C.D. Howe Institute

Commentary

www.cdhowe.org

No. 165, June 2002

Measuring the Performance of Government Training Programs

William P. Warburton
Rebecca N. Warburton

In this issue...

Studies have repeatedly found that most government job-training programs are ineffective. Yet there are examples of success, and the rewards from effective programs are potentially large. A key issue for policymakers is program evaluation: how to move beyond anecdotes and use solid evidence to avoid failure and replicate success.

The Study in Brief

Although studies have repeatedly found that most government job-training programs are ineffective, there are examples of success, and the rewards from effective programs are potentially large. A key issue for policymakers is program evaluation: how to move beyond anecdotes and use solid evidence to avoid failure and replicate success.

The first step in evaluating a training program is to define what outcomes will differ between those who went through the program (the "treatment group") and those who did not, and how. This task is harder than it seems, since people who enter and complete a program will likely differ from those who do not. A common source of error is confusing differences that are due to the program with those that existed beforehand between the two groups.

The ideal evaluation method from a theoretical point of view is to assign people into or out of the program at random. If the samples are properly chosen, large enough, and thoroughly monitored for impacts, random assignment will yield solid statistical information. But random assignment is sometimes ethically awkward or politically unpopular, may be more expensive and time consuming than less rigorous approaches, and is not widely used in job-training research, except in the United States.

Canada's performance in this area could be improved by using the federal-provincial division of powers to replicate the separation between the legislative and executive branches of government that seems to work in the United States. Ottawa could, for example, imitate the US Congress and require the provinces to use random assignment to test the results of federally funded programs.

Failing that, Canada needs better study of programs that do not use random assignment. Better observational studies could allow analysts to control for differences between the treatment group and others. Matching — trying to identify all key characteristics in which the two groups might differ and make only apples-to-apples comparisons — is one such method; regression analysis tries to do something similar. Higher-frequency (for example, monthly) data can reveal impacts (or lack of them) that lower frequency data may obscure. And observational studies need more individual information to help researchers detect and correct for differences between those who go through job-training programs and those who do not.

The Authors of This Issue

William P. Warburton is Adjunct Professor in the Department of Finance and Management Science, School of Business, University of Alberta. *Rebecca N. Warburton* is Assistant Professor at the School of Public Administration, University of Victoria.

* * * * * *

C.D. Howe Institute Commentary© is a periodic analysis of, and commentary on, current public policy issues. The manuscript was copy edited by Robyn Packard and prepared for publication by Marie Hubbs and Barry A. Norris. As with all Institute publications, the views expressed here are those of the authors, and do not necessarily reflect the opinions of the Institute's members or Board of Directors.

To order this publication, please contact: Renouf Publishing Co. Ltd., 5369 Canotek Rd., Unit 1, Ottawa K1J 9J3 (tel.: 613-745-2665; fax: 613-745-7660), Renouf's store at 71½ Sparks St., Ottawa (tel.: 613-238-8985), or the C.D. Howe Institute, 125 Adelaide St. E., Toronto M5C 1L7 (tel.: 416-865-1904; fax: 416-865-1866; e-mail: cdhowe@cdhowe.org).

$12.00; ISBN 0-88806-556-6
ISSN 0824-8001 (print); ISSN 1703-0765 (online)

James Heckman, 2000 Nobel laureate in economics, posed the question "How ineffective are current programs in moving people from welfare to work and in increasing their employment and earnings?" And he answered it: "Generally they are very ineffective" (1999, 29).

This is old news to many. On July 28, 1995, a headline in the *Globe and Mail* referring to the minister of Human Resources Development Canada (HRDC), proclaimed: "Axworthy loses faith in jobs training." And on April 6, 1996, *The Economist* reported, "Nobody seems to be saying that government-supported training is often a waste of money — nobody, that is, except researchers who have examined existing schemes" (p. 19).

A rigorous evaluation of US programs funded under the *Job Training Partnership Act* (JTPA), the main source of funding for training disadvantaged workers in the United States, found that the programs did not work. Participation in JTPA programs increased participants' incomes by less than US$5 per week[1] while making society poorer by more than US$200 per participant.[2]

Is the proper response to wipe out training programs, as recommended by, for example, David Hogberg (2001) of the American Public Interest Institute at Iowa Wesleyan College?[3] Not necessarily. The same studies that show that funding training does not *always* work also show that it does sometimes work and that it can be a very good investment for government. The impacts of JTPA-funded programs on various subgroups of participants and on various service strategies — for example, on-the-job training, classroom training, and job-search assistance — were measured separately in 16 sites. Although the studies found dismal results overall, they also found pockets of effectiveness. Some sites produced beneficial impacts while others did not; some service strategies produced bigger impacts than others; and some subgroups of participants benefited while others did not.[4]

Other programs that were also evaluated in different sites were shown to produce variable impacts. The US Manpower Development Research Corporation (MDRC), widely regarded as the preeminent random-assignment evaluator of welfare-to-work programs, conducted an extensive evaluation of California's Greater Avenues to Independence (GAIN) program. It found that, overall, GAIN increased costs to government, generating a return of US$0.76 for every US$1.00

The authors would like to thank Jeffrey Smith and Arthur Sweetman for very helpful comments on an earlier draft.

[1] Authors' calculation based on figures in Bloom et al. (1993, exhibit S1). Interestingly, the study does not indicate the overall disappointing impact: instead, it shows the impacts for subgroups selected so that the program is shown to be effective for at least one subgroup.

[2] Authors' calculation based on Bloom et al. (1997, tables 3 and 8). Benefits and costs were taken from their table 8, and sample sizes from table 3; the earnings gains reported in their table 3 were used in table 8. Note that all dollar amounts in this *Commentary* are in Canadian dollars unless otherwise specified.

[3] Certainly the Bush administration has some sympathy for this view. The Associated Press reported that Mitchell Daniels, President Bush's budget director, said that the government has too many job-training programs. See website: http://stacks.msnbc.com/news/664618.asp?cp1=1 accessed November 29, 2001.

[4] Bloom et al. (1993, 261) point out that they cannot reject the hypothesis that the differences in impacts across sites occur by chance. This is consistent with the position that the most likely explanation is that real differences exist across sites.

invested. But MDRC also reported that GAIN's Riverside County site increased participants' earnings by 49 percent while returning to government US$2.84 in increased tax revenue and decreased welfare payments for every $1.00 invested.[5] Positive results at some sites were swamped by negative results at others.

In addition to their impact on taxes, employment, and income, effective training programs can have important social benefits. MDRC reported the following effects of five welfare-to-work programs on children:

> All four programs that provided earnings supplements led to higher school achievement. Some of the programs also reduced behavior problems, increased positive social behavior, and/or improved children's overall health. (Morris et al. 2001, executive summary.)

If training programs not only increased government expenditure but also the incomes and well-being of participants and their families, we might well be willing to make the tradeoff. But the overall results from the JTPA and the GAIN studies indicate that, in most cases, training programs increase costs to government but have little effect on clients. And far from increasing well-being, training programs that do not work can exact a toll on participants, leaving them bitter and cynical after wasting their time.

Canada's national government training programs have never been subjected to rigorous evaluation.

We do not know if these US findings hold in Canada because our national government training programs have never been subjected to rigorous evaluation. But if the evidence from JTPA and other rigorous evaluations conducted in the United States holds for Canadian programs (and we have no proof that it does not), training programs do not improve Canadian society either. Despite this evidence, which probably contributed to the former minister's gloomy view of training, HRDC budgeted $2.8 billion on training in fiscal year 2002/03.

So is training a waste of money in Canada? Yes, overall, it probably is. But the lessons from the JTPA program's most effective site, "Site 4,"[6] and from Riverside County, California, show that it need not be. How can we pick out and expand Canada's Site 4s and Riversides, Canada's effective strategies and administrators, while weeding out poor programs and poor administrators? In short, how can we make government training programs work? That is the question we discuss in this *Commentary*.

In the next section, we explain the meaning of a program that "works," and we describe random assignment, the only method that estimates the impacts of a program — that is, tells us whether a program is working — without requiring us to make any additional assumptions.[7] We also go on to explain why random assignment cannot always be used and why random assignment does not produce all the answers.

[5] Riccio et al. (1994, tables 1 and 7 of the executive summary, which show results for single parents).

[6] The evaluation of JTPA programs does not identify the sites. Barnow (2000) identifies Site 4 as having the largest impacts.

[7] Random assignment is not a substitute for thought. For things that can go wrong in a random assignment study, see the section "Perceived Barriers to Using Random Assignment" later in this *Commentary*.

We then consider the political forces that work against the evaluation of Canadian government training programs and those in the United States that lead to exemplary estimates of program impact. All the evidence in the introduction is drawn from the United States, so the reader may wonder whether it applies in Canada. The sad truth is that almost no reliable evidence exists in this country, so we simply do not know. We also look at frequently used but ineffective methods of evaluating programs, and we suggest exploiting strengths in Canada's Constitution to achieve better results from our evaluation of training programs.

Then we provide a nuts-and-bolts strategy for improving our ability to measure the effectiveness of training programs over the next decade. We briefly review evidence that suggests that methods other than random assignment do not provide reliable evidence. We then describe a strategy for developing methods for reliably estimating the impacts of training programs so that we can make training programs effective.

In the final section, we discuss our conclusions.

Calculating the Impacts of Training Programs

Defining Program Impacts

A program works if it makes a difference for the participants, if it changes their lives in some way.

