get randomly assigned an incentive to participate. Thus, no one is excluded, but the incentive, when properly designed—learning about the impact of the incentive represents a side benefit of the design—induces exogenous variation in treatment status. The design identifies what the literature calls the local average treatment effect (LATE) rather than the average treatment effect on the treated. Put less technically, this design identifies the mean impact on those induced to participate by the incentive, but not the mean impact on all participants. Whether or not this parameter merits attention depends on the particular policy context. The final design consists of randomization at the margin, as in Black et al. (2003). This design does create a no-treatment control group, but only of individuals on the margin of participation. In the case of the Kentucky Worker Profiling and Reemployment Services System analyzed in Black et al. (2003), the margin consists of individuals whose predicted durations of benefit receipt put them in the last cell of treated individuals in a given local office in a given week. The state was willing to randomize these individuals but not those with long predicted spells. Like the randomized encouragement design, this design does not iden-

tify the average treatment effect on the treated, but it does identify the average impact of treatment for individuals at the margin of treatment. This parameter answers a different policy question of what the mean impact would be on individuals brought into the program by an increase in the number of slots. As with the randomized encouragement design, this parameter might have greater or lesser policy importance than the average treatment effect on the treated.

The push for random assignment evaluations of ALMPs (and other policies as well) ultimately has great value. For example, the zero (and sometimes negative) impact estimates for youth in the NJS led to large budget cuts in that program—cuts an order of magnitude larger than the cost of this (quite expensive) evaluation; see the discussion in Heckman and Krueger (2003). The experimental findings from the National Job Corps Study presented in Schochet, Burghardt, and McConnell (2008), which include positive impacts that fade out and so fail to pass a cost-benefit test given the high cost of the program, have led to some serious thinking about that popular and, prior to the evaluation, essentially untouchable program. Some of the IES experimental evaluation results, such as those for the Teach for America Program (Glazerman, Mayer, and Decker 2005), abstinence-only sex education programs (Trenholm et al. 2008), reading and mathematics software (Campuzano et al. 2009), and intensive teacher mentoring programs (Eisenberg et al. 2009), have had real impacts on expenditures and on the course of policy innovation and research. The Europeans can and should get in on this worthwhile game.

COST-BENEFIT ANALYSIS

Cost-benefit analysis combines impact estimates with information on program costs to produce a direct policy conclusion. In the case of impact estimates that capture the average effect of treatment on the treated, a comparison of the impacts with the average cost of the program provides a clear and direct message about the value of a program to the taxpayers who fund it. Historically, many U.S. evaluations have included at least rudimentary cost-benefit analyses. The cost-benefit analysis associated with the National Job Corps Study presented in

Schochet, Burghardt and McConnell (2006) represents a particularly fine example.

In contrast, one can look pretty hard and not find very many European ALMP evaluations that include serious cost-benefit analyses. Munch, Skipper, and Jespersen (2008) provide a notable Danish example, while Raaum, Torp, and Zhang (2002) do the same for Norway. Osikominu (2009) shows a more common situation, with only a very rudimentary comparison of costs and impacts. More generally, and despite these counterexamples, the modal European ALMP evaluation, at least in my experience, contains no cost-benefit analysis at all.

A number of reasons are given for the absence of cost-benefit analysis in European evaluations of ALMPs, the most common of which concerns the European focus on employment impacts, rather than earnings impacts, mainly for political reasons. This focus on employment has led to a lack of good administrative data on earnings in some countries, which makes cost-benefit analysis more challenging, as the researcher (or the literature more broadly) must come up with a compelling way to translate employment impacts into monetary units. In contrast, impacts on earnings, the most common case in North America, fit easily into a cost-benefit framework. Another reason sometimes given for the absence of cost-benefit analyses in Europe relates to the fact that the estimated employment impacts often turn out negative or zero or, in the bright and sunny cases, positive but small enough to make the negative result that would emerge from a serious cost-benefit analysis obvious in advance. This is the "why bother when the programs do not really work anyway" argument, and it has some sense to it.

