Practical guide to understanding returns to training investments

Adult Learning and Returns to Training Project

Arthur Sweetman | Marc Frenette | Karen Myers | Jean-Pierre Voyer

May 2014

SRDC SRSA

SOCIAL RESEARCH AND DEMONSTRATION CORPORATION | SOCIÉTÉ DE RECHERCHE SOCIALE APPLIQUÉE

The Social Research and Demonstration Corporation (SRDC) is a non-profit research organization, created specifically to develop, field test, and rigorously evaluate new programs. SRDC's two-part mission is to help policy-makers and practitioners identify policies and programs that improve the well-being of all Canadians, with a special concern for the effects on the disadvantaged, and to raise the standards of evidence that are used in assessing these policies.

Since its establishment in December 1991, SRDC has completed over 100 projects and studies for various federal and provincial departments, municipalities, as well as other public and non-profit organizations. SRDC has offices located in Ottawa, Toronto, and Vancouver.

For information on SRDC publications, contact

Social Research and Demonstration Corporation
55 Murray Street, Suite 400
Ottawa, Ontario K1N 5M3
613-237-4311 | 1-866-896-7732
info@srdc.org | www.srdc.org

Vancouver Office
128 West Pender Street, Suite 301
Vancouver, British Columbia V6B 1R8
604-601-4070 | 604-601-4080

Toronto Office
481 University Avenue, Suite 705
Toronto, Ontario M5G 2E9
416-593-0445

Published in 2014 by the Social Research and Demonstration Corporation

Table of contents

Introduction

Report purpose

This *Practical Guide to Understanding Returns to Investments* is one in a series of papers that have informed the development of an analytical framework for the *Adult Learning and Returns to Training Project.* The project is a three-year multi-disciplinary and collaborative effort to further the knowledge base of conceptual, analytical and methodological issues concerning the scope and measurement of adult learning activities and their associated financial and non-financial returns to individuals, firms and society at large.

The overall purpose of the project is to develop and test an analytical framework for estimating returns to various types of adult learning activities. The *Typology* report defined the various types of adult learning that will be considered in this project. The *What Matters What Should Count* report outlined a framework for conceptualizing the wider benefits associated with adult learning. While these two reports deal with the question of *what* should count, this report deals with the question of *how* to count. Simply put, how do we estimate returns to adult learning and training activities?

It is a generally accepted principle of public management that governments should invest if benefits exceed costs. But with this principle comes the more complicated question of *how should society assess the value of the vast array of interventions* that are either proposed or are operating in social policy arenas. More fundamentally, *how can impacts be catalogued* given as Gertler, Martinez, Premand, Rawlings, and Vermeersch (2011) point out, potential impacts may be broad — a mix of cognitive, behavioural, labour market and social dimensions; long term — perhaps the rest of the lives of participants; and variable over time — either increasing or decreasing with time from the initial intervention.

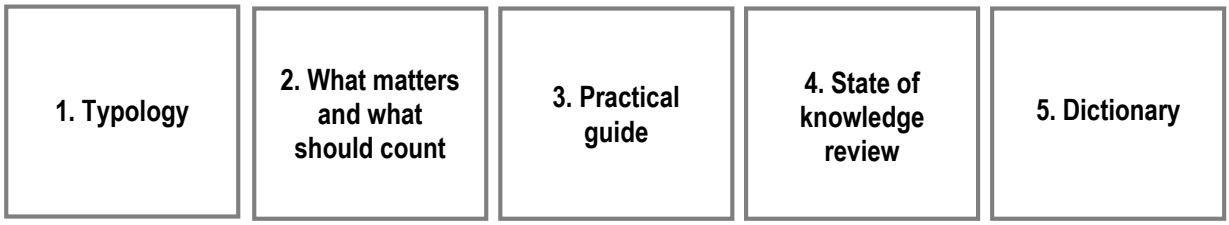
This report addresses this issue by proposing a *cost-benefit approach* as a useful overarching framework for comprehensively taking account of the full range of benefits and costs associated with adult learning interventions. We argue that even if a full cost-benefit analysis is not conducted, the approach imposes an intellectual discipline that allows analysts to explicitly recognize and value all the (social, financial, aesthetic, and other) costs and benefits of a program. The first section of this report provides a brief introduction to the cost-benefit analysis approach and discusses its potential application to social policy in general and adult learning more specifically.

We also argue that a cost-benefit analysis is a moot point if evidence of a causal benefit cannot be found. That said, determining what kinds of evidence should drive decision making is challenging. Evaluation studies come in all forms, vary on many dimensions, and sometimes conflict. The studies that bear on the question of what works are often scattered across different disciplines, are sometimes disseminated through inaccessible outlets, and can be of such questionable quality that interpretation is risky at best. How then can policy and practice be informed by such a fragmented knowledge base? In the second part of this report we outline a strategy for determining whether a study provides a credible

estimate of a casual impact. In this section, we present a hierarchy of evidence as a systematic way of ranking research designs aimed to estimate causal impacts according to their scientific validity.[1]

Companion reports

The analytical framework for the *Adult Learning and Returns to Training Project* consists of five companion reports. The first report is the *Typology*. The *Typology* report proposes a typology of **adult learning** activities created for this project. The second report, *What Matters and What Should Count*, provides a high-level conceptual framework for investigating a wide range of outcomes associated with various types of adult learning. The third piece is this *Practical Guide* which is intended to be a user-friendly guide to understanding key methodological issues in the literature on returns to adult learning. The fourth piece is a *State of Knowledge Review* that analyzes what we know and do not know about the wider impacts of adult learning. The fifth piece is a *Dictionary* report that provides definitions for key concepts from all of the companion reports.



[1] See Smith and Sweetman 2010 for a more extensive discussion of the current state of knowledge on alternative ways of estimating the causal effects of programs and policies.

Part 1: Assessing the value of adult learning interventions

Conceptualizing adult learning as an investment

Despite the widespread acceptance of the importance of human capital in a knowledge economy, adult learning programs are often considered as costs rather than investments. A motivating hypothesis for the *Adult Learning and Returns to Training Project* is that policies to encourage adult learning may not only improve individual life chances but also benefit the larger society and economy and generate public returns long after assistance has ended.

While the body of evidence needed to test this hypothesis with respect to adult learning is still in its infancy, there is considerable evidence to suggest that, in principle, well-designed social programs can be good investments.[2] Indeed, there are studies that suggest some social programs may produce substantial benefits that well exceed the cost. The widely cited Perry pre-school study is one of the best examples.

The Perry pre-school program was among the first carefully controlled social experiments to quantify the costs and benefits of early childhood education well into adulthood. Researchers followed young children through adulthood, demonstrating that a carefully designed pre-school experience generated a wealth of benefits including: high school graduation, future employment and earnings, federal taxes paid, less involvement in the criminal justice system and lower demands on the welfare system. According to a recent and rigorous study by Heckman et al. (2010), the estimated social rates of return for the program fall between seven and ten percent. As Heckman et al. point out, these estimates are above the historical return to equity.

While the Perry pre-school program provides a compelling example of the potential of human capital as a social investment, there are equally compelling examples of programs that do not work and may even provide a negative return on investment. For example, while a recent study for Washington State found several good investments that improve the effectiveness of the state's criminal justice system, the study also found programs that did not lower criminality and in fact had negative economic bottom lines.[3] The study also found another category of programs that produced some benefits but the cost of these programs was considerably greater than any savings realized. Other programs were cost-beneficial with certain groups of people in certain settings but not others.

As fiscal constraints grow more severe, being able to quantify the indirect benefits that accrue to taxpayers, who pay for effective social programs even though they may not directly benefit from them, may be as important to the ultimate survival of these programs as evidence of their impacts on those

[2] It is important to point out that social programs can also have direct consumption value and make other non-investment contributions to society.

[3] The study conducted by the Washington State Institute for Policy (Aos et al., 2011) found that several interventions produced benefit-to-cost ratios that exceeded twenty dollars of benefits for each dollar of taxpayer cost. That is, a dollar spent on these programs today could be expected to return to taxpayers and crime victims twenty or more dollars in the years ahead. The study also found that even programs that achieve relatively small reductions in crime can be cost-beneficial.

who benefit directly. This may be especially true in the case of adult learning programs where the immediate costs are relatively large and the payoff may not be as evident until much later in time.

What is cost–benefit analysis?

Cost–benefit analysis is the primary tool that economists employ to determine whether a particular policy, or policy proposal, promotes economic efficiency. Efficiency can be simply defined as getting the most value in a particular context from the resources available. The notion of efficiency includes technical efficiency, which means producing things of value in ways that involve giving up the smallest amounts of other things of value. More generally it is also concerned with the allocation of resources to generate the largest aggregate value, as assessed by summing individual valuations across all members of society (Weimer & Vining, 2009). A policy or policy alternative achieves optimal efficiency if no other policy can be identified that offers a larger excess of benefits over costs.

Conceptually, cost–benefit analysis is quite simple. It reduces all impacts of a proposed alternative to a common unit of measure, namely dollars. The purpose is not to price everything, but rather to order choices in a way that is informative about social choices for public decision makers. At the most general and comprehensive level, cost–benefit analysis is an aggregator of all impacts, to all affected parties, at all points in time. Once all impacts have been reduced to dollars, the evaluation rule is relatively straightforward. Choose the alternative that generates the largest aggregate net benefits (in dollars). Note that in its basic form, cost–benefit analysis places primary weight on economic efficiency, but it can also be modified to account for adverse wealth distribution effects by appropriately weighting the costs and benefits to individuals or groups.