A program works if it makes a difference for the participants, if it changes their lives in some way. We often hear the claim, "We know our program is working. Eighty percent of participants got jobs." But measuring outcomes — what happens to people after participating in the program — does not tell us whether it changed their lives. This point was driven home to a British Columbia politician in the early 1980s. Her staff accurately reported that 80 percent of welfare recipients going through a program became independent within three months and stayed independent. Understandably, she bragged about the program's success in the legislature. But her bragging backfired. The Opposition suggested that she offer the program to all clients so that the ministry's budget could be reduced by 80 percent in the following year.

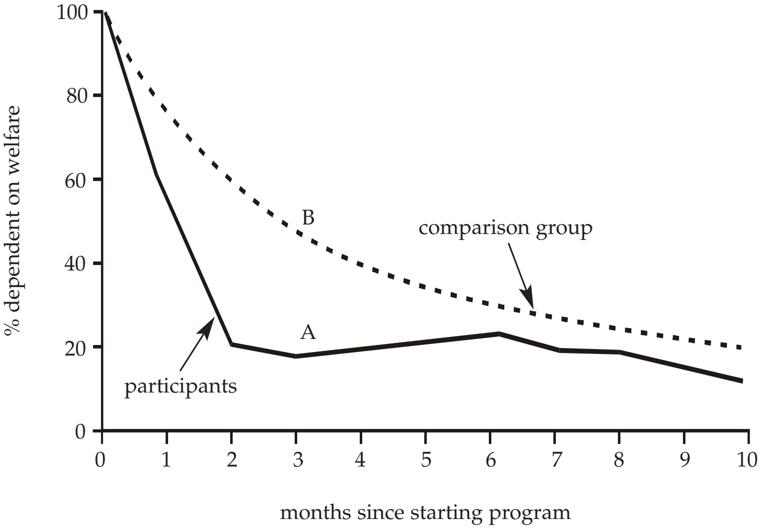
Her staff had gotten her into trouble by confusing outcomes with impacts. To find out if that program was working, she needed to know how it had changed the lives of the participants. She needed to know how many of the participants who became independent of welfare would have stayed on welfare without the program.

The impact of a program on subsequent welfare dependence is the difference between the number who become independent with the program and the number who would have become independent without the program.

This is illustrated by the figure on the next page. The solid line shows the subsequent welfare dependence of program participants. The dotted line shows the subsequent welfare dependence we expect these individuals would have experienced without the program. (Please suspend disbelief about the possibility of determining what the individuals would have experienced in the absence of the program until the next section.)

Putting the illustration into concrete terms, if 100 welfare recipients enroll in the program shown in the figure, three months later 20 of them remain dependent (point A). If they had not enrolled in the program, we would expect about 50 to

Welfare Dependency by Participants in a Hypothetical Training Program



Source: Authors' calculations.

remain dependent (point B). The difference, 30 people becoming independent, is the impact of the program. This means that the program has reduced the caseload by 30 (not 80) in three months.

The figure also shows that the impact of the program diminishes over time. Ten months after the start of the program, its impact is only 7 — that is, the caseload is only 7 people lower than it would have been in the absence of the program.

With an estimate of the program's impact, a politician can make statements about the benefits of the program in terms of, for example, reduced welfare caseloads and expenditure. The savings in month 3 equal the cost that the welfare agency would have borne if those 30 people had remained dependent. The total savings due to the program equal the sum of the savings in each month less the costs of operating the program. Note that a training program can save money even though the caseload is unchanged in the long run.

We could, in a similar way, calculate the impact of the program on incomes of the participants, success of their children in school, increased tax revenue, and so on. This is the type of information that ministers should bring both to the Treasury Board to secure funding and to citizens for political support for training.

The Difficulty of Estimating Impact

Estimating impacts is difficult because it is impossible to know what would have happened to any individual participant if he or she had not entered a program; yet this information is crucial for estimating impacts. Fortunately, the simplest of statistics, the average, enables us to estimate what would have happened to a group of individuals. The next four paragraphs explain how.

If we select a random sample from a population[8] and observe a characteristic — for example, height, age, or income — the average of this characteristic for the sample will be about the same as that for the population. The less the characteristic varies in the sample, the better will be the estimate of the average of the population. And the larger the sample, the more closely the average of the sample will correspond to the average of the population. Note that the statements were made (just as mathematicians made their proof of these statements) without regard to which characteristic we are averaging. The average height, weight, income, percentage with blue eyes, percentage wearing blue jeans, percentage married, or the percentage dependent on welfare will all be about the same for the group selected as for the

[8] We call the larger group from which the sample or subgroup is selected a "population" to avoid confusion in terms.

population from which they were selected *so long as the selection is random*. Random means that every member of the population must have an equal chance of being included in the sample.

This property of random samples enables us to estimate what would have happened to a group of program participants if they had not entered the program. It therefore meets the fundamental challenge of estimating impacts. Because the averages in each random sample will approximate the population average, all random samples will have similar average values for all characteristics.[9] If one of these groups participates in a program and the other does not, the averages for the nonparticipants will be a reliable estimate of what the averages for the participants would have been in the absence of the program. So the difference between the averages for the program participants (the *program group*) and the averages for the nonparticipants (the *control group*) indicates the impact of the program. The difference between the program group's average value of income and the control group's average value of income is an estimate of the impact of the program on income. The difference between the average value of social assistance received by the program group and the average value of social assistance received by the control group is an estimate of the impact of the program on social assistance receipt. Similarly, we can use the program and control groups to estimate the program's impact on any characteristic on which we have data, even if we had not thought of it before the program began.

Random assignment is the only method that can be proven mathematically to give unbiased estimates of the impact of programs.

This method of estimating impacts is usually referred to as *random assignment*. It is the only method that can be proven mathematically to give unbiased estimates of the impact of programs.[10] Random assignment has the additional advantage of being easy to understand. If participants in a program have been randomly assigned, estimating the program impact is as easy as calculating the average outcomes for the program and control groups and finding the difference between them.

How big must the selected groups or samples be? The answer depends on the variance of the characteristics in the population and on the degree of confidence required. We are all familiar with the property of averages as applied to poll results. For example, if we use a sample of 400 to estimate the proportion that will vote Liberal, then 19 times out of 20 the sample proportion will be within five percentage points of the true proportion of the population.[11] If we use a sample of 1,000, however, 19 times out of 20 the sample proportion will be within three percentage points of the true proportion.

Random assignment ensures that the control group has the same characteristics on average as the program group. If random assignment is not used, researchers must still estimate what would have happened if the program group had not

[9] We can make these averages as similar as we need simply by increasing the size of the group.

[10] The randomization does not have to be caused by the program. Two-stage techniques can take advantage of a random process that is unrelated to the exercise of estimating impacts. Nonetheless, estimates produced using two-stage techniques should be viewed with caution. See the discussion later in the paper under the heading "The Politics of Estimating Impacts."

[11] To make this calculation, we used the property that the variance of a proportion is $p(1 - p)/n$, where n is the size of the sample and p is the proportion. We use 50 percent for p in our calculation. The largest value for $p(1 - p)$ is when $p = 0.5$, so if the proportion is less than this, the confidence interval will be smaller.

participated in the program. To do this, they must find a comparison group of people who did not participate in the program, but who look as similar as possible to program participants in measured characteristics. (In this *Commentary*, we refer to nonrandomly selected control groups as "comparison groups.") Without random assignment, the estimate of program impact will be the difference between the outcomes of the participants and the outcomes of the comparison group.

Unfortunately, program participants, by virtue of their having chosen to participate, are more likely to have pre-existing characteristics, such as initiative and ambition, that are associated with success in the labor market. When program participants are allowed to volunteer, the difference in pre-existing characteristics together with participation in the program will improve the outcomes of the participants relative to the outcomes of the comparison group, resulting in an overestimate of program impact. The better the researcher controls for pre-existing differences, the lower the estimate of the impact of the program.[12]

Perceived Barriers to Using Random Assignment

Although random assignment is the only method that can be mathematically proven to give unbiased estimates of the impact of programs, using random assignment is not always possible. Researchers frequently encounter concerns about the feasibility, ethics, privacy, and cost of random assignment studies, and need to understand which objections are well-founded and which are not.

Researchers frequently encounter concerns about the feasibility, ethics, privacy, and cost of random assignment studies, and need to understand which objections are well-founded and which are not.

Feasibility

Using randomization to assess programs is usually feasible, since we can determine who is eligible for the program and randomly assign those eligible to enter or not enter the program. It is obviously not ethical or possible to assign some teenagers to complete Grade 12, and others to drop out after Grade 10. Yet we can and do assign people to stay-in-school programs, and when these are effective, the results can generate estimates of the impacts of life choices (see "Natural Experiments" later in this *Commentary* for a discussion of the limitation of random assignment studies that have incomplete randomization).

Ethical Issues

Where we have good reason to believe that a program is effective, and individuals have a legal or moral right to participate, random assignment is not appropriate; the reason is that in a random-assignment study some individuals who are otherwise eligible will not participate. This is not normally an issue in training programs, since usually there are many more individuals eligible for training than spaces available. The American National Research Council's Committee on Youth Employment Programs was asked to study the ethical issues in using random assignment to evaluate training programs for youth. The committee concluded, "[I]n situations in which program resources are scarce and program effectiveness

[12] The tendency of poorer quality studies to produce more optimistic results is documented in the health literature. See, for example, Sacks, Chalmers, and Smith (1982); and Schulz et al. (1995).

unproven, it [random assignment] is ethical" (Betsey, Hollister, and Papageorgiou 1985, 30).[13]

But even when random assignment is ethical, it may be politically unpopular. An individual excluded from a program is more likely to blame randomization than the limited budget. And even with a fixed budget, those who received no help would have preferred a different allocation mechanism — for example, that less expensive assistance be offered more widely.