The lack of good cost data also poses a barrier to serious cost-benefit analysis in many European contexts (and some North American ones as well). Ideally, one would have detailed data on both average and marginal costs for each service offered, broken down geographically in cases where costs varied substantially by, for example, location in a large city, a small city, or a rural area. Instead, researchers often have available little more than the program budget and the total number of persons served.

Both JTPA and WIA have attempted performance standards measures that included a cost component. These have faced real difficulties in assigning costs shared by JTPA or WIA and other programs, as when a variety of programs, often each having multiple funding sources, all

share a common physical location as a One-Stop center. These common cost allocation issues (and others) are real and challenging, and carry over directly from performance measures to the problem of creating meaningful cost information for use in cost-benefit calculations. At the same time, private firms face similar difficulties and a large literature and equally large body of empirical practice in accounting lay out reasonable ways to deal with them.

In addition to its value at informing decisions about keeping or dropping programs, cost-benefit analysis has the further benefit of encouraging thinking about important aspects of program design and evaluation, and of public policy more generally. First, it encourages thinking about the outcomes an ALMP will affect. A focus on outcomes other than just earnings, in particular on crime, represents one of the notable aspects of the Job Corp cost-benefit analysis highlighted earlier. Not only do impacts on crime account for much of the gross impact of the program, particularly in the short term, their presence tells us a lot about how the program works, and suggests other possible treatments that might well pass a cost-benefit analysis.

Thinking about outcomes and about the behavioral theory that links treatments to outcomes also leads to a salutary focus on the possible general equilibrium effects (which include spillovers or displacement effects) of programs. Johnson (1980) and Calmfors (1994) are classic references; see Lise, Seitz, and Smith (2004) and the citations it contains for pointers to the more recent (and still much too small) literature. While difficult to estimate, they deserve a place in cost-benefit analyses, if only in the form of a sensitivity analysis using informal estimates drawn from the broader literature.

Thinking about cost-benefit analysis in a serious way also highlights the importance of learning about the duration of program impacts. Most evaluations of ALMPs provide only a year or two of follow-up. The available evidence on longer-term impacts suggests that sometimes impacts remain remarkably steady over time for years after an intervention, as in the National Supported Work Demonstration (Couch 1992) and the National JTPA Study (GAO 1996); other times they fade out, as in the National Job Corps Study (Schochet, Burghardt, and McConnell 2008) and the California GAIN program (Hotz, Imbens, and Klerman, 2006); and other times they appear only belatedly, as in the evaluation of German training programs by Lechner, Miquel, and Wunsch

(2004). The absence of both a clear general empirical pattern and compelling theory on when estimates should persist and when not suggests the value of more frequently undertaking long-term follow-up, so as to minimize the impact of extrapolation of the sort described in Heckman, LaLonde, and Smith (1999).

Finally, paying attention to cost-benefit analysis focuses policy and research attention on two important parameters: the discount rate and the marginal social cost of public funds or "excess burden." Having a well-justified social discount rate for use in government budgeting and investment decisions represents a basic task of public finance economists. As noted in Heckman, LaLonde, and Smith (1999), the discount rate employed to bring future net impacts (and costs, if applicable) forward in time to the present can affect the outcome of a cost-benefit analysis. Also important, and routinely ignored in North American cost-benefit analyses (including otherwise exemplary ones like that from the Job Corp evaluation), is the fact that a dollar of government budget for ALMPs costs society more than a dollar, both because the operation of the tax system directly consumes real resources (all those cheery Internal Revenue Service agents have to get paid) and because all developed countries rely on distortionary tax systems. While estimates of the marginal social cost of public funds vary widely in the literature even for specific countries, and we would expect them to vary across countries due to differences in tax systems and tax rates and other institutional features, the estimates never equal zero and often reach magnitudes that suggest the policy importance of incorporating this factor into cost-benefit analyses and thereby into decisions about program existence and funding (see, e.g., Auerbach and Hines [2002] for a survey).

In sum, cost-benefit analysis represents a useful tool, both in a direct sense via its role in clarifying and systematizing decisions about program existence, expansion, or contraction, and indirectly via its direction of policy and research attention to important, but often neglected, issues of program design and impact and of public finance more broadly.