The concept of "returns" is used to summarize all of the factors described above (i.e., all costs and benefits, accounting for time). In its simplest formation (expressed as a percentage):

$$\text{Rate of Return (RoR)} = \frac{(\text{Benefits} - \text{Costs})*100\%}{\text{Costs}}$$

In practice, the analysis not only includes costs and benefits, but also their timing. Cost–benefit analysis aims to put all relevant costs and benefits on a common temporal footing. Adjustments are made for the time value of money, so that all flows of benefits and flows of project costs (which tend to occur at different points in time) are expressed on a common basis in terms of their "present value." This is often done by converting the future expected streams of costs and benefits into a present value amount using a suitable discount rate (see Box 1).

In addition, it is important to identify who bears the costs or benefits of the program since a program's effects can represent gains from one party or perspective and losses from another. Determining whose costs and benefits should count is referred to as standing. In the case of adult learning, program costs and benefits should be shown from at least four perspectives:

- Participant's (frequently called "private" since individuals are part of the "private sector")
- Firm's (also frequently called "private" since firms are part of the "private sector")
- Government's (frequently called "public" since government is part of the "public sector")
- All of society's (frequently called "social;" it encompasses all other perspectives)

It is also common to look at program costs and benefits for individuals who are non-participants, especially certain subgroups such as family members or children of participants. Sometimes effects on the neighbourhoods in which participants reside are also studied.

Box 1 Accounting for time

There are at least two reasons why time matters. First, individuals have subjective preferences for receiving benefits earlier rather than later. On the margin, individuals value consumption now more than later in life and some discount the future more heavily than others. Second, receiving more money (or other resources) sooner opens up new opportunities. If individuals earn more earlier on, they can invest immediately and enjoy more years of capital appreciation (and or interest/dividends). For these reasons, future earnings must be discounted in some way. The goal in such exercises is to compare apples to apples, or current dollars to current dollars. In practice, it is easier to discount based on financial market rates of return rather than by time preferences which vary from one individual to another and are not usually known. Doing so assumes financial market rate of returns reflect the average time preference of all society, which might fail to properly discount future earnings for some particular groups or individuals. See Burgess and Jenkins (2010) for a discussion of choosing appropriate discount rates.

Let's use a simplified example that assumes time indifference. Suppose individuals can earn a rate of return of r (e.g., 5% or r = 0.05) on their investments over a period of one year. In this case, it is clear that 10 dollars today is equivalent to (10 + 10*r), or simply 10*(1 + r) dollars one year from now. The reason is that the 10 dollars can be invested and earn a rate of r over the one-year period.

Now, let's reverse the situation. What is 10*(1 + r) dollars one year from now worth today? The answer is 10 dollars of course. This is intuitively clear, but the way we obtain this number is by dividing the dollars received one year from now (10) by (1 + r). So, to discount money (earnings) received one year from now, we simply need to divide that sum by (1 + r), where r is the annual rate of return. What about two years from now? If 10 dollars today is worth 10*(1 + r) dollars one year from now, then 10 dollars will be worth 10*(1 + r)*(1 + r) dollars two years from now. The reason is that 10*(1 + r) will earn a rate of return of r in the second year. To discount money or earnings received two years from now, we simply need to divide by (1 + r)*(1 + r), or $(1 + r)^2$. In general, we discount money received n periods in the future by dividing it by $(1 + r)^n$.

Of course, in reality, investments yield returns at a continuous rate. We've simplified things by showing the case when money earns returns at discrete points only. Nevertheless, the key points are the same with this simpler approach.

Putting all of these concepts together, the human capital model suggests that individuals will invest in training if the sum of discounted net benefits over a lifetime is positive. This is another way of saying that individuals will invest in training if it pays to do so. From this model, it is clear that younger workers have more incentives to invest in their human capital because they have more years to see their earnings grow. **Note:** In practice there is sometimes confusion between the "discount rate" that accounts for people being present-oriented and/or the investment value of funds/resources, and the "inflation rate" that makes a dollar today have greater purchasing power than a dollar in the future (assuming positive inflation). These are two distinct concepts and any empirical study needs to carefully identify how each is treated.

A life cycle model of human capital accumulation

Once the time-value of the flows of costs and benefits are understood, and the nature of the human life cycle taken into account, then a number of implications ensue. Most important among these are that a substantial number of pre-retirement (or pre-death) years are commonly required to recoup the initial costs (in terms of time, effort, financing, etc.) of a human capital investment. This implies that investments are best made early in life (or, more precisely farther away from labour force withdrawal), and it also implies that investments that reap benefits post-retirement/withdrawal (e.g., those with non-financial payoffs) are frequently easier to justify later in life. Both retirement and death are stochastic events in the sense that we do not know when they will occur (and retirement may be experienced in phases with, for example, part-time work preceding full retirement and/or it may be temporarily reversed with individuals rejoining the labour market after an initial withdrawal). Nevertheless, we have good estimates of the distribution of both and can make sensible predictions regarding the expected value of an adult education investment given a worker's age. Workers frequently recognize these realities and are reluctant to make large (in terms of time or costs) adult education investments late in life unless they also have substantial current consumption value.

In practice, some adult human capital investment combines both skills development and skills retention. Skills retention is increasingly an important justification for adult learning. As Green and Riddell (2013) illustrate using data for Canada, Norway, and the United States, literacy and numeracy skill deterioration across the life-cycle is a very real issue that has important implications for labour market productivity. (Differences in skills across birth cohorts are also measurable). The deterioration of basic skills may be particularly relevant for workers who are involuntarily displaced from their employment and even more so for workers who are also displaced from the industry and/or occupation in which they have accumulated valuable experience. (Thus workers in declining industries experience, on average, more severe losses than those who lose their job in a stable or expanding industry.) This is crucial since the basic skills are the foundations on which new skills associated with new industries and/or occupations are built. If adult education is valued not only for its role in skill development but also for that in skill retention, then its value is clearly increased for older workers.

A quite different aspect of the life-cycle hypothesis is founded on biological and physiological aspects of human development. With respect to language acquisition, for example, Bleakley and Chin (2004, 2008) build on what cognitive scientists – primarily psychologists and linguists – label the "critical period hypothesis" whereby children exposed to a new language during the critical period become fluent relatively easily, whereas those exposed later have much less certainty regarding attaining fluency. Language fluency among those who seriously take up a new language before the early teen years is almost identical to that for natives. However, proficiency drops off appreciably as the age of the learner increases beyond this threshold. They demonstrate the empirical importance of this phenomenon by comparing child and adult immigrants learning a new language in a receiving country – the differences are remarkable, but at the same time indicate that some level of ability to learn remains at all ages.

Challenges in applying cost–benefit analysis to social policy

Although cost–benefit analysis was initially developed to evaluate infrastructure investments, it has long been used to evaluate a wide range of investments including economic regulation, environmental initiatives and, especially in the United States, social policy with a particular emphasis on human capital investments.[4] The approach can encompass labour market, social and other program effects. In a "strict" cost–benefit analysis all costs and benefits (including ones that are not normally monetized such as, for example, aesthetic, moral and psychological costs and benefits) are converted into monetary equivalents, but in practice, program induced causal changes in inputs and outcomes of this type are sometimes only listed and not monetized.[5] In the United States it is common to include social benefits such as reductions in teen pregnancy, drug and tobacco use, and incarceration/recidivism rates, and increases in such activities as volunteer participation — these activities are frequently priced at the cost to the treasury/government with broader social benefits acknowledged but not explicitly monetized. In many situations, rather than undertaking a full-fledged cost–benefit analysis, it is common to do a cost-effectiveness study (to determine the most cost-effective approach to attaining a particular goal without explicitly monetizing non-financial outcomes).[6]

In recent years much attention has focused on estimating causal impacts, which are fundamental inputs to cost–benefit analysis, rather than conducting full cost–benefit analyses. In the absence of an appreciable positive impact (i.e., without an identifiable benefit), there is not much use in conducting a full cost–benefit analysis.[7] This emphasis follows from the considerable debate that frequently surrounds claims regarding the "value added" or "incremental benefit" of particular programs or initiatives. However, even if full cost–benefit analyses are infrequent, the conceptual framework is extremely useful in decision-making. It imposes an intellectual discipline that allows planners and evaluators to explicitly recognize and value all the (social, financial, aesthetic and other) costs and benefits of a program. It makes explicit aspects of a program's costs and benefits that are sometimes vague, implicit or neglected (which may sometimes be useful).

The application of cost–benefit analysis to social policy raises a number of issues that deserve special attention. In a recent report published by the Society for Benefit-Cost Analysis, Vining and Weimer (2009) address general considerations in conducting a cost–benefit analysis in social policy domains. They identify four key issues.

---

4 Relevant examples include Bloom et al. (1997) who review programs that are part of the US Job Training Partnership Act (JTPA), which had an appreciable adult education and training component, and Mathematica's evaluation reports of the US Job Corps education and training program for disadvantaged youth (e.g., McConnell & Glazerman, 2001 and Schocket, Burghardt, & Glazerman, 2001).

5 In practice, the greatest controversy regarding the monetization of costs and benefits are not usually generated by social policy initiatives, but by environmental ones.

6 Other related concepts are cost-utility analysis, and the estimation of incremental cost-effectiveness ratios. For a textbook length discussion of relevant issues from a sociological/public policy/psychological perspective see Rossi, Freeman, and Lipsey (2004).

7 While addressing education broadly, rather than focusing on adult education, the US Department of Education's "What Works Clearinghouse" (http://ies.ed.gov/ncee/wwc/) is an important source of information on impact assessment in the education context, and how it feeds into cost–benefit analysis.