Many believe that ethical issues are more pronounced in health care since the life-and-death results of decisions in medicine are often more apparent than decisions made for training programs.[14] Despite this, health care researchers use random assignment much more frequently than do researchers evaluating training programs. In most countries in the world, including Canada and the United States, new drugs and medical devices must be tested in randomized trials before they can be brought into general use. Because the outcomes (good health or death) are so important, it is important to know whether these drugs or devices work.[15] But this does not eliminate the political resistance to random assignment. Even where the drugs or devices are unproven, in some cases those allocated to the control group complain bitterly even when being allocated to the control group may result in longer life.[16]

Privacy

Estimating impacts of programs necessarily entails some privacy cost, but concerns about privacy are often exaggerated. With random assignment, information on outcomes for both participants and a control group is required. In the absence of random assignment, much more information is required since the researchers must select a comparison group that is as similar as possible to the participants. Estimating broader impacts (such as effects on health or children's schooling) requires linking databases, which affects the privacy of individuals as defined by the privacy commissioner. However, since linkage can be done with masked data without researchers ever knowing the names of participants, the actual risk to individual privacy is minimal.

Preventing access to data on privacy grounds can make the task of measuring the impacts of programs difficult, impossible, or simply much more expensive. Balancing privacy concerns and the gains from rigorous evaluation is essential, given the public interest in the effective use of public funds for programs intended to increase the employment and incomes of disadvantaged workers and improve the life chances of their children.

Preventing access to data on privacy grounds can make the task of measuring the impacts of programs difficult, impossible, or simply much more expensive.

[13] There is a long history of using chance to allocate resources. In the Old Testament of the Bible — for example, Num. 26:55 and Lev. 16:10 — goods were divided by lot.

[14] For discussion of this literature, see, for example, Altman and Bland (1999); Lumley and Bastian (1996).

[15] Perhaps the experience with thalidomide made the consequences of ignorance more visible in the world of medicine.

[16] For example, calcium channel blockers were, on the basis of animal studies, thought to prevent heart attacks, but clinical trials in humans found that they caused heart attacks in the treatment group. See the discussion in Sleight (1996).

Cost

High-quality random assignment studies generally take more time and money than studies of poorer quality, but the additional cost and time do not result from randomization. Overall, randomized studies probably cost less than nonrandomized studies. The Self Sufficiency Project (SSP) is Canada's best-known random-assignment study. It cost more than $50 million and will have taken more than ten years by the time the last report has been written. But the expensive features of the SSP were the extensive panel surveys, the payment system, and the analysis. The time taken to complete the study has resulted from extensive design work, pretesting the enrollment mechanism, spreading the intake evenly over a year to eliminate seasonal factors, looking at long-term impacts, and extensively cleaning and checking the survey and administrative data before they were turned over to other researchers for analysis.

Careful design, extensive analysis, and long-term follow-up take time and money whether random assignment is used or not. In fact, since program and control groups can be compared to determine program impact, randomization greatly simplifies data analysis, thus reducing analysis costs compared with those of nonrandom studies. Most important, the use of random assignment greatly increases the chances that the time and money spent on evaluation will produce reliable estimates of program impact. Without random assignment, there is a great danger of producing unconvincing results that require the study to be repeated, as Altman and Bland (1999) show.

Caveats in Interpreting Random Assignment Studies

Some care must be used in interpreting the results of random assignment studies. We list three areas for caution below. (For a more thorough discussion, see Heckman and Smith 1995; Burtless 1995.)

First, random assignment studies do not provide a complete answer. While the use of random assignment can produce reliable estimates of the impacts of programs on groups of individuals, it cannot provide information on how it affected people who were not part of the program. Sometimes programs can have larger effects on people who were not part of the program than on participants (Moffitt 1992). For example, if expanded funding were announced for welfare recipients to participate in postsecondary training, we would expect some people who formerly would have applied for student loans to now apply for welfare. This additional flow into welfare could exceed the impact of the program on the flow out of welfare. In this case, a random assignment study of postsecondary training for welfare recipients could find that it reduced dependence by participants, when in fact the program overall would increase welfare rolls by inducing more people to apply for welfare.

While the use of random assignment can produce reliable estimates of the impacts of programs on groups of individuals, it cannot provide information on how it affected people who were not part of the program.

Similarly, a random assignment study of a program that placed welfare recipients in private sector employment might find reduced welfare dependence among participants, even though the placed participants might have displaced other job seekers (who then became dependent on welfare). In this case, the net effect of the program could be zero; yet a random assignment study would show it to be effective. Program benefits can also be missed by the use of random assignment. If a community had a shortage of tugboat captains and a surplus of crew members, the training of a new captain could increase employment for crew members as well as providing employment for the person who took the training. This benefit would be missed if we looked only at the effect on the tugboat captains trained.

The second caution to be aware of in interpreting the results of random assignment programs is that their impacts may not be the same as those from similar programs that do not involve random assignment. This can occur for two main reasons. First, because random assignment affects the way in which service providers get their clients, they will be acutely aware that the impacts of their programs are being estimated. Consequently, service providers who believe that their programs are working well may be more likely to volunteer to participate in a random assignment study; once the program is running, they may exert additional effort knowing that they are being monitored. Second, the individuals who participate in random assignment studies may differ from those who participate in programs in which random assignment is not used. Generally, program administrators tend to select individuals they think will benefit most from the program, while random assignment brings in participants typical of the average member of the population. If administrators can accurately identify and select those individuals best able to benefit from the program, then random assignment studies, which measure the effect on average members of the population, will show lower program impacts than studies not using random assignment.

The third caution is that random assignment can be poorly implemented, creating systematic differences between the program and control groups, and biasing estimates of program impact.[17] The initial assignment might not be completely random, thereby leading to unintended systematic differences between those assigned to program and control groups. If study procedures are not monitored, people assigned to the participant group may not participate in the program, and people assigned to the control group may be allowed to participate. This will tend to reduce the estimated program impact.

Random assignment is often unpopular with program administrators because it explicitly interferes with or replaces their usual selection and referral process, which is based on staff and client judgment.

The importance of properly concealing the allocation of participants into the program or control group is well known in health care studies (Day and Altman 2000) but underestimated in studies of training programs. Random assignment is often unpopular with program administrators because it explicitly interferes with or replaces their usual selection and referral process, which is based on staff and client judgment. Staff may see judgment-based referral as an important part of a program. Where a program is new, and no referral process has been established, it may be easier to set up a random referral process than in an ongoing program where the referral process is well-established. Where randomization is used in an established program, staff may be highly motivated to intentionally subvert the randomization scheme — for instance, by discerning the random assignment scheme in order to help particular clients be assigned to the treatment group. In such cases it is essential to use a rigorous and nondiscernable method of randomization.

Allocation concealment is also important during program implementation and follow-up. In order to obtain unbiased data on immediate and follow-up outcomes,

[17] Unbiased impacts can be recovered from the data under plausible conditions. See Bloom (1984). See also the discussions in Heckman, Smith, and Taber (1998); and Heckman et al. (2000).

it is essential that those doing the outcome assessment and follow-up not know which subjects were program participants and which were in the control group. When program operators and evaluators know the allocation (to program or control group) of subjects, they may unintentionally bias their outcome assessment — usually toward positive outcomes for participants and negative outcomes for controls.

Losses to follow-up — the attrition over time in the subjects that evaluators are actually able to locate and contact in order to collect follow-up data — can also be biased when allocation concealment is lacking. Suppose that difficult-to-contact subjects tend to have poorer outcomes than more successful subjects (those who are dead or in prison, for instance, are harder to locate). When evaluators know which subjects were program participants and which were controls, they may be more diligent in following up the controls, which will result in finding (and recording outcomes for) a higher proportion of the unsuccessful controls than of the unsuccessful participants, again biasing results so that program effectiveness is overstated (Schulz et al. 1995).[18]

A final caution: the presence of random assignment does not ensure that follow-up survey data are collected reliably. As with any other study, participants who feel good about a program — perhaps because they benefited from it — may be more likely to provide information about their outcomes. If surveys are used to assess program outcomes, successful program participants are more likely to reply than unsuccessful ones; therefore, resulting estimates of program impact will be biased upward.

The Politics of Estimating Impacts

Random assignment is not routinely used in Canada to estimate the impacts of training programs for disadvantaged people. As discussed earlier, this is not for practical, ethical, privacy, methodological, or financial reasons. Some administrators object, but this hardly seems to be a sufficient reason for not using random assignment. Does the answer lie with politics?

The reason for the paucity of random assignment studies of training programs lies not with the politics of poverty, but with the politics of accountability.

The "politics of poverty" might be blamed for the paucity of random assignment studies of training programs; perhaps because such programs serve disadvantaged people, it is regarded as unimportant to know whether the programs work. In our opinion, however, the reason lies not with the politics of poverty, but with the politics of accountability.

The basic politics of training programs are straightforward. Three groups of people are affected by training programs: taxpayers, service providers, and clients. The political force of taxpayers manifests itself in demands for "accountability." In Canada, about 1½ percent of federal government expenditure is spent on training, so those concerned with taxation are not likely to devote much time specifically to training programs.[19] Instead political forces are likely to be expressed through a general requirement for accountability, diffused over all programs run by government.

[18] On average, randomized trials that have not used appropriate levels of blinding show larger treatment effects than properly blinded studies.

[19] Total government expenditure in fiscal year 2002/03 is budgeted at almost $170.4 billion; expenditure on training is budgeted at $2.8 billion (from website www.tbs-sct.gc.ca/tb/estimate/20022003/002_e.html).

We would expect service providers to advocate the training programs they provide out of pure self-interest. But even without this financial incentive, service providers are unlikely to be able to provide objective evidence. Providers routinely see the good outcomes from their programs (recall the comment, "Eighty percent of our clients get jobs!") and cannot see what happens to comparable people who do not receive training.