ORGANIZING EVALUATION RESEARCH

Surprisingly little research seeks to document and explain differences in the quantity and quality of ALMP evaluation across countries. I am aware of Riddell (1991) and not much else. Given the heterogeneity in both quality and quantity obvious even to the most casual observer, this gap in the literature comes as a surprise. Filling the gap represents a worthy task for researchers. Because of this gap, my remarks here rely mainly on my own observations as a scholar studying evaluation methods, a provider of evaluation short courses to graduate students at various locations in Europe, a referee and editor handling academic evaluations, and an occasional evaluation consultant as well as on discussions with friends in the academic and policy worlds. The lack of quantitative evidence on national variation in quality and quantity necessitates the following caveat: I am well aware that low-quality research, such as PriceWaterhouseCoopers (2004, p. 15), with its smiley faces and confusion of outcome levels and impacts, or Gregory (2000), with its distinctive "sites of oppression matrix" evaluation tool, appear everywhere, including the United States and Canada, because of the universal demand for evaluation reports that promote the views of interested parties while providing an appearance of technical understanding and objectivity sufficient to fool the reading public.

I will argue that differences in the quality and quantity of evaluation research across countries result from much more than simply differences in the industrial organization of the evaluation industry, but those differences play a role and make a good place to start my discussion. The evaluation industry in the United States combines government, private for-profit firms, private nonprofit firms, and academia in remarkable and complex ways that differ across program types. For ALMPs, both nonprofit and for-profit firms, operating on contract to the USDOL, have undertaken many of the evaluations of large programs such as JTPA, WIA, the Job Corps, the Trade Adjustment Act, and so on. Additional evaluation work is performed by academics operating with research funding from places like the National Science Foundation or private foundations; this work often uses data from the original USDOL-funded evaluations, as with the long series of papers by Heckman and various coauthors using the data from the NJS; see Heckman

et al. (1998) for an example. Other evaluation work, including process evaluation work, is also often contracted out to a somewhat wider set of firms than the small number of large firms (e.g., Abt, Mathematica, MDRC, etc.) with the capacity to undertake large evaluations. These firms compete in both the product market and the labor market; at least in regard to economists, they compete for the same newly minted doctorates as academic economics departments just outside the top 20. Some evaluation work is also done in-house at the USDOL, whose staff includes people trained in economics at the doctoral level. A similar pattern holds in the education world, though probably with more academic involvement in the actual performance of the evaluation work, as opposed to simply advising or undertaking secondary analyses using the data generated by evaluations conducted by others.

What makes the European evaluation market different from the North American ones? First, some European countries have an important player in their markets that is absent in the United States in the form of (mostly or entirely) government-supported research institutes devoted to labor market policy and evaluation that operate (more or less) at "arm's length" from the government itself. I have in mind here the IFAU in Sweden and the various institutes in Germany (e.g., the ZEW in Mannheim, the IZA in Bonn, the DIW in Berlin, and the RWI in Essen). My understanding is that these institutes both have base funding and do work on contract. They maintain a remarkable degree of independence, in the sense that they routinely report evaluation results indicating that ALMPs have zero or even negative impacts (and other more humorous but still somewhat embarrassing-to-the-government findings such as paternal leave being more common during hunting season and such like).

Neither the United States nor Canada has any direct analog to these institutes. The GAO does some work along the lines of process and implementation evaluation, but not much in the way of econometric impact evaluation.[5] The closest analogue in Canada, the Auditor General, is even less like the European Institutes. The U.S. Congressional Research Service largely confines itself to literature surveys. While I could imagine the Canadians setting up something like the IFAU, I find it hard to imagine the United States doing so, in part because it would present real competition to the various DC think tanks. These institutes

represent a valuable component of the European scene, and countries that do not have them ought to reconsider.