Comprehensiveness

A standard principle of the cost–benefit approach is that analyses should be comprehensive in terms of taking account of all valued effects in predicting net benefits. The comprehensiveness principle poses a particular challenge in the social policy domain since social program effects are generally effects which are difficult to predict and value. Social programs often have effects that spillover from one domain to another. For example, investments in adult learning are often designed to affect labour market performance but may have spillover effects such as improved health outcomes. Learning programs may also generate effects that are likely to persist over long periods of time but are difficult to predict. For example, an adult learning program may produce intergenerational effects such as improved health and educational outcomes for the children of adult learners in early childhood. And these effects may persist for decades after the investment, requiring predictions of a chain of effects leading from early cognitive development through formal school achievement all the way to labour market entry and career advancement. Adult learning may also involve the reduction of negative externalities, such as crime, which require valuation. As Vining and Weimer (2009) point out, the metrics, or "shadow prices," for a number of these externalities are currently uncertain. The recent work by Heckman et al. (2010) on the returns to the Perry Preschool program provides an excellent example. The authors present an extensive sensitivity analysis showing significant consequences of alternative assumptions about the social costs of crime for the estimated rate of return.

As we demonstrate in the *State of Knowledge* report, the body of evidence on adult learning has not addressed the question of societal impacts. Most studies focus on either private financial outcomes estimated under very short time horizons or even fall back on estimating intermediate outcomes that they hope will predict final outcomes. A further challenge is that, as with many social policy areas, adult learning programs likely work well for some groups but not others. Without careful subgroup analysis, studies may come to incorrect conclusions about program effectiveness. In addition, there may be considerable unobserved heterogeneity at the individual level. This may result in a situation where an impact estimate is not statistically significantly different from zero, but the standard errors are so large that the "truth" could be zero, or could be a relatively high ROI.

Uncertainty

Uncertainty — both in terms of prediction of effects and valuation of shadow prices — is a major challenge to applying cost–benefit analysis to social policy domains. Comprehensiveness requires the valuation of all effects including those that might not achieve conventional levels of statistical significance in particular studies. The challenge is particularly great when a single study provides a number of important estimates of effects and few of these effects have statistical significance under the rules of multiple comparisons. Vining and Weimer (2009) argue that because of these uncertainties, sensitivity analysis is an important step in the process. Sensitivity analysis is a tool for testing the robustness of findings to inherent uncertainties and the need for assumptions. The idea is to simply replace unknown or uncertain parameters with alternative values drawn from a plausible distribution. Researchers might, for example, conduct sensitivity analysis over alternative specifications of the discount rate and estimated magnitude of each impact. The conclusions of any cost–benefit analysis

should be explicit about whether results are sensitive to particular parameters, especially when they are associated with uncertainty, differences of opinion, or both.

Non-standard behaviour

Social policies quite often involve behaviours that do not necessarily conform to the principles of neoclassical economics that underpin welfare economics generally and cost–benefit analysis specifically. The case of addiction is a particularly compelling example. Should consumption that satisfies addiction be treated the same way as consumption that contributes to utility under the assumption of positive marginal utility? As Vining and Weimer (2009) point out, the application of cost–benefit analysis to policies that affect substance abuse often requires an answer to this question.

Distributional goals

In contrast to many other policy areas, concerns about equity legitimately motivate the adoption of many social policies. While the standard cost–benefit analysis framework is based solely on the value of efficiency, it is possible to elicit people's willingness-to-pay for various sorts of re-distributional outcomes. For example, it may be possible to estimate how much Canadians are willing to pay to move a child or family above the poverty line. Equity effects valued in this way may be included in cost–benefit analyses through a broader definition of efficiency. This approach is often taken in cost–benefit analyses of environmental policies. These studies may include not only people's willingness to pay for the consumption (use) of private goods, but also their willingness to pay for consumption (non-use) of public goods, including more equitable redistributions. This approach is not common in social policy, partly because there are not yet good shadow price estimates of these external effects (but see, for example, Blomquist et al., 2009). Another approach is to embed a standard analysis within a multi-goal framework that takes account of equity as a second value. The consideration of equity may involve a trade-off between these two values. However, there may be some fortuitous situations in which well-targeted social policies increase both efficiency and equity. (See Box 2.)

Box 2 Efficiency versus equity

Not all social programs should be subject to a cost–benefit test. But as Weimer and Vining (2009) argue, well-crafted cost–benefit analysis can help avoid routine dismissal of social programs as wasteful welfare spending when in reality they may be worthy investments. In this context, judicious use of cost–benefit analysis can supplement rather than replace the value of compassion in the formulation of social policy and may help protect programs. Moreover, balancing efficiency and equity does not always involve a trade-off. As Rogers (2003) points out, under certain conditions redistributive policies may enhance efficiency. For example subsidizing vaccination of children from low-income families to protect them against a communicable disease may increase the consumption of a good with a positive externality and promote a fairer distribution of preventative health care. Encouraging adults to invest in their human capital development later in life may be another area where the efficiency-equity double dividend is significant.

In addition, there are reasons to consider cost–benefit analysis even when efficiency is not the primary goal. Even in situations in which we wish to pursue a goal such as assistance to the poor as a matter of principle, efficiency can help identify better choices among possible means for achieving that goal. If cost–benefit analysis could identify policies that would produce identical gains then it would be sensible to choose the one that brought the lowest cost so that society would have more resources available for other uses. Concerns about the efficiency-equity trade-off are central to a wide range of policies that promote human capital investment. We agree with Weimer and Vining (2009) that we can go a long way to crafting better social policies if we can answer two questions: First, which policies and programs are or are likely to be efficient; second, for those programs that are deemed valuable for redistributive reasons, what alternatives involve the largest net benefits?

Conducting cost–benefit analysis

Although the idea behind cost–benefit analysis is simple, underlying it are extensive conceptual foundations drawn from microeconomics, welfare economics as well as a variety of widely accepted techniques for actually estimating costs and benefits. Every cost–benefit analysis is different regarding the appropriate methodologies and required assumptions. There are nevertheless some fundamental concepts that are common to most applications of cost–benefit analysis. An exposition of these conceptual foundations and techniques requires book-length treatment (see, for example, Boardman et al., 2006 or Weimer & Vining, 2009, for application to a range of social policies domains). Here we provide only a brief overview of the basic steps.

1. Assess evidence on what works

The first step is to produce estimates of policies and programs that have been shown to improve outcomes of interest. This involves carefully analyzing all high-quality research to identify those interventions that have best achieved outcomes and which ones have not. In this context, high quality research refers to studies with strong evaluation designs that are able to isolate and assess a program's impact. Isolating a program's impact requires an estimation of outcomes after the program, *and* an estimation of what the outcomes *would have been* at an equivalent point in time if the program had not happened. In other words it requires estimating the *counterfactual.* We discuss this point in detail in

the next section. As we also discuss in the next section, ideally a meta-analytic framework is used to systematically assess the entire research literature on a given topic. However, to do this effectively you need a body of evidence on what works and doesn't work in an area. In Canada most adult learning programs have not been rigorously evaluated. As we discuss in the *State of Knowledge* report, there are major gaps in our knowledge base. These gaps make is difficult to credibly estimate returns to learning investments.

Once program impacts are known, the next step is to project impacts over time. The time span for the analysis should be long enough to cover any differences in impacts between the alternative and the status quo. For infrastructure projects this is usually the predicted useful life of the facility. Social policy analysts rarely have such clear guidance. In addition, the empirical evidence often covers only a few years at best. Extrapolations are therefore sometimes necessary. Some policies, like adult education, may have significant impacts not only beyond participation in the program but for the next generation as well. While disagreement may exist about the most appropriate planning horizon — cost–benefit analysis requires specification of some duration, and the assumption should be made explicit, as it can significantly affect study results.

2. Calculate costs and benefits

The next task is to determine how much it costs to produce the outcomes identified in Step 1 and how much it is worth to citizens to achieve these outcomes. To answer these questions we need an economic model that provides internally consistent bottom line measures with standard financial statistics: net present values, benefit cost ratios and returns on investment. Standard approaches present estimates from three distinct perspectives: the benefits that accrue solely to program participants; those received by taxpayers; and any other measurable (non-participant and non-taxpayer) benefits. The sum of these perspectives provides a total view on whether a program produces benefits that exceed costs. Although these are reconciled in the encompassing social perspective, what may appear as a cost from one limited perspective (e.g., a tuition payment for a participant) could be a benefit from a different perspective (e.g., that same tuition payment for a provider). Of course, from a social perspective a tuition payment is simply a transfer between parties, and neither a cost nor a benefit at all, whereas the work that the student and instructor are not undertaking because of the time allocated to training is a real cost for society; that is, an important real cost to society is the opportunity cost of the participants' time.

The precise model needs to be tailored to a specific program area. For example, in the area of criminal prevention programs a key question is for every dollar spent on a program, can rates of future criminal activity be reduced to avoid at least that amount in downstream criminal justice costs? In other words, by spending a taxpayer dollar now on a program, will more than one taxpayer dollar be saved in the years ahead? In this case we could also consider the crime victim's perspective. If a program can reduce rates of future criminal offending, not only will taxpayers receive benefits but there will also be fewer crime victims.

Many social programs also produce important benefits that are best thought of as avoided costs: delinquency and crime or morbidity and mortality. In the case of adult learning, we would expect to see a mixture of incremental benefits as well as avoided costs such as decreased reliance on the

government transfer system. This emphasizes the importance of the counterfactual as a conceptual and empirical tool. The primary aim of most human capital investments is to increase the labour productivity of the target population in some way. Additionally, this increased productivity may induce further indirect benefits, such as improvements in the health, self-esteem, and happiness of participants and the welfare of participants' children. Even more indirectly, there might also be reduced administrative costs in those agencies that administer income transfer programs. These indirect effects are also real benefits or costs, but they are much more difficult to measure (Greenberg & Knight, 2007). For suggestive studies on how some of these impacts might be measured see Grogger, Karoly, and Klerman (2002) and Morris (2001). Social interventions often generate more varied and complex impacts than other interventions. Identifying and measuring the full range of impacts is difficult but necessary.