Finally, we would expect clients to advocate programs for three reasons: first, they appreciate the effort of the service providers; second, they do not know what would have happened to them in the absence of the programs and so may ascribe good outcomes to the program; and third, in some cases, they receive money for participating in the program.

So, as Machiavelli warned in *The Prince*, politicians should be wary of changing training programs, "because the innovator has for enemies all those who have done well under the old conditions, and lukewarm defenders in those who may do well under the new." Given these political forces, it is perhaps surprising that there are any reliable random assignment estimates of training program impacts.

The political forces against reliably estimating the impacts of training programs seem to triumph in every country except the United States.

The political forces against reliably estimating the impacts of training programs seem to triumph in every country except the United States. Beatrice Reubens (1980, 132) concluded, "[C]ompared to the American practice, the evaluation of youth programs in most other countries has been infrequent, unsystematic, and often methodologically dubious." As another example, in 1986, the Ontario government commissioned a review of the literature relating to employment and training programs. The study found that "[i]n spite of the fairly extensive list of Canadian institutions and experts contacted during the information gathering process, it became evident that most, if not all, of the information pertains to the US" (Social Program Evaluation Group 1986, 3). This observation still remains true. An ironic illustration comes from the literature related to programs for promoting self-employment. Two primary methods are used to promote self-employment, the British model and the French model, so named because of their widespread use in those two countries over the past 20 years. But to find reliable estimates of their impacts, we have to go to the United States, where they have been tested using random assignment in two different sites.

Evaluating Programs in the United States

With the separation of powers in the United States, the executive branch runs the programs, but the legislative branch provides the money and demands the accountability. In the case of the *Comprehensive Employment and Training Act* (CETA) programs (the predecessor to the JTPA), Congress did not demand that the impacts be estimated using random assignment. Data on the programs were collected and the Department of Labor (part of the executive branch) commissioned studies of the impacts (Bryant and Rupp 1987). But in 1980, Ronald Reagan, a Republican, was elected president while the Democrats controlled Congress. So, in addition to the effect of separation of powers (Congress embarrasses the president, not itself, if it finds a program to be ineffective), Congress had a political incentive to show that a program run by the other party was ineffective. Perhaps in response to this additional political incentive, the Congressional Budget Office no longer merely required the Department of Labor to evaluate its programs but commissioned a study of its own (Barnow 1987, 158). This led to substantial discussion, which was finally resolved by referring the matter to a blue ribbon panel, which recommended, first, that future programs be evaluated using random assignment; and second, that research into sources of bias be conducted so that methods other than random assignment could be used to provide reliable estimates in the future (Stromsdorfer et al. 1985).

Congress thus built into JTPA the requirement that the Department of Labor evaluate JTPA programs using random assignment. In addition, Congress funded the Center for Social Program Evaluation at the University of Chicago to investigate sources of bias.

This separation of powers does not inevitably lead to reliable estimates of program impact. It may be coincidental that, when the executive and legislative branches of government were controlled by the same party for many years, the legislative branch did not require the executive branch to evaluate its programs using random assignment. Nonetheless, contending political forces in the United States have generated the vast majority of the world's random assignment estimates of program impact, in contrast to parliamentary systems, which have generated almost none.[20]

Evaluating Programs in Canada

In parliamentary systems, both the agency that provides the money and the agency that runs the program are overseen by cabinet and the prime minister or premier, thereby severely reducing the political incentive to estimate the impacts of a program. The program managers will, on the basis of anecdotal evidence, sing its praises. Current and subsequent ministers, acting on this information, may claim ownership of the program. It takes four or five years to produce rigorous estimates of program impact, and by that time the government will have significant backtracking to do if it finds that the program does not achieve its intended objectives. Consequently, finding a low or negative program impact does not have the political payoff in Canada that it does in the United States. The best the politician who commissions the study can hope for is that it confirms the claims that have been made all along. A much more likely outcome is that the Opposition is handed information that allows it to assert that the government has been misleading the public and/or is incompetent.

The best the politician who commissions a program evaluation study can hope for is that it confirms the claims that have been made all along.

Perhaps because of these disincentives, Canada has no generally accepted standards for assessing program impacts. In the absence of these standards, the Opposition or the auditor general can declare that a program is ineffective or that an evaluation is inadequate, but program advocates will reply with anecdotes about success stories or glowing survey results. The public may be interested in good use of public money, but can hardly be expected to be interested in experts bickering over abstruse evaluation methods.

[20] Canada distinguishes itself with the Self Sufficiency Project and the Earnings Supplement Project, which were evaluated with random assignment.

Conducting Unreliable Evaluations

Program managers can often evade the diffused pressure from taxpayers for accountability, avoiding evaluation of any kind, simply by arguing against random assignment studies on ethical, privacy, or financial grounds. These objections are not well-founded, but nonetheless often succeed. Where evaluation cannot be avoided, we often see evaluations that appear to provide accountability but do not provide reliable estimates of program impacts.

As we saw above, unreliable estimates are more likely to overstate than understate the impacts of programs. Absent clear standards or strong political pressure for rigorous evaluation, political forces actually encourage inaccurate — hence favorable — program evaluations. Let us look briefly at the pitfalls of evaluations that do not accurately estimate program impacts.

Absent clear standards or strong political pressure for rigorous evaluation, political forces actually encourage inaccurate — hence favorable — program evaluations.

Client-Centered Evaluation

A client-centered evaluation is intended to meet the needs of clients and program managers. Surely that is a worthwhile evaluation goal — or is it? The difference between approval of a program in a client-centered evaluation conducted by experts and the popular understanding of the word *evaluation* is illustrated by an evaluation of an experiential education program. The program was evaluated by Michael Q. Patton, former president of the American Evaluation Society, who teaches evaluation and is the author of many books on how to conduct client-centered evaluations. Patton counsels evaluators to ask program administrators what information is needed and to collect that information in an evaluation. But estimating impacts is of limited use to program administrators for three reasons. First, they are rarely in a position to make decisions about program funding, which require information on program impacts. Second, they may face losing their programs — and their own job — if their programs are found to have lower-than-expected impacts.[21] Third, most program administrators believe they already "know" that their programs are effective from personal experience — they have seen the successes.[22]

For all these reasons, client-centered evaluations tend to focus on process rather than impacts. Impacts may not be completely ignored, but the focus of the evaluation shifts from impacts to process. With few resources devoted to assessing impacts, the reported impacts will probably be higher than if more rigorous methods were used. And the report can still be thick, full of abstruse jargon and backed by impressive credentials.

In the case of the experiential education program, Patton rounds out his client-centered evaluation by concluding, on the basis of the assertions of participants and others associated with the program, that the program was effective. He notes that "many of those administrators returned to their colleges to spearhead curriculum

[21] Even when a program administrator believes a program is ineffective, he or she is more likely to support fixing it than scrapping it. Estimates of impact are not necessary to fix a program.

[22] Recall that programs that make no impact can still have many success stories — that is, they can have many participants who experience good outcomes. Participants themselves may attribute their good outcomes to the program, even in cases where a random assignment study finds that members of a control group have just as many positive outcomes.

reform" (1997, 66). The evaluation contained no consideration of what would have happened in the absence of the intervention.

There is no question that such an evaluation would be useful to a program administrator. But former US Senator William Proxmire, looking at the same program from the perspective of a funder, was not convinced by the assertions of effectiveness and gave it a Golden Fleece Award for being a flagrant waste of taxpayers' money.[23]

The use of evaluation methods that inadequately measure impacts has become institutionalized in Canada. The 12 attributes of effectiveness of the Canadian Comprehensive Auditing Foundation (CCAF) is the bible for many evaluators in Canada (1987). But of the 12 attributes, only two — Number 4: Achievement of Intended Results, and Number 6: Secondary Impacts — deal with program impacts; moreover, the attributes do not specify the need for rigorous measurement. An evaluator following the CCAF's guide could consider more than 80 percent of the attributes and not discuss program impacts at all. An evaluator could even consider every attribute but still include only nonrigorous measures of program impact.

Harnessing Canadian Political Forces to Improve Evaluations

As discussed earlier, the separation of political responsibility between funding programs and delivering them has led to exemplary estimates of program impacts in the United States. Canada, with its federal-provincial Constitution, is also blessed (some might say cursed) with a separation of powers that makes it possible to separate the political forces associated with giving the money (and demanding accountability) from the political forces associated with running programs. The feasibility of such a separation is being tested by the Labour Market Development Agreements (LMDA), under which the federal government provides the money and provincial governments provide services. Transfer of responsibility only to some provinces (only Quebec and Alberta have full devolution) limits this ability to separate political forces since, to date, accountability mechanisms remain similar for fully devolved and co-managed provinces.

The separation of powers provides an opportunity for the federal government to require, by law, that the provinces estimate the impacts of their programs using random assignment.

The separation of powers provides an opportunity for the federal government to require, by law, that the provinces estimate the impacts of their programs using random assignment, just as the US Congress has a history of legislating that the executive estimate the impacts of their programs using random assignment.[24]

Improving the Measuring of Impacts

Even if we can develop the political will to estimate program impacts, we cannot always use random assignment, so we will still face the question of how to produce

[23] The Golden Fleece Award, instituted by Senator Proxmire in 1975, singles out US federal programs that most Americans agreed were outrageous and wasteful (see website: www.taxpayer.net/awards/goldenfleece/about.htm#summary, accessed April 3, 2002).