Size represents a second important contrast between the evaluation market in the United States and that in Europe (and in Canada, for that matter). Size has two relevant dimensions here. The first is the simple magnitude of evaluation research going on. The United States spends quite a lot of money on evaluation in a number of policy areas, including for programs that it funds in developing countries. To the extent that evaluation firms, whether for-profit or not-for-profit, have economies of scale over some range, a larger market can support more firms and thus allow more competition between firms. The second dimension of size concerns the number of potential clients for evaluation research firms. My sense is that evaluation firms in the United States face many more potential clients both at the national level (where they might deal with the departments of labor, education, housing and urban development, health and human services, homeland security, transportation, agriculture, and so on, and in some cases even separate parts of particular departments), as well as the development banks, states and larger cities, and private foundations. This diversity of potential clients reduces the dependence of the firm on repeated interactions with a single client and thus, I think, reduces the potential costs associated with catering to the truth rather than to the client agency. Firms in smaller European countries with highly centralized governments and no private foundations may face a much, much smaller number of potential clients and thus face much stronger pressure to bend to the client's wishes of the moment.

One easy way to increase the size of the European evaluation market is for that market to become truly European rather than national. At present, I am aware of very little evaluation work that happens across boundaries in Europe. Transforming small national markets into a much larger European market would allow greater competition between providers and would give firms more freedom to avoid clients seeking a particular answer rather than necessarily the correct answer. I think entry by the major U.S. firms into the European market would aid in these developments. This has happened in a very limited way in the UK, with MDRC playing a role in the experimental evaluation of the Employment Retention and Advancement Demonstration (Miller et al. 2008). More activity on this front would, in my view, bring great benefits.[6]

In this context, the Association for Public Policy and Management (APPAM) is important because it fosters interactions between academics, government consumers and producers of evaluation research, evaluation firms, and policy people interested in the results of evaluations. Bringing these groups together, both via the annual meetings and via APPAM's publications and other activities, represents an important contribution not duplicated, to my knowledge, by any European organization. Efforts to replicate APPAM in Europe, with some linkages and occasional joint conferences as with the Society for Labor Economics in North America and its younger European compatriot the European Association of Labor Economists, would add value.

Finally, you have to want it. At a narrow level, this means having at least some people in government who care about evidence more than they care about the party line or about their narrow bureaucratic imperatives of budget increase and career advancement. It needs to encompass both the levels of administration that change at election time and those that do not. It also means that some people at both levels have to understand enough about evaluation to know what to ask for and to evaluate what gets produced in response. I think the U.S. practice of having serious academics spend brief stints in the national administration, say, as chief economist at the USDOL or on the Council of Economic Advisers, plays an important role in the (very much relative) success the United States has had on this dimension, and commend such institutions to European governments. The temporary nature of the appointments matters here precisely because you do not want the academics to assimilate into the bureaucratic culture. Rather, you want them to maintain their outsider perspective and their academic devotion to getting the right answer (helped along by the threat of ridicule from their university friends and colleagues if they sell out).

The George W. Bush administration provides a useful illustration here. At the Labor Department, evaluation research became a low priority during this administration. More broadly, the department had such a poor reputation in regard to its interest in evidence that it could not manage to fill the chief economist position with a serious academic economist (for eight years!). Contrast this to the distinguished list of chief economists under Clinton, which included Larry Katz and Alan Krueger. In contrast, less than one mile away, the U.S. Department of Education—in particular, the IES under Russ Whitehurst—made a seri-

ous run at transforming the entire field of educational policy evaluation through a program of experimental and high-quality nonexperimental evaluations, as well as the funding of a training grant program to create a generation of new, quantitative, serious education policy evaluators with disciplinary roots at least partially outside of traditional schools of education (see the discussion in IES [2008]). How do you create more places like IES? I must confess that I do not have a good answer here, but we should be thinking about it, because doing so has a very high payoff indeed.

More broadly, the demand for serious program evaluation has to come from somewhere. It can come from leaders within government. It can come from actors outside government, such as the media and public intellectuals. It can come from the general public. But it must come from somewhere. Casual empiricism suggests a link at the country level between the quality and quantity of evaluation and the imprint of neoclassical economics. Countries with long neoclassical traditions, including the UK, the Netherlands, and the Nordic countries, are pretty much the same as those with long traditions of serious research devoted to the evaluation of social programs. Looking within countries, Germany has gotten serious about empirical evaluation research only in the last 15 years or so, a time period that coincides with the triumph of neoclassical economics within academic economics in that country. This observed link between the demand for evaluation and neoclassical economics might reflect a causal relationship. Alternatively, both demand for serious policy evaluation and the dominance of neoclassical economics may reflect broader and deeper differences across countries in individualism, deference to authority, the importance of social class, average education, and so on. Regardless of whether the current relationship reflects causality or not, one might argue that increasing the number of individuals trained in economics, particularly a practical version of economics rather than just high theory or theoretical econometrics, at both the undergraduate and graduate levels might represent a long-term strategy for increasing the demand for quality policy evaluation, as well as the ability to supply it with domestic labor. Who knows, it might even improve European agricultural policy as well!