Another challenge in measuring non-monetary outcomes relates to assigning a dollar value to them. In the private sector, cost–benefit analysis is necessarily a monetary counting exercise. However, public cost–benefit analysis involves non-monetary costs and benefits. What is the value of reduced crime to society? The answer is not clear. Economists sometimes look for answers through price systems (e.g., how much individuals are willing to pay to insure their property or to defend it with an alarm system, a fence, or a guard dog). However, the public already provides security against various crimes, and individuals who are content with that level of protection will "free-ride" and pay zero dollars for additional protection. Many of these individuals may have actually been willing to pay an amount less than what they are currently paying in taxes for protection. In these cases, it is impossible to know the value of crime reduction. Similar issues are prevalent in evaluating the benefits of health and other non-monetary outcomes.

Discounting is another standard feature of cost–benefit analysis that can affect results. When costs and benefits occur at different points in time, discounting makes adjustments to allow comparisons across time. As discussed in Box 1, benefits accrued in the future are worth less today. Discounting converts all future costs and benefits into their present value. The cost–benefit criteria are then a question of whether the present value net benefits are positive. Both the timing of impacts and the discount rate itself have an important effect. For instance, with a discount rate of five percent, a benefit of one hundred dollars in ten years has a present value of approximately sixty-one dollars, and if it occurs in fifty years, the present value drops to less than nine dollars. With a discount rate of eight percent, the present values are even lower (about forty-six dollars and two dollars, respectively). Even these simple examples demonstrate the significant effect that discounting can have: benefits that occur far into the future count much less against costs that are incurred earlier. There is general consensus about the need to discount, but the question of what discount rate to use is less clear. Some economists argue that discount rates should reflect market transactions in which people reveal how they actually make trade-offs across time. Others argue in favour of lower discount rates that reflect more normative judgments about how societies ought to place more weight on the future. In practice, discount rates tend to be picked according to average market rates of return on financial instruments, such as government bonds, with maturity corresponding to the expected duration of the impact under study (see Burgess & Jenkins, 2010).

A further challenge that is especially problematic in the area of adult learning is how to estimate costs and benefits associated with interventions that are quite modest in magnitude. Take the case of an

adult learning program that costs $1,000 per person. Even a 10% ROI — which is generally considered high to reasonable for a social program — would only amount to $100. This situation is further complicated if this return accrues over a number of years). How do we measure a $1,100 increase ($1,000 to recoup the investment, and $100 to obtain the return on investment) in annual earnings (with $1,100 potentially spread over a few years)? Understanding the timing, nature and magnitude of the expected return can help in designing an evaluation strategy by identifying which strategies are likely not worth pursuing since the expected benefit, even when it is economically and socially large, may be too difficult/costly to measure in practice.

3. Perform sensitivity analysis

The third analytical step involves testing the robustness of the estimates. Because of considerable uncertainties with applying cost–benefit analysis to social policy domains, sensitivity analysis is an important step in the process. Sensitivity analysis is a tool for testing how conclusions might change when assumptions are altered. One approach is to perform a "Monte Carlo simulation" in which the key factors in calculations are systematically altered. The idea is to simply replace unknown or uncertain parameters with alternative values drawn from a plausible distribution. The purpose of the sensitivity analysis is to determine the odds that a particular approach will at least break-even.

4. Provide a portfolio analysis of a combination of policy options

From a government perspective it is probably more useful to compare results from one program to another, rather than solely focusing on the absolute value of any particular benefit-to-cost ratio. For example the Washington State Institute for Policy prepares "Consumer Report" like lists of what works and what does not work, ranked by cost–benefit estimates. The Institute estimates the degree to which a portfolio of policies is likely to affect big picture outcomes such as crime or high school graduation rates. For example in their 2001 report they estimate how a combination of prevention, juvenile justice and adult corrections programs could influence Washington's crime rate, the need to build prisons, and overall state and local criminal justice spending. This step moves away from lists of what works to a strategic analysis of ways to improve state-wide outcomes (Aos et al., 2011).

Part 2: Hierarchy of evidence for estimating program impacts

Policymakers need a wide range of information to inform their decision making. For example, to identify the scope and severity of a problem such as poverty, policy analysts use government statistics on poverty rates, levels, and durations. These data respond to the question: "What is going on?" To obtain information on risk factors that are associated with poverty spells, one would ideally use longitudinal studies that track families over time. This type of data can answer questions such as "how did this problem occur?" Determining "what works" to reduce poverty, however, requires a different type of scientific evidence. In this case, data from outcome or summative evaluations, or those studies that have tested the impact of some intervention on an outcome measure of poverty, are necessary. Consistent with the overall focus of this project on returns to adult learning, this section addresses the quality of research studies that address the last question: learning what works.

Learning what works requires more than examining the isolated results of one or two evaluations. Each evaluation study is part of a cumulative process in constructing knowledge about interventions (Lipsey, 1997). The challenge is how to engage in this process when studies are often scattered across different disciplines and vary considerably in terms of quality. How then can policy and practice be informed by such a fragmented knowledge base? What study, or set of studies, if any at all, ought to be used to influence policy?

As a starting point for addressing this challenge, in this section we present what is referred to in the literature as a *hierarchy of evidence*. A hierarchy of evidence offers a systematic way of ranking the strength of a study's findings. Specific research designs — combinations of the data for analysis, the institutional context and/or relevant theory, and the methodological approach(es) employed to extract information from the data — are ranked according to their scientific validity.[8] Formally, such hierarchies are not rankings of "findings" or of "evidence," but of the credibility of the data, context and methodology employed to obtain that evidence. That is, whether a study finds a particular initiative or program to have, for example, a large positive, modest or even detrimental effect is not relevant for its ranking in the hierarchy of evidence; rather, the hierarchy attempts to judge whether the findings, whatever they may be, are believable/credible.[9]

Hierarchies are usually developed relative to a specific scientific issue or related cluster of issues, such as the estimation of causal impacts in the case of the hierarchy presented here. This differs on several dimensions from the policymakers' decision regarding whether a program should be funded, which also involves moral, social, political and economic judgments. The purpose of a hierarchy is not for

[8] Of course, as Daly et al. (2007) point out, in practice judging the quality of evidence of an intervention should go beyond a discussion of research methods to consider an array of contextual factors such as the generalizability of the evidence across contexts such as cultures, points in the business cycle, rural/urban settings, etc.

[9] To ascertain how "credible" a particular statement or piece of evidence is requires judgment; it is a statement regarding how plausible/believable/convincing a, specific analysis, statement or data source is as measured by the norms of scientific evidence. On one dimension, a hierarchy of evidence is a formal tool for evaluating the credibility of evaluations/studies/research projects. For further discussion of the notion of credibility see Chapter 1 of *Statistics for Social Change* (Horwitz & Ferleger, 1988).

making policy decisions *directly*, but to help policymakers by informing them about the degree to which a particular piece of evidence should be included in the decision-making process as a function of the quality of that evidence. The evidence-based approach is probably most fully developed in the area of medicine and related health fields where very clear "hierarchies of evidence" are well established. However this approach is increasingly being applied to social sciences domains.[10] As mentioned earlier, an important and carefully developed framework in the context of education is the US Department of Education's/Institute of Education Science's "What Works Clearinghouse" (http://ies.ed.gov/ncee/wwc/). Another important project systematically evaluating the quality of social policy initiatives is the Campbell Collaboration (http://www.campbellcollaboration.org/), which is quite similar to the extremely well-known Cochrane collaboration (http://www.cochrane.org/) for health care.

We argue that hierarchies are useful tools when comparing the quality of evidence between two or more interventions, or different studies of the same intervention. Note that the proposed hierarchy theoretically includes both quantitative methods and qualitative methods but it focuses exclusively on adjudicating the quality of studies that address *whether* any impact has occurred. A hierarchy interested in adjudicating the question of *why* a particular outcome occurred would likely have a very different ranking of methodologies. For example, research designs that are excellent at identifying *whether* a program had a positive impact might well not be particularly well suited to understanding *why* it had a beneficial impact. (Some very high quality research designs for measuring impacts are referred to as "black box" methods because they give virtually no insight into what is happening inside the program. In the extreme, an approach answers one important research question very well, but ignores other important questions.) This is why a large-scale research program or evaluation addressing multiple policy questions usually draws on multiple lines of evidence. Ideally each line of evidence is of high quality judged with respect to a hierarchy of evidence relevant to the particular research question being addressed.

Some important considerations

Before we delve into a discussion of each type of study it is worth emphasizing some additional points. First, the hierarchy described in this section is intended to be used when different studies with the *same* broad research question are using *different* research approaches (Daly et al., 2007). Under this context, one is comparing and ranking on some common ground. Second, no study can be better than the data on which it is based. If the data, for example, have substantial measurement error or attrition, or do not measure variables that are directly relevant to the policy issue at hand, then there are limits to what can be learned. Econometric techniques and statistical methodologies can be used for extracting information from data and good techniques do so in an efficient and unbiased way, but they do not create information. Even an extremely sophisticated econometric technique applied to data with enormous amounts of measurement error cannot produce highly credible results. Similarly, an extremely sophisticated interview technique will not produce useful results if the set of people

[10] Examples of discussions of such hierarchies in health include: Evans (2003), Brighton et al. (2003), Daly et al. (2007), and Guyatt et al. (2008). While there are differences across proposed hierarchies, the basic structure across alternative versions is remarkably similar with the main differences relating to the subject matter to which the hierarchy is applied.

interviewed is not appropriate to the research question and/or are not knowledgeable about the topic of the interview. In other words, the information needs to be in the underlying data to begin with. There is also a need to ensure that the results are interpreted appropriately. It is not unheard of for empirical data to be misinterpreted as answering a particular policy question when in fact it is not addressing that issue at all. Third, an important caveat is that using a hierarchy of knowledge is useful in that it formalizes ideas and provides a common basis for discussion, but it is not a replacement for thinking. Judgment is always required since poorly executed studies in the highest-ranking category, or ones with inappropriate data, may be less credible than well executed studies with a research design ranked lower in the hierarchy. The relevance of the measured impacts at hand to the policy question under discussion is also always an issue.