[24] Section 452(d)(1)(a) of the JTPA states: "Evaluations conducted under paragraph (1) shall utilize sound statistical methods and techniques for the behavioral and social sciences, including random assignment methodologies if feasible."

reliable estimates of program impacts in the absence of random assignment. That question is considered in this section, which has four parts. The first describes techniques that can be used to estimate the impacts of programs in the absence of random assignment. The second describes some techniques that have been proposed as substitutes for random assignment, but which have been shown not to work. The third part reviews selected papers that compare the results from random assignment studies with estimates made using observational techniques. The final section describes the conditions that researchers have discovered are necessary (although not sufficient) for producing unbiased estimates of program impact.

Alternatives to Random Assignment

A study that does not use random assignment is referred to as an *observational study*. In such a study, service providers and clients determine who participates in a program. Researchers must select as a comparison group nonparticipants who look most like program participants and use statistical techniques to control for any observed pre-existing differences between the treatment and comparison groups. The difference in outcomes between the program and comparison groups, after controlling for the effects of as many pre-existing differences between them as possible, provides an estimate of program impact.

Observational studies vary in complexity and in the techniques used to conduct them. They may be done well or badly, and techniques used for them may be appropriate or inappropriate, but the accuracy of the studies will suffer whenever the comparison group differs from the participants in a way that has not been noticed or measured.[25] Unfortunately, unobserved pre-existing differences between participants and nonparticipants are very common. For instance, individuals who are more highly motivated are generally felt to be more likely to enter training programs and more likely to become employed. If, after receiving training, participants do better than nonparticipants, we will be left wondering whether participants did better because of the training or because they were more highly motivated to begin with. Estimates that are wrong because they falsely attribute the impacts of characteristics of participants to programs suffer from *selection bias*.

Estimates that are wrong because they falsely attribute the impacts of characteristics of participants to programs suffer from selection bias.

Rigorously implemented randomization with good blinding is the only way to prevent selection bias. Within studies that have randomly selected participants, gradations of reliability often exist because of variations in the rigor with which randomization has been implemented.[26] In some cases, all the individuals selected to participate (the treatment group) will participate and none of those not selected

[25] Some of these differences may be readily apparent to program administrators, even if they are not quantified and known to the researcher, they will still cause bias. For example, individuals who are incapacitated in some way — for example, those with a serious physical illness — are less likely to take training and also less likely to move into employment on their own. Differences in subsequent employment between program participants and a comparison group of nonparticipants might result from higher rates of illness in the comparison group rather than from the program itself. Researchers rarely have medical information on program participants; yet, without it, estimates will be biased.

[26] The best studies are properly blinded in both treatment and follow-up, have complete reporting, ensure proper randomization, and do not have dropouts or crossovers.

(the control group) will do so. In such rigorous random assignment studies, the difference in the outcomes of the treatment and control groups equal the average impact of the program.

In other cases, randomization is incomplete, the proportion of the treatment group that participates is less than 100 percent, and the proportion of the control group that participates is greater than zero. When that happens, participant and control groups both include participants and nonparticipants. Under plausible assumptions, we can still estimate program impact by dividing the difference in average outcomes for the treatment and control groups by the difference in percentage treated (Bloom 1984).

It is unlikely, however, that a program has the same effect on everyone. Instead some people probably benefit more than others from some programs. The difference in the average outcomes of the program and control groups is an estimate of the average impact across all members of the program group. When randomization is incomplete, the difference in the average outcomes of the program and control groups divided by the difference in percentage participating is an estimate of the average affect on the additional members of the treatment group receiving the treatment.

Natural Experiments

In some cases, researchers can take advantage of *natural experiments* — events that affect the likelihood of receiving some treatment but are not correlated with unobserved personal characteristics that might influence the effects of the treatment. For instance, in the United States during the Vietnam War, birthdays were selected randomly to determine who would be drafted. Because (at least initially) men could avoid the draft by continuing their education, men with "high-probability-of-draft" birthdays stayed in school longer on average. Angrist and Krueger (1992) used the random selection of birthdays for the draft (which affected some men's likelihood of staying in school) to estimate the impact of education on income.[27]

Natural experiments can be viewed as random assignment studies with incomplete randomization.

Natural experiments can be viewed as random assignment studies with incomplete randomization: estimates of impact are recovered by dividing the difference in average outcomes of program and control groups by the difference in participation; and the impact generated is the average impact for the additional people affected. Unfortunately, many natural experiments create only small differences in the level of treatment received by the program and control groups. If the difference in proportion treated is only 1 percent, the program and control group outcomes will differ by only 1 percent of the average program impact. This affects the reliability of the estimates.

[27] Birthdays for people affected by the Vietnam War–era draft provide an excellent example of a natural experiment because the birthdays were randomly selected for the draft, so the instrument was clearly not correlated with the unmeasured characteristics of the people affected by the treatment. However, the effect of the draft was to increase schooling by only about 5 percent, so the estimated impact was not applicable generally. In addition, the treatment caused more than one outcome: for some, having a high-draft birthday meant staying in school to avoid going to war; for others, it meant going to war. Techniques that are not based on random assignment must be used to sort out these two impacts.

Recall that the average values of the program and control groups will be similar but not identical. Some variables may be correlated with both the treatment and the outcome of interest to researchers, and they will create small differences in the outcomes of the program and control groups in addition to the difference caused by the treatment. While the effect of the treatment on the difference gets smaller as the percentage affected by the treatment gets smaller, the differences due to chance will be constant. For this reason, program impacts get harder to detect the smaller the percentage treated.

We should note two cautions relating to natural experiments. First, without careful statistical analysis, researchers can be fooled when the percentages treated are small (Bound, Jaeger, and Baker 1995). Second, as with imperfectly constructed experiments, the estimates produced do not correspond to the impact of eliminating the program but rather to the impact of increasing or decreasing the level of treatment slightly. The difference in outcomes between the program and control groups is caused by the higher treatment percentage in the program group. The extra participants often differ from other participants, however, so that the impacts — which result only from the additional participants — measured by the natural experiment may be substantially different from the overall impacts (for all participants).

Natural experiments are not a reliable or complete alternative to random assignment because of the rarity with which they occur.

Natural experiments provide an opportunity to produce unbiased estimates of the impact of an intervention. However, they are not a reliable or complete alternative to random assignment because of the rarity with which they occur.

Matching

An experimental control group is by design, on average, similar to the program group in all characteristics. If we do not have a control group, we can draw a comparison group that is similar to participants in all observed characteristics. Although researchers have proposed a number of methods of selecting comparison groups over the years (see, for example, Dickinson, Johnson, and West 1986), most currently use the propensity score method developed by Rosenbaum and Rubin in 1983. Those researchers showed that a comparison group constructed by selecting, for each participant, a comparison group member with the closest probability of program participation (“propensity score”) will give consistent[28] estimates of program impact when all the factors that affect both program participation and the outcome of interest are known and measured. Dehejia and Wahba (1999) and Smith and Todd (2001) use this method.

Nearest-neighbor matching is the term used when a comparison group member is selected for each participant. A disadvantage of nearest-neighbor matching is that it throws out some information. As noted earlier, the precision of an estimate increases with the number of units in the sample; but nearest-neighbor estimates limit the sample size of the comparison group to the size of the program group. Selecting more than one nearest neighbor can ameliorate this. Heckman, Ichimura, and Todd

[28] In general sample size does not affect bias. However, some estimators can only be shown to accurately reflect the true impact when the sample size approaches infinity. Such estimators are said to be consistent. For the definition of consistent estimator, see website: www.statistics.com/content/glossary/c/consistest.html.

(1997; 1998) give a method for using all members of the potential comparison group, weighting the observations in proportion to how well they match the participants.

Regression

Regression analysis is another statistical technique for adjusting for measured differences between program and comparison groups and under certain circumstances can be shown to produce identical results to matching.[29] To use regression analysis, the researcher must first posit a mathematical relationship (generally referred to as a functional form) between the program and comparison groups' characteristics and the outcome of interest. For example, a researcher might posit that

$$\text{Income} = a_0 + a_1 \times \text{last year's income} + a_2 \times \text{age} + a_3 \times \text{program participation} + a_4 \times \text{education} + a_5 \times \text{age of youngest child} + e \text{ (random factors that we cannot measure).}$$

Then the researcher will use regression analysis to find the values for $a_0, a_1,\ldots, a_5$ that best predict income. Although this is a fairly typical functional form, it immediately raises questions, some serious, some not serious. In this specification, the researcher has put in a variable "program participation." This variable will take the value 1 if the individual participated in the program, 0 otherwise. In other words, it says that program participation increases income by an amount a_3. This formulation constrains the effect to be the same on all individuals, something that we would not expect to be true *a priori*. It turns out that the best predictor of the effect of a program on individual income when the actual impacts vary from individual to individual will be the average impact. So even though the functional form is incorrect, regression analysis will provide a useful result. As noted above, while the average impact is a useful number for calculating the impact of canceling the program altogether, it may not be a very good guide to the impact of increasing or decreasing the size of the program.

The best predictor of the effect of a program on individual income when the actual impacts vary from individual to individual will be the average impact.

A more serious problem occurs when some factor not included in the regression analysis affects both program participation and the outcome of interest. For example, enthusiastic people may be more likely to both participate in a training program and have higher incomes in the absence of the program. Failure to control for "enthusiasm" in either regression analysis or matching will result in some of the effect of enthusiasm on income being falsely ascribed to the program. For both regression and matching, collecting and controlling for more information on the participants and the potential comparison group is the only way to reduce selection bias due to unobserved characteristics.

Regression Plus Matching

Regression analysis can be shown to give the most accurate estimates of program impact when the functional form is correctly specified; it controls for even small

[29] Such circumstances require the functional form used in the regression analysis to be correct and the average values for each characteristic to be the same for the treatment as for the control group.

differences in the measured characteristics of the individuals. Matching requires no assumptions about functional form, but does not control for the small differences between members of the participant and comparison groups that typically exist after matching. Both regression and matching will be biased if unobserved characteristics affect both the likelihood of program participation and the outcome of interest. So when researchers choose between regression and matching, it seems that they should use matching when concerned about functional form, and regression when concerned that the matching may be incomplete.