CONCLUDING REMARKS

This chapter has touched on three important areas where the European Social Fund can learn from the North American experience in evaluating ALMPs. I have argued that current European practice lies very far from the point where the marginal value of additional experimental evaluations would equal their marginal cost. I have also argued that Europe would benefit from much greater attention to careful cost-benefit analysis following evaluation. Such analyses would allow the evaluation results to provide more guidance to policy and, more broadly, would increase our understanding of how policy works and so aid in the design of future policies. Finally, I have argued that much room remains for improving the organization of evaluation in Europe. The European environment includes distinctive and valuable aspects not present in North America, but could usefully incorporate aspects of the North American experience as it seeks to improve the overall quality of European evaluations.

Notes

My thoughts on the issues discussed in this chapter have benefited from my interactions with a number of scholars over the years, including (but not limited to) Jim Heckman, Dan Black, Michael Lechner, Carolyn Heinrich, Burt Barnow, Lars Skipper, and Arthur Sweetman. I am very grateful for those interactions, and for comments from Jessica Goldberg, but, of course, retain all responsibility for the (occasionally provocative) views expressed here.

1. I use *North American* in the Canadian manner to mean the United States and Canada but not Mexico.
2. See Smith (2000, 2004) for broad nontechnical surveys of evaluation methodology. See Heckman, LaLonde, and Smith (1999), Imbens and Wooldridge (2009), and Blundell and Costa-Dias (2009) for somewhat more technical surveys. See Heckman and Abbring (2007) and Heckman and Vytlacil (2007a,b) for recent technical overviews.
3. I distinguish here between social experiments and both laboratory experiments under fully controlled conditions and the small-scale field experiments that have taken the development literature by storm over the last decade. For discussions and categorizations, see, e.g., Levitt and List (2009) and Banerjee and Duflo (2009).
4. The What Works Clearinghouse can be found at http://ies.ed.gov/ncee/wwc/.

5. For an exception, see GAO (1996), which presents long-term impact estimates for the JTPA experiment using administrative data.
6. This same point applies to Canada as well.

References

Angelucci, Manuela, and Giacomo di Giorgio. 2009. "Indirect Effects of an Aid Program: How Do Cash Transfers Affect Ineligibles' Consumption?" *American Economic Review* 99(1): 486–508.

Arni, Patrick, Rafael Lalive, and Jan van Ours. 2009. "How Effective Are Unemployment Benefit Sanctions? Looking Beyond Unemployment Exit." IZA Discussion Paper No. 4509. Bonn: IZA.

Auerbach, Alan, and James Hines. 2002. "Taxation and Economic Efficiency." In *Handbook of Public Finance, Volume 3*, Alan Auerbach and Martin Feldstein, eds. Amsterdam: North-Holland, pp. 1347–1421.

Banerjee, Abhijit, and Esther Duflo. 2009. "The Experimental Approach to Development Economics." *Annual Review of Economics* 1: 151–178.

Barnow, Burt, and Jeffrey Smith. 2004. "Performance Management of U.S. Job Training Programs: Lessons from the Job Training Partnership Act." *Public Finance and Management* 4(3): 247–287.

Bergemann, Annette, and Gerard van den Berg. 2008. "Active Labor Market Policy Effects for Women in Europe—A Survey." *Annales d'Economie et de Statistique* 91/92: 385–408.

Bitler, Marianne, Jonah Gelbach, and Hilary Hoynes. 2006. "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments." *American Economic Review* 96(4): 988–1012.

Black, Dan, Jeffrey Smith, Mark Berger, and Brett Noel. 2003. "Is the Threat of Reemployment Services More Effective than the Services Themselves? Evidence from Random Assignment in the UI System." *American Economic Review* 93(4): 1313–1327.