Fourth, while the hierarchy is intended to reflect the quality of a well-executed study of each type listed, it is not intended to reflect the frequency with which studies are undertaken in practice. The most commonly employed approaches are not necessarily of the highest quality. Most studies of adult learning fall into the middle (lower middle) or lower tiers of the hierarchy.[11] This is not surprising since commissioning research that uses a design that can credibly establish causality requires additional investments that may not always be possible and low cost serendipitously or naturally occurring exogenous variation is not common. Thus it is worth emphasizing that the workhorses (most commonly used techniques) of empirical studies are simple (cross-) tabulations and ordinary least squares regressions. As such we discuss a specific regression specification, the so-called "earnings equation" (sometimes called a Mincer equation in honour of Jacob Mincer), in some detail in Appendix A.

Finally, it is worth emphasizing that because our hierarchy is designed to adjudicate among studies that estimate casual impact, the hierarchy is organized primarily around the features of a research design that are fundamental to estimating casual impact. As our discussion in this section explains, we identify two key features: a source of exogenous variation and/or a credible comparison group or counterfactual (see below for a full discussion of these terms). However, while our hierarchy ranks studies primarily on the absence or presence of these two features, we acknowledge that these features are only one aspect of a quality research design. For example, a study may employ a credible comparison group but fail to use outcomes measures that are highly correlated with the true outcomes the intervention seeks to affect. Thus our hierarchy is best understood as a way to rank the strength of evidence with respect to causal impacts assuming that other criteria for high quality research designs have been met.

Proposed hierarchy of evidence

Most fundamentally, addressing the question of what works for who requires identifying "*causal impacts*," not simply correlations. Assuming there is a need the justification for a particular government intervention/program is its associated *value added* or net benefit. If government action cannot generate a positive benefit, then the rationale for that action is called into question. At the heart of a net benefit is a causal impact. Without a positive causal impact, there can be no benefit no matter how low the cost

[11] See the aforementioned World Bank methodology handbooks, especially Khandker, Koolwal, and Samad (2010), for a detailed discussion of relevant issues.

nor how important the issue. Moreover, an impact is not simply the outcome that follows after the intervention or program. This is important because the outcomes that individuals experience are usually not the result of their program status alone (i.e., whether they were eligible to take part in the program or not, and whether they participate) but are also likely to be influenced by the characteristics of the individuals themselves and the context at the time the program is operating.[12] Deriving *impacts* requires a convincing/credible *counterfactual*. The counterfactual is an estimate of what the relevant outcome would be in the absence of the intervention or program (the treatment). It is the difference in outcomes with and without treatment that is the impact (i.e., the observed outcome minus the counterfactual outcome).[13] Consistent with established practice, our proposed hierarchy of evidence evaluates studies primarily on their ability to deliver a convincing counterfactual and thus produce valid estimates of a program's causal impact (see Table 1).

At the top of almost any hierarchy are Systematic Reviews of relevant research/evaluation findings (although they may be labelled differently in different contexts — sometimes, for example, they are called Literature Reviews). It is not the case that the quality of evidence presented by a review need be extremely good. Rather, they present the best available evidence and are stronger than any individual study, but a Review can be no better than the underlying studies that are synthesized in it.

In terms of individual studies (that is, those providing analysis of primary data rather than analysis of existing studies), it is generally accepted that the best estimates of counterfactuals have some source of "exogenous variation" in the selection into treatment (e.g., the assignment to a particular adult training program as opposed to an alternative program; or the assignment to one, as opposed to another, approach to pedagogy). Thus an exogenous source of variation is the defining characteristic of our upper tier. The word exogenous is employed since the variation in question comes from "outside the system" and is "outside of the control" of the (potential) participant. In a *randomized experiment* (sometimes called a social experiment), the random assignment process creates the source of exogenous variation. As reflected in our hierarchy of evidence, it is generally accepted that this approach is the gold standard for evaluation.

A second category of research designs that falls in this upper tier is called *natural experiments*. Like randomized experiments, natural experiments also exploit exogenous variation, but in this case variation is not manipulated by the researcher but rather the manipulation of individuals' behaviour/choices occurs "naturally."[14] The participant participates and the non-participant does not

12 These individual characteristics represent additional causal factors (or *covariates*) in determining outcomes. If individuals who participate in a program differ from individuals who do not participate in terms of important characteristics at baseline, then any observed differences in outcomes will be due both to program participation and to differences between the groups in terms of these other causal factors.

13 Note that some techniques, such as difference-in-differences applied to data from a social experiment, normally have a counterfactual generated by a control group that that is extremely obvious. Other techniques, such as instrumental variables (e.g., two-stage-least-squares) have a counterfactual that is much less obvious. In the latter case, the counterfactual is extracted from the data by the estimation technique.

14 In the economics literature the term "quasi-experiment" is sometimes employed as a synonym for "natural experiment." However, other disciplines use the phrase "quasi-experiment" to describe

participate in a particular program not exclusively because of his or her choice but because of the context (frequently the institutional setting). In practice, an individual is not typically required to participate, but the context induces or incentivizes the person to do so (or not to do so), perhaps by changing the cost of participation. Non-participants form the "comparison group" and estimates of their outcomes provide the counterfactual or an estimate of what the relevant outcome would be in the absence of treatment.

Note that while the terms "comparison group" and "control group" are sometimes used interchangeably, we recommend reserving the term control group for situations where the research design includes some type of "controlled" exogenous randomization — i.e., where there is actually some investigator determined control. The term "comparison group" can be used when the researcher does not control how individuals participate in the program in question. In practice, the literature uses these terms in a wide variety of situations suggesting substantial disagreement in the definitions of the terms "comparison" and "control" group. Readers need to be careful to ensure that they understand how an author is using these phrases. A particularly gray area involves natural experiments where typically there is not investigator driven "control" but there is a source of exogenous variation that may be controlled by, for example, some feature the institutional context, or there may be little control. Moreover, sometimes there is, for example, control regarding eligibility for a particular program (control regarding who is offered an adult training program), but no control regarding who among those offered treatment elects to participate in the program. In this case, there may well be a control group for policy questions regarding the "intention to treat," but a comparison group with a source of exogenous variation regarding the "treatment effect."

The middle tier of the hierarchy involves a range of approaches, the defining characteristic of which are the existence of a well-defined comparison group(s) usually based on observable characteristics. This tier also employs well measured and relevant outcomes (as opposed to inputs or outputs). Most analyses in this tier are referred to as *correlational studies* since they do not exploit an exogenous source of variation but still aim to (sometimes implicitly) create a comparison from which to estimate the counterfactual. In this case, the counterfactual usually takes the form of a comparison group that is comprised of individuals who "look like" those who receive the program based on observable characteristics, and are sometimes the treated individuals themselves prior to the treatment (as in a before-after comparison).[15] Causality can be inferred if one believes that selection on the observed characteristics employed in the study is credible, or at least plausible. A judgment call is required and this requires institutional knowledge to be brought to bear. Correlational designs may be cross-sectional (data collected at one point in time) or longitudinal (data collected at two or more points in

substantially different situations, including ones where there is no source of exogenous variation but only a comparison group selected based on observable characteristics. Given the tremendous diversity of definitions that appear to exist for the phrase "quasi-experiment" readers need to be very careful in interpreting it.

15 "Observable characteristics" have two limitations. First, there are always some characteristics such as motivation that are not observed in the dataset but are likely relevant to the program impact. Secondly, there are often limitations to the measures that do exist. For example, while we may know that an individual is a high school graduate we do not commonly know whether the person had very high or low marks, and/or what courses the individual took.

time). They can contribute to our knowledge base by providing descriptions of how outcomes of interest change over time as individuals begin to, or decline to, participate in a program.

For those not familiar with this terminology, this use of the word "correlational" can be misleading since such a study might well employ ordinary least squares (OLS) or some other sophisticated (e.g., propensity score matching), or less sophisticated (e.g., a simple After less Before difference in averages), framework for estimation. In the OLS context the regression coefficients are sometimes referred to as "conditional correlations".[16] That is, they represent the correlation between the dependent (Y) variable and the independent (X) variable in question, conditional on the other independent variables in the regression. Regression, in this situation, is a descriptive tool that summarizes the relationships in the data and does not to make a causal statement. The nomenclature is meant to express the basic idea that without a source of exogenous variation or some other type of causal identification (perhaps from theory and/or knowledge of the context) OLS regression coefficients and other types of analysis are telling us about observed relationships in the data.

Of course, opinions can differ regarding the credibility/plausibility of various sources of identification. In some situations one reader may choose to interpret a particular result as causal based on what is known about the context/institutions, while another reader with a different interpretation of the same information may view the same results as correlational. In both cases the causality, or lack of causality, is in the interpretation. If a reader does not feel justified in interpreting the results as causal, then it is a correlational study. Of course, a study that all readers agree is causal might well also use a regression in the formal analysis, so it is not the use, or the lack of use, of a sophisticated empirical technique that generates causality but the entire research design (data, institutional context/theory, and the estimation technique).

A specific strength of correlational studies with longitudinal designs is that researchers can use statistical techniques such as fixed effects to control for time-invariant individual characteristics that are not observed in the dataset. However it is important to note that techniques such as fixed effects do not allow researchers to control for individual characteristics such as motivation if they vary over time. In addition, unless there is a comparison group (i.e., non-participants), longitudinal designs are limited in that they cannot account for any wider trends in outcomes that may be occurring in society. Even with a comparison group and no time-invariant unobserved heterogeneity, however, longitudinal designs do not necessarily generate credible causal estimates of program impacts unless there is some exogenous source of variation behind program participation (in which case they are no longer correlational). For example, individuals may choose to participate in a training program because they know it will benefit them — program participation is determined by the program outcome in this case.