Fortunately, it is not an either/or choice. A researcher can draw a matched comparison group and then run a regression to control for remaining differences. Rubin (1979) found that regression plus matching worked better than either on its own. When the participants are very similar to the comparison group in their measured characteristics, as is the case for a matched comparison group, misspecification of the functional form is not serious. And regression will control for any remaining difference between the program and the comparison groups. So the combination of regression and matching yields benefits from both approaches.

Two Contenders Not Recommended

The Heckman Two-Step Method

James Heckman (1976) showed that, under special circumstances, the impacts of interventions could be estimated without specific knowledge of the factors that affected program participation.[30] Unfortunately, the method can generate unpredictable results when the special circumstances do not hold; and it is impossible to test whether the appropriate circumstances do hold (Goldberger 1983). Knowledgeable researchers such as Heckman no longer use this method, except in conjunction with a natural experiment. But because of its apparently miraculous properties, it is still in use today.

Difference in Differences

The method known as difference in differences *is based on the premise that the characteristics that differ between participants and nonparticipants are fixed.*

Another method, known as *difference in differences*, which is appealing in its simplicity, is based on the premise that the characteristics that differ between participants and nonparticipants are fixed. If that were the case, we could estimate the impact of a program without using random assignment. Suppose we were interested in the impact of a program on income. An individual's characteristics (including the characteristics of the labor market that apply to that individual) would determine his or her income. After taking training, the individual's income will be affected by the same characteristics plus training. In this special case, the change in income from before to after training is the impact of training. (Simple before-and-after studies rely on this assumption.) We could relax this assumption somewhat by allowing changes in the labor market to affect income as well, although the effect must be the same for participants and nonparticipants. Then changes in the income of the nonparticipants serve as a measure for the effect of changes in the labor market.

[30] The special circumstances required jointly normally distributed error terms.

The effect of training becomes the difference between the change in income of the participants and the change in income of the nonparticipants.

Although appealing, the basic premise of this method — that differences between the participants and nonparticipants are fixed — has been shown to be false. Training participants are more likely than others to have suffered a preprogram dip in earnings, and random assignment studies have shown that even without any training, this earnings dip is transitory.[31] The difference between incomes of training participants and other workers therefore increases before training (the preprogram dip) and then decreases later *with or without training*. These changes over time in the difference between the incomes of program participants and the incomes of others will seriously bias difference-in-differences estimates.

Comparing the Results of Observational Studies and Random Assignment Studies

In theory, we can control for all pre-existing characteristics, and therefore random assignment should not be necessary.

In theory, we can control for all pre-existing characteristics, and therefore random assignment should not be necessary. In practice, however, observational studies and random assignment studies rarely generate the same estimates of program impact.

Early studies (Fraker and Maynard 1987; LaLonde 1986) compare estimates of the impact of the National Supported Work Demonstration Project using, on the one hand, the control group and, on the other, observational techniques and a variety of comparison groups. Their conclusions are not favorable for observational techniques. LaLonde, for example, concludes, "even when the econometric estimates pass conventional specification tests, they still fail to replicate the experimentally determined results" (p. 617).

Friedlander and Robins (1994) compare estimates of four random assignment studies of training programs for welfare recipients with estimates made using comparison groups drawn from the same jurisdiction, before the introduction of the program, and estimates made using comparison groups from other jurisdictions. Their analysis is hampered by a lack of explanatory variables — for instance, they had only one year of preprogram earnings. They conclude that "statistical matching or a specification test alone will be unable to markedly reduce the uncertainty surrounding that kind of non-experimental estimate" (p. 18).

Cain et al. (1993) have the most optimistic finding. They compare observational and random assignment estimates of the impact of the Homemaker-Home Health Aide Demonstration and find that their observational study produced results that are very similar to those from the random assignment study. Their study is remarkable in three ways. First, they draw their potential comparison group from the dropouts and rejectees, people who were clearly eligible for and interested in the program. Second, they have extensive — five-year — histories of employment from tax records for all participants and potential comparison group members. And third, the program

[31] Heckman, LaLonde, and Smith (1999, fig. 1) show that program participants experience, on average, a dip in their earnings because some of them lose their jobs before they enroll in training. Average earnings of program participants decline at about 10 percent per month going into the dip, but the experience of the control group shows that the dip is transitory. Without help from any program, average earnings of the control group doubles in some months as they come out of the dip.

had substantial impacts. They conclude, "[T]his paper offers a number of findings that lend some optimism to the search for valid nonexperimental methods" (p. 27).

Heckman et al. (1998) compare observational and random assignment estimates of the impact of the programs funded under the JTPA. At first blush, it seems that their conclusions are pessimistic. They report bias of at least 83 percent of program impact in their observational estimates. As noted earlier, however, the impacts of JTPA-funded programs were very small — about $5 per week. For men, the subset analyzed in Heckman et al. (ibid.), the impacts were even smaller — about $4 per week. So, although in percentage terms the bias is large, in absolute value it is quite small. If bias is not proportional to impact and if we can use the results (in Heckman et al.) as a guide, bias would be small in proportion to a program that increased earnings by $50 per week.

The most recent optimistic results by Dehejia and Wahba (1999) have been shown by Smith and Todd (2001) not to be robust.

Lessons Learned

The simple and obvious conclusion is that, when measuring the impact of programs, we cannot rely on estimates that have not been made using random assignment. We also know that, in one case at least, an observational study was able to reproduce the random assignment results. Before we can expect researchers to have confidence in observational estimates, we will have to replicate random assignment results many times. The differences among the studies that have attempted to do this in the past give us clues as to what may be successful in the future.

Before we can expect researchers to have confidence in observational estimates, we will have to replicate random assignment results many times.

In comparing Heckman et al. (1998) and Cain et al. (1993), we can see that the former may have been hampered by the source of their potential comparison group. Their comparison group comprises people who lived in the same neighborhoods as the participants and who were found, in a screening interview, to be eligible for the program. The result is a group of nonparticipants who are not closely comparable to the participants (see Heckman et al. 1998, fig. 2). In addition, that study uses a relatively short history of employment and earnings that was gathered after the fact and so may have errors due to faulty recollection.

By contrast, Cain et al. (1993) use administrative data that was extensive and free from recall problems. They also draw their comparison groups from people who were at least part way through the selection process.

Heckman, LaLonde, and Smith (1999) identify three additional conditions whose absence led to unpredictability in the observational studies that earlier researchers had compared with random assignment studies. First, the comparison group must be drawn from the same labor market as the program participants. Second, the same data source must be used for both program participants and the comparison group. Specifically, if the data are collected by survey, the same survey questionnaire must be used for both. Third, the analysis must be restricted to participants for whom there are comparable nonparticipants.

Another lesson learned from attempts to estimate impacts using observational methods is that monthly, not annual, data must be used. Data that are too coarse and that can mask the pre-program dip will lead to bias. In observational studies, the individuals selected for comparison groups match the participants in all characteristics, including earnings and employment, in the period just before entry into the program. With annual data, the period "just before entry into the program" will be the previous year, up to 11 months before program participation and, for many, before they lost their jobs. So if we have only annual data, participants who were employed in the year before they entered the program are compared with nonparticipants who were also employed in the previous year. But because programs are for the unemployed, participants will differ from nonparticipants in that the former are unemployed; this difference will lead to substantial bias (Warburton 1996a).

One final important source of bias is nonresponse in the surveys that collect information on outcomes of the participants and the comparison group. Typically, the interviewer fails to contact some people, while others refuse to answer the questions. If respondents from the program group are different in some unobserved way from respondents from the comparison group, we may falsely attribute the impact of these differences to the program. This problem is exactly analogous to the problem of selection bias, where researchers do not know the circumstances under which a nonresponse bias would be serious. In 1979, the US Office of Management and Budget proposed a standard requiring that the survey research firm contact 75 percent of the treatment and control groups (Smith 1999, 14). Because the reasons for nonresponse vary, this percentage might be unnecessarily high in some cases and not high enough in others.[31] Warburton (1996b) finds that a survey with a 75 percent response rate generated positive and statistically significant estimates of program impact, whereas full information (based on monthly administrative data) indicated that the programs actually had no impact. This discrepancy indicates that even 75 percent might not be high enough for evaluations of employment and training programs. Therefore, response rates should always be reported in evaluations, and serious efforts should be made to improve response rates and/or identify the circumstances in which nonresponse is not important.

Response rates should always be reported in evaluations, and serious efforts should be made to improve response rates and/or identify the circumstances in which nonresponse is not important.

Conclusions

Our examination in this *Commentary* leads to five conclusions. The primary conclusion is that random assignment is the only method certain to produce reliable estimates of the impacts of employment and training programs. So, to have confidence in estimates of the impacts of training programs, we must use random assignment.

Second, the federal-provincial separation of powers in Canada could provide the political force that leads to accurate estimates of the impacts of training programs. To paraphrase David Hume (in *A Treatise of Human Nature*, 1740), Reason is and ought only to be the slave of politics, and can never pretend to any other office than to serve and obey them. If there is no political will to estimate the impact of training programs, then reason, by itself, will not cause them to be evaluated.

Third, society needs standards for the evaluation of training programs in the absence of random assignment. A first cut at a flowchart for evaluation studies is

[31] If people do not respond for reasons related to the program, 75 percent might not be high enough. For example, people who believed that the program helped them might be more willing to take the time to respond to a survey, or people who are working might be more difficult to contact. If nonresponse is random, then the response rate will not matter much.