Blundell, Richard, and Monica Costa-Dias. 2009. "Alternative Approaches to Evaluation in Empirical Microeconomics." *Journal of Human Resources* 44(3): 565–640.

Boockmann, Bernhard, Stephan Thomsen, and Thomas Walter. 2009. "Intensifying the Use of Benefit Sanctions—An Effective Tool to Shorten Welfare Receipt and Speed Up Transitions to Employment?" Unpublished manuscript. Universität Magdeburg, Magdeburg, Germany.

Burtless, Gary. 1995. "The Case for Randomized Field Trials in Economic and Policy Research." *Journal of Economic Perspectives* 9(2): 63–84.

Burtless, Gary, and Larry Orr. 1986. "Are Classical Experiments Needed for Manpower Policy?" *Journal of Human Resources* 21(4): 606–639.

Buurman, Margaretha, and Robert Dur. 2008. "Incentives and the Sorting of Altruistic Agents into Street-Level Bureaucracies." IZA Discussion Paper No. 3847. Bonn: IZA.

Calmfors, Lars. 1994. "Active Labour Market Policy and Unemployment: A Framework for the Analysis of Crucial Design Features." *OECD Economic Studies* 22: 7–47.

Campuzano, Larissa, Mark Dynarski, Roberto Agodini, and Kristina Rall. 2009. *Effectiveness of Reading and Mathematics Software Products Findings from Two Student Cohorts.* NCEE 2009-4041. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Card, David, Jochen Kluve, and Andrea Weber. 2009. "Active Labor Market Policy Evaluations: A Meta-Analysis." IZA Discussion Paper No. 4002. Bonn: IZA.

Couch, Kenneth. 1992. "New Evidence on the Long-Term Effects of Employment Training Programs." *Journal of Labor Economics* 10(4): 380–388.

Dehejia, Rajeev, and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association.* 94(448): 1053–1062.

———. 2002. "Propensity Score Matching Methods for Non-experimental Causal Studies." *Review of Economics and Statistics* 84(1): 151–161.

Djebbari, Habiba, and Jeffrey Smith. 2008. "Heterogeneous Program Impacts: Experimental Evidence from the PROGRESA Program." *Journal of Econometrics* 145(1–2): 64–80.

Dolton, Peter, and Jeffrey Smith. 2010. "The Econometric Evaluation of the New Deal for Lone Parents." Unpublished manuscript. University of Michigan, Ann Arbor.

Doolittle, Frederick, and Linda Traeger. 1990. *Implementing the National JTPA Study*. New York: Manpower Demonstration Research Corporation.

Eisenberg, Erik, Steven Glazerman, Martha Bleeker, Amy Johnson, Julieta Lugo-Gil, Mary Grider, and Sarah Dolfin. 2009. *Impacts of Comprehensive Teacher Induction: Results from the Second Year of a Randomized Controlled Study,* NCEE 2009-4072. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Fraker, Thomas, and Rebecca Maynard. 1987. "The Adequacy of Comparison Group Designs for Evaluation of Employment-Related Programs." *Journal of Human Resources* 22(2): 194–227.

Fredriksson, Peter, and Per Johansson. 2008. "Program Evaluation and Random Program Starts." *Journal of Business and Economic Statistics* 26(4): 435–445.

Friedlander, Daniel, and Philip Robins. 1995. "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods." *American Economic Review* 85(4): 923–937.

Glazerman, Steven, Daniel Mayer, and Paul Decker. 2005. "Alternative Routes to Teaching: The Impacts of Teach for America on Student Achievement and Other Outcomes." *Journal of Policy Analysis and Management* 25(1): 75–96.

Government Accountability Office (GAO). 1996. *Job Training Partnership Act: Long-Term Earnings and Employment Outcomes*. Report HEHS-96-40. Washington, DC: U.S. Government Printing Office.

Greenberg, David, and Mark Shroder. 2004. *Digest of the Social Experiments.* 3rd ed. Washington, DC: Urban Institute Press.

Gregory, Amanda. 2000. "Problematizing Participation: A Critical Review of Approaches to Participation in Evaluation Theory." *Evaluation* 6(2): 179–199.