[16] Formally, of course, a regression coefficient is related to, but distinct from a correlation. A correlation is normalized to vary between plus and minus one, and has no units. In contrast, the value that a regression coefficient can take is not bounded, and a coefficient has units – for example, in a regression of an unadjusted measure of earnings (Y) on a linear measure of years of school (X), so that Y= a + bX + e, the regression coefficient b has units of "dollars per year" ($/yr) and measures the increased earnings associated with each additional year of education. In a correlational study this is an association (a correlation; conditional on the other variables in the regression if there are any) and it does not have a causal interpretation. The additional earnings associated with each additional year of schooling may be caused by that schooling and/or caused by higher average beneficial levels of unobserved innate ability, motivation, parental connections/encouragement, or other factors possessed by individuals with higher levels of education.

In this way, correlational longitudinal designs are distinct from longitudinal designs that draw variation in program participation from some exogenous source, which we will refer to as "difference-in-differences" (discussed below).

Case studies are another type of research design. We note that this term is used very broadly ranging from highly quantitative studies that may exploit sources of exogenous variation, to purely qualitative ones with no sense of a counterfactual. We distinguish between varieties of case studies by using additional descriptive terms. Case studies are focused on a specific site or set of sites. This is both a considerable strength in that it can allow a more in-depth and richer analysis but also a weakness in that it also limits generalizability. Single case studies typically provide insight into specific situations and to explore at a level of detail that is not feasible in a broader study. Single case studies are usually less generalizable than broader studies because the sample sizes tend to be small and the focus is very specific to one situation. Some case study designs do aim to generate findings that are generalizable by starting from a theoretical framework that has been developed from earlier research. The generalizability of a case study can also be increased by using a multi-case approach in which one can replicate the findings from one study to another. Case studies with counterfactuals or comparison groups, and which involve good outcome measures, are generally of more value in terms of estimating impacts than those without such features. Similarly, among those without a counterfactual, studies with outcome measures are typically superior to those with only input and/or output measures.

Client/participant satisfaction or evaluation surveys or focus groups can provide useful information, especially if well designed, at relatively low cost. Clear and unbiased (not "leading") questions about issues respondents are knowledgeable about can elicit valuable information and can provide input into estimating an impact if the questions are phrased so that they measure not only the outcome, but the value added, or change in outcomes that the participant believes resulted from the treatment. This can sometimes be phrased in terms of improvements that could be made. As is well known in this area of research, care needs to be taken to minimize, or at least recognize, any biases that may exist. Similar to surveys of participants, surveys of or interviews with experts can provide some measures of impacts. But, again, there is a need to be wary of potential biases, which may result from reputational or ideological sources as much, or more in some contexts, than financial ones. Also, as with participant surveys, the nature of the queries matter and obtaining information of the value added is necessary.

An important issue in evaluating the quality of evidence not discussed thus far is frequently associated with the concepts of "efficacy" and "effectiveness." An efficacious program is one that works well under ideal circumstances, whereas an effective one works well in the "real world." It is commonly taken that social experiments provide good measures of the efficacy of a program, but sometimes questions are raised about how well they measure effectiveness. In contrast, natural or quasi-experiments are sometimes thought to provide better measures of effectiveness, although they are more opportunistic and can be narrower in scope. Related to this are caveats about short versus long run impacts, and changes in the quality of delivery over time as managers and program workers gain experience and expertise making a particular estimate less relevant in the future. These are all issues that are relevant to the context and interpretation of research results that need to be considered, but are not always well addressed in a simple hierarchy of evidence. Nevertheless, the hierarchy provides a point of departure in adjudicating across alternative pieces of evidence.

With these caveats in mind, we now turn to more detailed discussion of various approaches in each tier. The focus here is on the intuition behind the approaches, as opposed to the detailed mathematics behind the techniques. More extensive practical discussions of these approaches, together with tips for application in the context of social policy, are provided by two World Bank training manuals by Khandker, Koolwal, and Samad (2010), and Gertler, Martinez, Premand, Rawlings, and Vermeersch (2011). A relatively non-technical textbook discussion of some of the issues in a broad context from a sociology/psychology perspective is provided by Rossi, Freeman, and Lipsey (2004), a more technical treatment in the economics literature is by Imbens and Wooldridge (2009), and Smith and Sweetman (2010) provide a discussion focused on government action for a developed economy. Older, but very useful general discussions are by Angrist and Krueger (1999), and a discussion focussing on formal education, but with many elements that are relevant to adult education, is by Card (1999).

Table 1 Proposed hierarchy of evidence for research designs estimating causal impacts

Type	Design features	Evidence quality
Systematic reviews and meta-analysis	Use established approach to synthesize all quality research evidence (esp., "upper tier" studies) on a specific issue.	Strongest evidence but only as strong as underlying evidence.
Upper Tier – Individual studies with randomization/credible source of exogenous variation		
Randomized experiments	Well-designed with sufficient sample size.	Very strong evidence.
Natural experiments	High quality source of exogenous variation generating comparison group that provides credible approach to estimating counterfactual. Well-designed pre-post measures of outcomes and well-measured and appropriate data with large sample. Employs techniques such as regression discontinuity, instrumental variables, difference-in-differences or propensity score matching.	Very strong evidence if the source of exogenous variation is credible and if appropriate econometric/ statistical technique(s) is (are) employed to extract the information from the data.
Middle Tier – Limited or no source of exogenous variation, but with credible comparison group/counterfactual		
Some control in the assignment of treatment	Limited source of exogenous variation or some control of selection process (e.g., program administrator, perhaps non-randomly, assigns treatment; different sites follow different procedures; or individuals select into limited range of options). Well-designed pre-post outcome measures; dynamic pre-treatment measures; well measured, appropriate data with large sample. Employs techniques such as difference-in-differences and/or propensity score matching, or an appropriate regression technique.	Studies is this tier produce evidence that ranges from *very strong* and *strong* to *moderate* depending on specific design features. All other things being equal, studies with some control in assignment of treatment are generally ranked higher than studies without control.
Correlational studies including studies relying on selection on observables and case studies with a comparison group	Reasonable approach to estimating counterfactual; well-designed pre-post measures of outcomes; large sample and rich set of covariates. Quality of the comparison group is critical. Employ techniques such as difference-in-differences; population correlation designs; propensity scoring matching, hierarchical linear modeling, structural equation modeling, and OLS regression. Longitudinal designs may use techniques like fixed effects.	Studies without any exogenous variation but with a credible comparison group/counterfactual generally produce evidence that ranges from *very strong* and *strong* to *moderate* depending on specific design features.
Lower Tier – Studies without measured comparison groups/counterfactuals		
Studies without comparison group	Credible case selection, explicit causal logic model and analytical strategy, understanding of the process, quality outcome measures.	Evidence should be considered suggestive.
Participant satisfaction	Collect feedback from participants on value of intervention. Better quality studies ask about "value added" or change in relevant outcomes following from treatment, rather than only eliciting measures or opinions regarding satisfaction, inputs, outputs, processes, or outcomes.	Care needs to be taken to understand potential biases and interpret the findings accordingly.
Expert opinions	Respected organizations or individuals, explicit rationale for opinion.	
Exploratory case studies	Less credible/explicit: case selection criteria, theory of change, analytical strategy or outcome measures. Does not have good quality (or any) outcome measures. May rely on measures of inputs or outputs.	Evidence should be considered suggestive.

***Caveat emptor*: Some studies claim, a high quality source of exogenous variation or rich covariates for matching, but the reader must decide for herself whether these claims are credible. A study using an instrumental variable approach with a poor source of exogenous variation may be best discarded. Even the most sophisticated technique cannot extract information from data that is not there to start with.**

Systematic reviews

In systematic reviews, researchers attempt to gather relevant evaluative studies, critically appraise them, and come to judgments about what works using explicit, transparent, state-of-the-art methods (Davies, 1999; Nutley, Davies, & Tilley, 2000). In contrast to traditional syntheses, a systematic review will include detail about each stage of the decision process, including the question that guided the review, the criteria for studies to be included, and the methods used to search for and screen evaluation reports. It will also detail how analyses were done and how conclusions were reached. The main advantage of this type of review is that it provides a reliable and comprehensive statement about what works based on the full body of underlying research. Since the review draws on numerous studies it is better than any single study on its own, but it cannot answer questions not well answered in the literature on which it relies. However, by drawing attention to the underlying nature of specific conflicts in findings, systematic reviews can go beyond generic claims such as "more research is needed" to provide a specific research agenda. Reviews also ensure that evaluations — which may have been long forgotten — are consistently used. Of course, as Petrosino and Lavenberg (2007) point out the final statement of a review may be "we know little or nothing — proceed with caution." This can guide funding agencies and researchers toward an agenda for a new generation of evaluation studies. This can also include feedback to funding agencies where additional studies would be critical to implement.