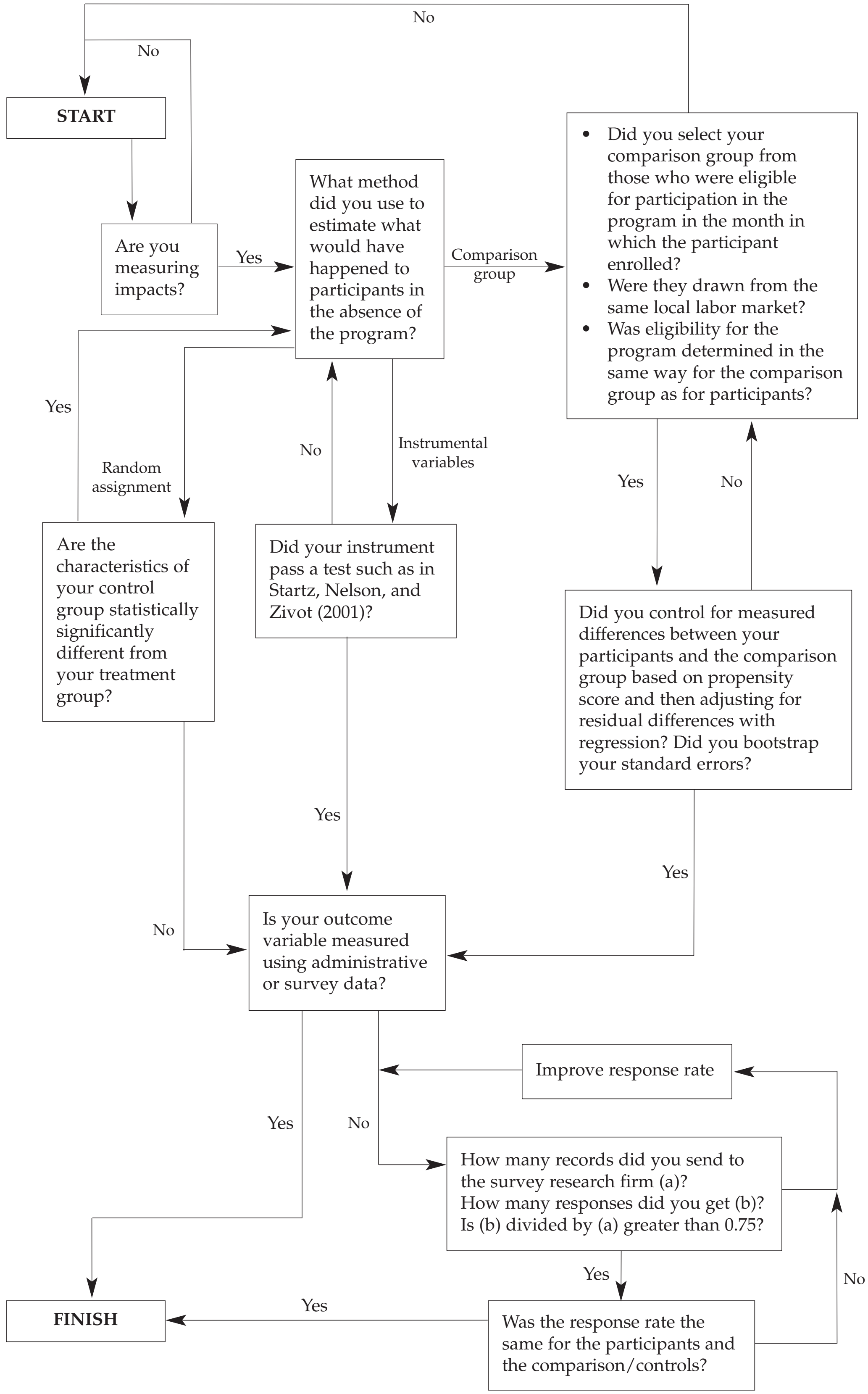
included as an appendix to this *Commentary*. The flowchart needs improvement, but that can only be done by conducting random assignment studies, then comparing the results with those produced by observational studies, identifying the sources of differences, and comparing again. Each initial attempt to replicate the results of a random assignment study should be made without the researcher's knowing the results of the random assignment study so that there is no possibility of trying numerous techniques until, perhaps by chance, one works.

Fourth, we need to develop standards for response rates to surveys. To do this, we must compare the results produced using administrative data with the results of surveys that for different reasons, yield various response rates; in doing this, we can identify the circumstances under which these data sources produce comparable results.

Finally, methods that do not use random assignment clearly require identifiable personal information. Even though the data can be masked and need never be made public, the use of personal information entails a small privacy risk. To preserve research uses of personal data, it is essential that researchers comply with privacy safeguards that ensure that data approved for research use are not also used for unauthorized administrative purposes such as audit or enforcement. Canada must balance these small risks to privacy against the critical need for reliable information on the impacts of training and other social programs. If knowledge of the true impacts of programs leads to improved training programs, evidence from the limited number of existing random assignment studies suggests that the payoff could be reduced poverty, reduced taxes, and improved health and well-being of our children. The alternative is to continue as we are, spending billions of dollars per year on training programs that the preponderance of evidence suggests simply do not work.

If knowledge of the true impacts of programs leads to improved training programs, the payoff could be reduced poverty, reduced taxes, and improved health and well-being of our children.

Appendix: Flowchart for Studies of Training Programs


No
No
START
Are you measuring impacts?
Yes
What method did you use to estimate what would have happened to participants in the absence of the program?
Comparison group
• Did you select your comparison group from those who were eligible for participation in the program in the month in which the participant enrolled?
• Were they drawn from the same local labor market?
• Was eligibility for the program determined in the same way for the comparison group as for participants?
Yes
Random assignment
No
Instrumental variables
Yes
No
Are the characteristics of your control group statistically significantly different from your treatment group?
Did your instrument pass a test such as in Startz, Nelson, and Zivot (2001)?
Did you control for measured differences between your participants and the comparison group based on propensity score and then adjusting for residual differences with regression? Did you bootstrap your standard errors?
Yes
Yes
No
Is your outcome variable measured using administrative or survey data?
Improve response rate
Yes
No
How many records did you send to the survey research firm (a)?
How many responses did you get (b)?
Is (b) divided by (a) greater than 0.75?
Yes
No
FINISH
Yes
Was the response rate the same for the participants and the comparison/controls?

References

Altman, Douglas G., and J. Martin Bland. 1999. “Treatment Allocation in Controlled Trials: Why Randomize?” *British Medical Journal* 318: 1209.

Angrist, Joshua D., and Alan Krueger. 1992. “Estimating the Payoff to Schooling Using the Vietnam-Era Draft Lottery.” NBER Working Paper 4067. Cambridge, Mass.: National Bureau of Economic Research.

Barnow, Burt S. 1987. “The Impact of CETA Programs on Earnings: A Review of the Literature.” *Journal of Human Resources* 22 (2): 157–193.

———. 2000. “Exploring the Relationship between Performance Management and Program Impact: A Case Study of the Job Training Partnership Act.” *Journal of Policy Analysis and Management* 19 (1): 118–141.

Betsey, Charles L., Robinson G. Hollister, Jr., and Mary R. Papageorgiou. 1985. *Youth Employment and Training Programs: The YEDPA Years*. Washington, DC: National Academy Press.

Bloom, Howard S. 1984. “Accounting for No-Shows in Experimental Evaluation Designs.” *Evaluation Review* 8: 225–246.

Bloom, Howard S., et al. 1993. *The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months*. Research and Evaluation Report Series 93-C. Washington, DC: Department of Labor. Available from website: wdr.doleta.gov/opr/FULLTEXT/1993_23.pdf; accessed April 20, 2002.

———, et al. 1997. “The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study.” *Journal of Human Resources* 32 (3): 549–576.

Bound J., D. Jaeger, and R. Baker. 1995. “Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogenous Explanatory Variable Is Weak.” *Journal of the American Statistical Association* 90: 443–450.

Bryant, Edward C., and Kalman Rupp. 1987. “Evaluating the Impact of CETA on Participant Earnings.” *Evaluation Review* 11 (4): 473–492.

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

Cain, Glen, et al. 1993. “Using Data on Applicants to Training Programs to Measure the Program’s Effects on Earnings. Discussion Paper 1015-93. Madison: University of Wisconsin, Institute for Research on Poverty.

Canadian Comprehensive Auditing Foundation. 1987. *Effectiveness: Reporting and Auditing in the Public Sector*. Ottawa: CCAF.

Day S.J., and D.G. Altman. 2000. “Blinding in Clinical Trials and Other Studies.” *British Medical Journal* 321: 504.

Dehejia, R.H., and S. Wahba. 1999. “Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs.” *Journal of the American Statistical Association* 94: 1053–1062.

Dickinson, Katherine P., Terry R. Johnson, and Richard W. West. 1986. “An Analysis of the Impact of CETA Programs on Participants’ Earnings.” *Journal of Human Resources* 21: 64–91.

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

Friedlander, Daniel, and Philip K. Robins. 1994. “Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods.” Working paper. October.

Goldberger, Arthur S. 1983. “Abnormal Selection Bias.” In S. Karlin, T. Amemiya, and L. Goodman, eds., *Studies in Econometrics, Time Series and Multivariate Statistics*. Stamford, Conn.: Academic Press.

Heckman, James J. 1976. “The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for such Models.” *Annals of Economic and Social Measurement* 5: 475–492.

———. 1999. “Policies to Foster Human Capital.” Presentation to Aaron Wildavsky Forum. Available from website: http://lily.src.uchicago.edu/papers/labor/Wildavsky.pdf; accessed April 20, 2002.

———, et al. 2000. “Substitution and Drop Out Bias in Social Experiments: A Study of an Influential Social Experiment.” *Quarterly Journal of Economics* 115 (2): 651–694.