Heckman, James J., and Jaap Abbring. 2007. "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation." In *Handbook of Econometrics, Volume 6B,* James J. Heckman and Edward Leamer, eds. Amsterdam: Elsevier, pp. 5145–5303.

Heckman, James J., Carolyn Heinrich, and Jeffrey Smith. 2002. "The Performance of Performance Standards." *Journal of Human Resources* 37(4): 778–811.

Heckman, James J., Neil Hohmann, Jeffrey Smith, and Michael Khoo. 2000. "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment." *Quarterly Journal of Economics* 115(2): 651–694.

Heckman, James J., and V. Joseph Hotz. 1989. "Choosing among Alternative Methods of Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84(408): 862–874.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66(5): 1017–1098.

Heckman, James J., and Alan Krueger. 2003. *Inequality in America: What Role for Human Capital Policies*. Cambridge, MA: MIT Press.

Heckman, James J., and Jeffrey Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives* 9(2): 85–110.

Heckman, James J., Robert LaLonde, and Jeffrey Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics, Volume 3A*, Orley Ashenfelter and David Card, eds. Amsterdam: North Holland, pp. 1865–2097.

Heckman, James J., Jeffrey Smith, and Nancy Clements. 1997. "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies* 64(4): 487–535.

Heckman, James J., Jeffrey Smith, and Christopher Taber. 1998. "Accounting for Dropouts in Evaluations of Social Programs." *Review of Economics and Statistics* 80(1): 1–14.

Heckman, James J., and Edward Vytlacil. 2007a. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In *Handbook of Econometrics, Volume 6B*, James J. Heckman and Edward Leamer, eds. Amsterdam: Elsevier, pp. 4779–4874.

———. 2007b. "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Economic Estimators to Evaluate Social Programs and to Forecast Their Effects in New Environments." In *Handbook of Econometrics, Volume 6B*, James J. Heckman and Edward Leamer, eds. Amsterdam: Elsevier, pp. 4875–5144.

Heinrich, Carolyn, Peter Mueser, Kenneth Troske, Kyung-Seong Jeon, and Daver Kahvecioglu. 2009. "New Estimates of Public Employment and Training Program Net Impacts: A Nonexperimental Evaluation of the Workforce Investment Act Program." IZA Discussion Paper No. 4569. Bonn: IZA.

Hirano, Kei, Guido Imbens, Donald Rubin, and Xiao-Hua Zhou. 2000. "Assessing the Effect of an Influenza Vaccine in an Encouragement Design." *Biostatistics* 1: 69–88.

Hollenbeck, Kevin. 2009. "Workforce Investment Act (WIA) Net Impact Estimates and Rates of Return." Unpublished manuscript. W.E. Upjohn Institute for Employment Research, Kalamazoo, MI. See http://research.upjohn.org/confpapers/2/ (accessed April 7, 2011).

Hotz, V. Joseph, Guido Imbens, and Jacob Klerman. 2006. "Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Re-Analysis of the California GAIN Program." *Journal of Labor Economics* 24(3): 521–566.

Imbens, Guido, and Jeffrey Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47(1): 5–86.

Institute of Education Sciences, U.S. Department of Education (IES). 2008. *Rigor and Relevance Redux: Director's Biennial Report to Congress.* IES 2009-6010. Washington, DC: IES.

Johnson, George. 1980. "The Theory of Labor Market Intervention." *Economica* 47(187): 309–329.

Kluve, Jochen. 2006. "The Effectiveness of European Active Labor Market Policies." IZA Discussion Paper No. 2018. Bonn: IZA.

LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4): 604–620.

Lechner, Michael, Ruth Miquel, and Conny Wunsch. 2004. "Long-Run Effects of Public Sector Sponsored Training in West Germany." IZA Discussion Paper No. 1443. Bonn: IZA.

Lechner, Michael, and Jeffrey Smith. 2007. "What Is the Value Added by Caseworkers?" *Labour Economics* 14(2): 135–151.

Levitt, Steven, and John List. 2009. "Field Experiments in Economics: The Past, the Present, and the Future." *European Economic Review* 53(1): 1–18.