Randomized social experiments

Experimental methods are generally viewed as the gold standard in social program evaluation. The defining characteristic of a randomized experiment is the use of a random assignment design by which participants in the research project are assigned at random to a treatment group that is eligible to receive the intervention being tested or to a control group that is not eligible.[17] This is the only method that is guaranteed to eliminate selection bias and thereby produce unbiased estimates of program impacts in large samples. The process of random assignment ensures that there are no systematic differences between the program group and the comparison group. Random assignment eliminates this form of selection bias by ensuring that the program and control groups are the same in terms of all characteristics — observed and unobserved, measured and unmeasured. For example, in large enough samples the groups are statistically identical in terms of their motivation to participate in the program, their demographic characteristics and past life experiences. They differ only in that one group is eligible for the program and the other is not. Therefore, any differences that are observed over time in the

[17] With respect to terminology, in an ideal social experiment random selection refers to the assignment of individuals to the experimental group from the population under study. It is relevant for external validity since random selection ensures that the experimental group looks like the relevant population and it speaks to the generalizability of the experiment. Random assignment to treatment takes individuals from the experimental group and randomly assigns each to either the treatment group, which participates in the program, or the control group, which does not receive treatment (sophisticated social experiments may have several treatments, but the basic idea is the same). Random assignment ensures that the characteristics of the treatment and control groups are the same in large samples and speaks to the internal validity of the experiment. Of random selection and random assignment, random selection is frequently the more difficult of the criteria of a social experiment to satisfy.

experiences of the two groups (and that exceed the statistical fluctuations that can occur due to chance, which decrease as the sample size increases) can be attributed with confidence to the program.[18]

Of course, judgment is always required in analyzing evidence, and an extremely good quality observational study may be superior to a randomized controlled trial that is poorly executed. Concato et al. (2000) argue that in some contexts high quality observational studies should be considered on par with randomized controlled trials, which is important given the ethical and cost issues frequently associated with randomized controlled trials. However, their argument is controversial and limited with respect to context. In the context of the US Department of Education/Institute of Education Science's "What Works Clearinghouse" (http://ies.ed.gov/ncee/wwc/) some experts make a similar argument suggesting that although they are observational rather than controlled, regression discontinuity designs should be viewed on a par with social experiments. Importantly, this is not a claim for all regression based approaches, nor even for all credible sources of exogenous variation, but only for credible regression discontinuity designs.

More generally, there are limitations to social experiments. First, random assignment may not produce equivalent treatment and control groups resulting in large standard errors when the sample is small. When there are no other options but to work with a small sample, researchers often choose to randomly assign participants to treatment and control groups after initially grouping them according to certain baseline characteristics. (This is commonly referred to as stratifying the sample, or blocking.) This approach is common in the medical literature, where sample size is often an issue, and in social sciences especially when the unit of interest is the individual but randomization can only occur for clusters of individuals such as classrooms or schools (see, e.g., Imai, King, & Nall, 2009). The advantage of this approach is that it ensures that treatment and control groups share a minimum number of characteristics and increases the power of statistical tests.

A second limitation is that individuals must agree to participate in the experiment (i.e., to join the experimental group and accept the opportunity to potentially receive the program/treatment or be assigned to the control group). They are only likely to do so if they believe they will benefit from the treatment. Since non-participants may be different, experimental impacts may not be externally valid (i.e., not applicable to those who did not agree to participate). However, among that group, the experiment is able to generate credible impact estimates because of random assignment to the

[18] Note that for the identification of a causal impact we formally need not only randomization, but also that the treatment is (almost) exclusively obtained as a result of the random assignment to treatment. Broadly speaking, if this is not the case one speaks of problems with substitution and/or dropout bias. This criterion would fail if, for example, as a result of being randomly assigned into the control group a large number of individuals consciously sought out treatment from alternative sources. This would imply that the control group is, at least partly, subject to the treatment and not really very well "controlled." This is a threat to internal validity in medical research, where individuals with a particular diagnosis involved in a medical trial who are assigned to the control group (i.e., they will not be treated) seek out alternative similar treatments for their medical condition. In pharmaceutical research this is the reason why many trials are "blind" or "double blind," but of course "blinding" participants is not normally feasible for social experiments. For those with econometric training, similar to the context of two-stage-least-squares what is required is both validity and predictive power in the first stage equation ("real" randomization or naturally occurring assignment that is "as good as" randomization, and predictive power that is not "weak"), *and* exclusion from the second stage (the instrument cannot be correlated with the error term in the second stage).

treatment or control group. This refers to internal validity (i.e., applicable to those who receive the program offer or seek treatment).

Furthermore, not all those randomly assigned to the program/treatment group opt to actually take up the program or treatment, especially if accessing the benefits the program offers involves some action from the participant, For example, in the *learn*$ave experiment, participants to the program group were offered a matching grant on their savings towards future education. While the offer to participants was the matching grant, it was still conditional on participants performing the act of saving. As discussed earlier, when measuring the impact of a program, it is common to focus on all treatment and control group members to avoid violating the intention-to-treat principle. However, it may still be useful to know what the impact of the treatment on the treated is for those who take up the treatment, or what it would be if the entire treatment group participated in the program (i.e., accepted the treatment). This speaks to the potential large scale effect of the program, which policy analysts can strive to achieve by trying to increase program take-up rates. While estimates of the treatment on the treated are sometimes available for those who elect to take up the offer (e.g., using Bloom's (1984) technique or an instrumental variables approach), unfortunately little can usually be done other than tracking the treatment take-up rates and dosage (to provide context to the impacts) to estimate the treatment's impact for those who elect not to take up the offer.

Social experiments are often criticized because they deny programs or services to the control group. Several counterarguments can be made. First, it is unclear if the program works (hence the experiment). Second, program denial leaves the control group in the same situation as before. In fact, control group members could even be compensated for not receiving the treatment, although great care must be taken to ensure that the compensation does not bias the experimental design. Third, if programs are not properly tested, taxpayers may be left to foot the bill for ineffective (or even detrimental) programs. They too should be considered.

Natural experiments[19]

In the absence of well-designed experiments, researchers can sometimes locate a context where there is both a source of exogenous variation (that mimics random assignment) and available data that measures relevant characteristics and outcomes. If such a context, and such data, can be located then there are several econometric techniques that address particular types of identification problems available to assist researchers in appropriately extracting useful information from the data. These

[19] Note that natural experiments, and the correlational studies discussed next, both fall into the category known as observational studies since there is no experimenter induced control over the assignment of individuals to treatment. However, the terminology in the literature is nonstandard and we think it more natural in this context to categorize natural experiment studies together with experimental studies and contrast them to situations where there is no source of exogenous variation. We think it is easiest to use terminology closest to its natural meaning so that natural experiments are both observational studies and have a source of exogenous variation. In accord with much of the literature, we use the term "correlational studies" for observational studies where there is no source of exogenous variation and the researcher is simply looking at correlations, sometimes conditional on covariates, between dependent and independent variables. We are, however, fully aware that the language has not yet become standardized across disciplines, and it even sometimes differs within disciplines, so readers need to be concerned to understand exactly which definition an author is employing for a particular technical term.

include regression discontinuity estimators, instrumental variable regression, and difference-in-differences approaches. The idea behind these techniques is to try to "mimic" the conditions of an experimental design. Researchers do this by searching for "natural experiments," or situations in nature that closely replicate the random assignment of agents into program and comparison groups.

Unfortunately, researchers who pursue natural experiments are like hunter-gatherers taking what they can find with little control over what is available. Pursuing this metaphor, in contrast researchers who pursue social experiments are more like farmers who plant seeds for particular crops in particular soil, tend those crops, and then reap the harvest. While farmers do not have control over all aspects of the cultivation process, such as the weather, they know that if they plant a particular seed that they will reap a particular crop and they have dramatically more choice and control than hunter-gatherers over what they will produce. Nevertheless, the natural experiment approach can be quite useful, since it represents a relatively inexpensive way to build up aspects of the evidence base over time. Moreover, and this is where the hunter-gatherer versus agriculturalist metaphor breaks down, natural experiments can sometimes provide better insight into the effectiveness of programs, as opposed to their efficacy, since natural experiments are observing how things occur in the "real world." Further, there appear to be a large number of sources of exogenous variation that can be exploited using appropriate data to address interesting and relevant policy questions. Given the low cost of pursuing this type of project, it makes obvious sense to "keep one's eyes open" for interesting and policy relevant situations where there is a natural experiment to be exploited to produce extremely credible information about a program even if this does imply pursuing topics provided by "nature" as opposed to selecting and controlling the topic to study.

Regression discontinuity

Regression discontinuity estimators have been described as closely mimicking the conditions of random assignment. The idea behind the approach is quite simple, and can be illustrated by the following example. Suppose a government training program is offered to firms if they are deemed to be "in need." The government's definition of "need" is based on the proportion of the firm's workforce with low educational attainment. The program is costly to implement in a firm setting because of start-up costs, so only firms where at least 40% of its workforce have low-education will qualify for the program. If the fraction of low-educated workers in a firm is 39.99%, then they do not qualify for the program.

How can a researcher exploit this situation to study the impact of training on some outcome? The answer becomes obvious by answering the following question, "In the absence of the training program, would there really be any meaningful difference between a firm with 39.99% of its workforce with low-skills compared to a firm with 40.00%?" If the answer is no, then a researcher can simply compare these two firms (and many others situated very close to the cut-off) when the program is offered. One group will have been offered the treatment (those above 40%), but not the other. However, both groups will otherwise be similar. Random assignment could probably do no better in creating equivalent groups. The counterfactual in this case comes from assuming that observations on one side of the discontinuity are similar to those on the other side, perhaps after accounting for trends in the relevant variable(s).

The key drawback of discontinuity designs is that estimates are only valid around the cut-off point. This is a very narrow segment of the population, and results from studies using this approach should be interpreted with this in mind. Interestingly, a key advantage of the discontinuity design is also that the estimate is valid around the cut-off point. In some cases the impact of the program change on individuals at, or near, the cut-off point is exactly the right impact estimate for the policy question being posed. For example, if the government is interested in introducing a slight increase (or decrease) in the age cut-off for eligibility for a particular program, then it is the impact near the current cut off that is relevant for that decision and this is exactly the estimate provided by a regression discontinuity design. As with most research designs, this approach is extremely useful for certain types of exogenous variation and sources of data, and some policy questions, and it is not very useful in other contexts.

Instrument variables

Another quasi-experimental approach is the method of instrumental variables. If we regressed earnings (or log earnings) on some measure of training and various control variables, the training coefficient would be biased because training may itself depend on earnings (i.e., it may be endogenous to earnings). But what if we identified a factor that was exogenous to earnings, but was highly correlated with training? It turns out that we could use it as a form of instrument to study the relationship between training and earnings.