———, et al. 1998. “Characterizing Selection Bias Using Experimental Data.” *Econometrica* 66 (5): 1017–1098.

———, H. Ichimura, and P. Todd. 1997. “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program.” *Review of Economic Studies* 64 (4): 605–654.

———, H. Ichimura, and P. Todd. 1998. “Matching as an Econometric Evaluation Estimator.” *Review of Economic Studies* 65 (2) 261–294.

———, Robert LaLonde, and Jeffrey Smith. 1999. “The Economics and Econometrics of Active Labor Market Programs.” In Orley Ashenfelter and David Card, eds. *Handbook of Labor Economics*, vol. 3A. Amsterdam: North-Holland.

———, and Jeffrey A. Smith. 1995. “Assessing the Case for Social Experiments.” *Journal of Economic Perspectives* 9 (2): 85–110.

———, and Jeffrey A. 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.

———, Jeffrey Smith, and Christopher Taber. 1998. “Accounting for Dropouts in Evaluations of Social Programs.” *Review of Economics and Statistics* 80 (1): 1–14.

Hogberg, David. 2001. “The 2002 Federal Budget: Let’s Do Some Cuttin’!” *Institute Brief* 8 (24). Available from website: www.limitedgovernment.org/pubs/brf/brf8-24.PDF.

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

Lumley, Judith, and Hilda Bastian. 1996. “Competing or Complementary?” *International Journal of Technology Assessment in Health Care* 12 (2): 247–263.

Moffitt, Robert. 1992. “Evaluation Methods for Program Entry Effects.” In C. Manski and I. Garfinkel, eds., *Evaluating Welfare and Training Programs*. Cambridge, Mass.: Harvard University Press.

Morris, Pamela A., et al. 2001. *How Welfare and Work Policies Affect Children: A Synthesis of Research*. New York: Manpower Demonstration Research Corporation. Available from website: www.mdrc.org/Reports2001/NGChildSynth/NG-childSynth.pdf; accessed April 20, 2002.

Nelson, Charles R., and Richard Startz. 1990. “Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator.” *Econometrica* 54 (4): 967–976.

Patton, Michael Q. 1997. *Utilization-Focused Evaluation*. Thousand Oaks, Cal.: Sage Publications.

Reubens, Beatrice G. 1980. “Review of Foreign Experience.” In Bernard E. Anderson and Isabel V. Sawhill, eds., *Youth Employment and Public Policy*. Englewood Cliffs, NJ: Prentice Hall.

Riccio, James, et al. 1994. *GAIN: Benefits, Costs, and Three-Year Impacts of a Welfare-to-Work Program: California’s Greater Avenues for Independence Program*. New York: Manpower Demonstration Research Corporation.

Rosenbaum, P.R., and D.B. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70: 41–55.

Rubin, Donald B. 1979. “Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies.” *Journal of the American Statistical Association* 74: 318–328.

Sacks, H.S., T.C. Chalmers, and H. Smith. 1982. “Randomized Versus Historical Controls for Clinical Trials.” *American Journal of Medicine* 72: 233–240.

Schulz, K.F., et al. 1995. “Empirical Evidence of Bias: Dimensions of Methodological Quality Associated with Estimates of Treatment Effects in Controlled Trials.” *Journal of the American Medical Association* 273: 408–412.

Sleight, P. 1996. “Calcium Antagonists During and After Myocardial Infarction.” *Drugs* 51 (2): 216–225.

Smith, Jeffrey, and Petra Todd. 2001. "Reconciling Conflicting Evidence on the Performance of Matching Estimators." *American Economic Review* 91 (2): 112–118.

Smith, Tom W. 1999. "Developing Nonresponse Standards." Presentation to International Conference on Survey Nonresponse, Portland. Available from website: www.norc.uchicago.edu/online/nonre.htm; accessed April 20, 2002.

Social Program Evaluation Group. 1986. "Review of the Literature Related to the Evaluation of Employment Programs for Social Assistance Recipients." Kingston, Ont.: Queen's University. Unpublished.

Startz, Richard, Charles R. Nelson, and Eric Zivot. 2001. "Improved Inference for the Instrumental Variable Estimator." Available from website: www.econ.washington. edu/econprelim/users/startz/working_papers/ImprovedIV.pdf.

Stromsdorfer, Ernst W., et al. 1985. *Recommendations of the Job Training Longitudinal Survey Research Advisory Panel*. Washington, DC: Department of Labor, Employment and Training Administration.

Warburton, William P. 1996a. "What Went Wrong in the CETA Evaluations?" *Canadian Journal of Economics* 29 (special issue, part 1): 105–108.

———. 1996b. "Estimating the Impact of Selected Programs on Participants' Subsequent Welfare Dependence and Employment in British Columbia." PhD thesis. London: University of London.

Recent Issues of *C.D. Howe Institute Commentary*

No. 164, May 2002 Hartt, Stanley H., and Patrick J. Monahan. “The Charter and Health Care: Guaranteeing Timely Access to Health Care for Canadians.” 28 pp.; $12.00; ISBN 0-88806-557-4.

No. 163, May 2002 Aba, Shay, Wolfe D. Goodman, and Jack M. Mintz. “Funding Public Provision of Private Health: The Case for a Copayment Contribution through the Tax System.” 20 pp.; $12.00; ISBN 0-88806-550-7.

No. 162, April 2002 Dobson, Wendy. “Shaping the Future of the North American Economic Space: A Framework for Action.” 32 pp.; $12.00; ISBN 0-88806-551-5.

No. 161, March 2002 Palmer, John P. “Bread and Circuses: The Local Benefits of Sports and Cultural Businesses.” 18 pp.; $12.00; ISBN 0-88806-545-0.

No. 160, February 2002 Slack, Enid. “Municipal Finance and the Pattern of Urban Growth.” 25 pp.; $10.00; ISBN 0-88806-544-2.

No. 159, February 2002 Jaccard, Mark. “California Shorts a Circuit: Should Canadians Trust the Wiring Diagram?” 27 pp.; $10.00; ISBN 0-88806-546-9.

No. 158, February 2002 Laidler, David, and Shay Aba. “Productivity and the Dollar: Commodities and the Exchange Rate Connection.” 15 pp.; $10.00; ISBN 0-88806-547-7.

No. 157, January 2002 Donaldson, Cam, Craig Mitton, and Gillian Currie. “Managing Medicare: The Prerequisite to Spending or Reform.” 20 pp.; $10.00; ISBN 0-88806-540-X.

No. 156, November 2001 Richards, John. “Neighbors Matter: Poor Neighborhoods and Urban Aboriginal Policy.” 37 pp.; $10.00; ISBN 0-88806-542-6.

No. 155, November 2001 Finnie, Ross. “Measuring the Load, Easing the Burden: Canada’s Student Loan Programs and the Revitalization of Canadian Postsecondary Education.” 32 pp.; $10.00; ISBN 0-88806-538-8.

No. 154, November 2001 Bird, Richard M., and Kenneth J. McKenzie. “Taxing Business: A Provincial Affair?” 32 pp.; $10.00; ISBN 0-88806-539-6.

No. 153, May 2001 Chant, John F. “Main Street or Bay Street: The Only Choices?” 24 pp.; $10.00; ISBN 0-88806-532-9.

No. 152, May 2001 Prentice, Barry E., and Tamara Thomson. “An Electronic System for Railcar Market Access.” 22 pp.; $10.00; ISBN 0-88806-530-2.

No. 151, April 2001 Donaldson, Cam, Gillian Currie, and Craig Mitton. “Integrating Canada’s Dis-Integrated Health Care System: Lessons from Abroad.” 24 pp.; $10.00; ISBN 0-88806-526-4.

No. 150, March 2001 Bish, Robert L. “Local Government Amalgamations: Discredited Nineteenth-Century Ideals Alive in the Twenty-First.” 35 pp.; $10.00; ISBN 0-88806-525-6.

No. 149, February 2001 Kesselman, Jonathan, and Finn Poschmann. “A New Option for Retirement Savings: Tax-Prepaid Savings Plans.” 39 pp.; $10.00; ISBN 0-88806-524-8.

No. 148, February 2001 Robson, William B.P. “Will the Baby Boomers Bust the Health Budget? Demographic Change and Health Care Financing Reform.” 29 pp.; $10.00; ISBN 0-88806-523-X.

The C.D. Howe Institute

The C.D. Howe Institute is a national, nonpartisan, nonprofit organization that aims to improve Canadians' standard of living by fostering sound economic and social policy.

The Institute promotes the application of independent research and analysis to major economic and social issues affecting the quality of life of Canadians in all regions of the country. It takes a global perspective by considering the impact of international factors on Canada and bringing insights from other jurisdictions to the discussion of Canadian public policy. Policy recommendations in the Institute's publications are founded on quality research conducted by leading experts and subject to rigorous peer review. The Institute communicates clearly the analysis and recommendations arising from its work to the general public, the media, academia, experts, and policymakers.

The Institute was created in 1973 by a merger of the Private Planning Association of Canada (PPAC) and the C.D. Howe Memorial Foundation. The PPAC, formed in 1958 by business and labor leaders, undertook research and educational activities on economic policy issues. The Foundation was created in 1961 to memorialize the late Rt. Hon. Clarence Decatur Howe, who served Canada as Minister of Trade and Commerce, among other elected capacities, between 1935 and 1957. The Foundation became a separate entity in 1981.

The Institute encourages participation in and support of its activities from business, organized labor, associations, the professions, and interested individuals. For further information, please contact the Institute's Development Officer.

The Chairman of the Institute is Kent Jespersen; Jack M. Mintz is President and Chief Executive Officer.