Lise, Jeremy, Shannon Seitz, and Jeffrey Smith. 2004. "Equilibrium Policy Experiments and the Evaluation of Social Programs." NBER Working Paper No. 10283. Cambridge, MA: NBER.

Miller, Cynthia, Helen Bewley, Verity Campbell-Barr, Richard Dorsett, Gayle Hamilton, Lesley Hoggart, Tatiana Homonoff, Alan Marsh, Kathryn Ray, James Riccio, and Sandra Vegeris. 2008. *Employment Retention and Advancement (ERA) Demonstration: Implementation and Second-Year Impacts for New Deal 25 Plus Customers in the UK*. New York: MDRC.

Munch, Jakob, Lars Skipper, and Svend Jespersen. 2008. "Costs and Benefits of Danish Active Labour Market Programmes." *Labour Economics* 15(5): 859–884.

Osikominu, Aderonke. 2009. "Quick Job Entry or Long-Term Human Capital Development? The Dynamic Effects of Alternative Training Schemes." Unpublished manuscript. Breisgau, Germany: Albert-Ludwigs Universität Freiburg.

PriceWaterhouseCoopers. 2004. *New Deal 25+ Evaluation Report No. 9*. London: Department for Employment and Learning. http://www.delni.gov.uk/new_deal25plus_final_report.pdf (accessed November 20, 2010).

Raaum, Oddbjørn, Hege Torp, and Tao Zhang. 2002. "Do Individual Programme Effects Exceed the Costs? Norwegian Evidence on Long Run Effects of Labour Market Training." Memorandum No. 15/2002. Oslo: Department of Economics, University of Oslo.

Radin, Beryl. 2006. *Challenging the Performance Movement: Accountability, Complexity and Democratic Values*. Washington, DC: Georgetown University Press.

Riddell, Craig. 1991. "Evaluation of Manpower and Training Programs: The North American Experience." In *The Evaluation of Manpower, Training*

and Social Programs: The State of a Complex Art. Paris: OECD, pp. 43–72.
Schochet, Peter, John Burghardt, and Sheena McConnell. 2006. *National Job Corps Study and Longer-Term Follow-Up Study: Impact and Benefit-Cost Findings Using Survey and Summary Earnings Records Data, Final Report*. Washington, DC: Mathematica Policy Research.
———. 2008. "Does Job Corps Work? Impact Findings from the National Job Corps Study." *American Economic Review* 98(5): 1864–1886.
Sianesi, Barbara. 2004. "An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s." *Review of Economics and Statistics* 86(1): 133–155.
Smith, Jeffrey. 2000. "A Critical Survey of Empirical Methods for Evaluating Employment and Training Programs." *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 136(3): 247–268.
———. 2004. "Evaluating Local Economic Development Policies: Theory and Practice." In *Evaluating Local Economic and Employment Development: How to Assess What Works among Programmes and Policies*, Alistair Nolan and Ging Wong, eds. Paris: OECD, pp. 287–332.
Smith, Jeffrey, and Jonas Staghøj. 2009. "Using Statistical Treatment Rules for Assignment of Participants in Labor Market Programs." Unpublished manuscript. University of Michigan, Ann Arbor.
Smith, Jeffrey, and Petra Todd. 2005a. "Does Matching Overcome LaLonde's Critique of Nonexperimental Methods?" *Journal of Econometrics* 125(1–2): 305–353.
———. 2005b. "Rejoinder." *Journal of Econometrics* 125(1–2): 365–375.
Svarer, Michael. 2007. "The Effect of Sanctions on the Job Finding Rate: Evidence from Denmark." IZA Discussion Paper No. 3015. Bonn: IZA.
Todd, Petra, and Kenneth Wolpin. 2006. "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility." *American Economic Review* 96(5): 1384–1417.
Trenholm, Christopher, Barbara Devaney, Kenneth Fortson, Melissa Clark, Lisa Quay, and Justin Wheeler. 2008. "Impacts of Abstinence Education on Teen Sexual Activity, Risk of Pregnancy, and Risk of Sexually Transmitted Diseases." *Journal of Policy Analysis and Management* 27(2): 255–276.