Let's suppose that distance to nearest post-secondary institution serves as an instrument. This is presumably correlated with training decisions (since greater distance involves greater costs), but for the moment, assume it does not have an independent effect on earnings (other than through its role of influencing training). Rather than regressing earnings on training, researchers would regress earnings on the level of training that results from the distance to post-secondary that individuals face. Mechanically, researchers begin by regressing in a first stage the training variable on distance to school, plus all of the same control variables in the earnings equations (the second stage). Based on the first stage equation, the level of training can be "predicted" (the amount of training we would expect would be generated given the distance faced by individuals and the value of the control variables). Variation in the predicted training variable would be driven (or determined) in part by variation in distance to school. In this case the counterfactual comes from individuals living at different distances to a training center. Those who live far from the center (i.e., have high cost of attendance — the counterfactual) are compared to those who live closer (i.e., have a low cost of attendance — the treated group). In the second stage, the researcher would then regress earnings on the predicted value of training, plus the control variables. The interpretation of the training coefficient in this case would be the earnings effect of training that is driven only by variation in distance to school. In other words, if we incite people to train more or less by altering the distance they face, this is the impact we might expect among those who changed their training decisions. This is a local average treatment effect (LATE). See Appendix B for a more detailed discussion of various types of treatment effects. For policy purposes, studies that use instruments that are potentially amenable to policy are the most useful, since they indicate what might be the impact of the treatment among those incited to change their training decisions based on a change in the instrument (something policymakers may have some control over).

The distance instrument we describe may indeed be correlated with training decisions, but it may also have an independent effect on earnings (that is, an effect on earnings other that what we would expect

through its effect on training). If distance to postsecondary also has an independent effect on earnings, then it is not a valid instrument.

In general, for a variable to be a credible instrument it needs to meet two key criteria. First, it needs to have an appreciable predictive power for the treatment. That is, it cannot be a "weak instrument." If it turned out that there was very little difference in postsecondary attendance probabilities as a function of distance to the nearest postsecondary institution from one's place of residence, then the instrument would not be a very good predictor of the treatment (postsecondary attendance), and it would be a weak instrument. Fortunately, this requirement is relatively simple to test empirically, and the simplest approach is to perform an F-test of the coefficient(s) of the instrument(s) in a regression together with other relevant covariates on the treatment variable. See Imbens and Wooldridge (2009) for an extended discussion.

The second criteria for an instrument, as discussed, is that it needs to exclusively affect the assignment to treatment, and have no direct effect on the outcome. This is sometimes referred to as an "exclusion" restriction and there is no way to test it empirically. It is an identifying assumption of the model being estimated and it is up to the researcher to convince the reader of the credibility of this assumption based on such sources as program or economic theory and/or the institutional context. If the author fails to convince the reader regarding the credibility of the exclusion restriction, that is if the author cannot justify the source of exogenous variation, then the reader will not believe that the results of the analysis estimate causal impacts. To illustrate these ideas, consider the rationale behind random assignment, which can be viewed as the perfect instrumental variable. Random assignment is an extremely strong predictor of treatment, so it is not a weak instrument and it meets the first criteria. Also, by design it is perfectly correlated with treatment but not correlated with the outcome by any other means since it is random and in a properly "controlled" experiment the control group cannot access treatment by any other means, so any change in the outcome can be entirely attributed to the impact of the treatment and not to any direct effect from the instrument, so it meets the second criteria. A "good instrument" does not need to be perfectly correlated with the treatment (while it cannot be "weak" there is a large gap between perfectly correlated and weak), but it does need to be credibly excluded from having a direct effect on the final outcome independent of the treatment.

Difference-in-differences around a source of exogenous variation

A third natural experiment research design is called difference-in-differences around a source of exogenous variation. Of course, this is only a natural experiment if the first difference is around some exogenously occurring event or change of some type. It is entirely feasible to apply a difference-in-differences econometric/statistical methodology to data not containing a source of exogenous variation; in this latter case it is not a natural experiment and is likely correlational study with a well-defined comparison group (discussed elsewhere in the guide). A research design is not an econometric/statistical technique but, as mentioned, the combination of data, context and technique. The same technique applied to different data and/or in different contexts might be part of quite different research designs.

Difference-in-differences are most commonly applied in longitudinal studies although differences may be taken around other variable such as age groups in a common year. They usually provide more

credible evidence than longitudinal methods such as fixed effects models, which can be viewed as (first) difference models.[20] The intuition behind the difference-in-differences approach is quite simple, but there is an important distinction that needs to be made depending on the level of aggregation at which the source of exogenous variation operates. In studying policy changes, the source of exogenous variation is most commonly not at the level of the individual, but at the level of some identifiable aggregate, such as province of residence, EI Exhaustees in 2010, or those who initially enrolled in mechanical engineering at McGill University between 2000 and 2010. In other situations, the source of exogenous variation is at the level of the individual. Understanding the source of exogenous variation, and the level of aggregation at which it operates, is important both in estimating and interpreting the results. We focus on the common case where the source of exogenous variation is at the level of the province and the analyst exploits interprovincial policy differences.

Suppose a province introduces a training grant. Policy analysts might be interested to know if the grant helped raise training levels (and perhaps earnings). One way to do this is to calculate the difference in the outcome before and after the grant was introduced in the province, and compare it to the difference observed in provinces where the grant was not introduced. The province without the grant is used as a comparison group (some might say control group, although the researcher has no control in this example so we prefer the term comparison) to account for any background trends that might be occurring in society. Identification in a difference-in-differences context relies in large part on the comparison group having a "common trend" with the treatment group. If there is a common trend, and if the difference was greater in the province where the grant was introduced, then it is likely because of the grant.

There are several caveats in interpreting difference-in-difference estimators. First, the approach assumes that the policy change (the treatment) was randomly assigned to a certain province. That is, it assumes that the policy change is exogenous, at least with respect to individuals. This may not be the case. It is up to the analyst/researcher to convince readers regarding the exogeneity of the source of variation employed in the analysis. Exogeneity is not always the case, as programs are often initiated by government if they believe it will work in their context. A good description of the program and comparison groups is required in this case, although it is virtually impossible to convincingly argue this possibility away. Second, policies are often bundled together as part of a larger reform. If this is the case, the observed trends may be due to one or more of the policies. A full description of the policy environment is thus an important complement to a difference-in-difference analysis. Third, the aforementioned common trend assumption may be violated and any gap that emerges between provinces during the period under study may simply be a continuation of a longer trend. To rule out this possibility, researchers need to demonstrate that the program and comparison groups followed similar trends prior to the introduction of the policy. Although the common trend assumption can never be fully proven, evidence can be provided regarding its plausibility. In particular, if sufficient data are available "falsification" tests are frequently undertaken in periods adjacent to, but excluding, the period of the policy change. If a statistically significant difference-in-differences estimate is observed where one ought not to exist, then this suggests that there is a specification problem and that the

[20] Formally, a first difference model and a fixed effect model are numerically identical if there are only two periods of data. However, they differ as the number of periods of data grows.

common trend assumption is likely invalid. Fourth, standard errors (used for tests of statistical significance) need to account for the fact that the data are clustered (say, at the provincial level) and the outcomes are potentially serially correlated (correlated over time). If not taken into account, both of these issues can lead to "false negative" statistical tests — a situation where the test shows a significant result when there is in fact no real program impact (known as a type II error).

Difference-in-differences designs can be very effective if a credible comparison group can be found and the study can find an appropriate variable across which to estimate differences. One strong point of difference-in-differences designs is that falsification tests can be undertaken if data is available away from the differencing point. However, similar to the evaluation of instruments in instrumental variables designs, readers need to critically ascertain the quality of the comparison group in difference-in-differences designs.

Overall our assessment of the literature on adult learning is that the field could benefit substantially from a greater focus on research designs with natural experiments. Exploiting sources of exogenous variation to gain "glimpses at causality" should become a more central part of the research and evaluation culture. Given the relative rarity of policy changes, the challenge of course is to "spot the natural experiment."

Studies with credible counterfactuals/comparison groups based on observable characteristics

The key difference between natural experiment studies and correlational studies employing comparison groups to estimate counterfactual outcomes for the treated group is that correlational studies lack an (high quality) exogenous source of variation to exploit. Correlational studies still aim to construct a counterfactual. Correlational studies use a vast array of techniques such as fixed effects, matching estimators, hierarchical linear modeling, structural equation modeling and OLS or other types of regression. Sometimes the counterfactual is implicit rather than explicit, but it is better if the authors make the counterfactual explicit so that its quality can be more easily considered. Studies with longitudinal designs look at the change in outcomes for those individuals who undertake a program before and after treatment. The problem of course is that we do not know what the outcomes for these individuals would be in the absence of treatment.

Techniques often associated with stronger correlational designs are difference-in-differences, fixed effects (i.e., first differences), and matching estimators (frequently in combination with difference-in-differences or first difference models). For example, individuals who choose to train are followed before and after they train and their outcomes are compared to those of a benchmark group who did not train. The key is to obtain several data points before and after the training episode so that long-term outcomes may be observed. The approach allows individual fixed characteristics to be held constant however it is not able to account for unobserved heterogeneity that varies over time. As with the difference-in-difference approach, the pre-training trends are important to observe and report. The matching estimators approach or propensity score matching, involves constructing a comparison group of individuals with very similar characteristics to those in the program group. However, while individuals are matched on observable characteristics, important differences in unobserved characteristics may exist.

While there are numerous additional techniques, simple correlation designs using cross-sectional data and basic regression are in fact probably the most common approach in the research and evaluation literature. This type of design simply looks at differences in outcomes across individuals where some individuals obtain a treatment and others do not. For example, studies of differences in earnings across groups of individuals with different levels of education would fall into this category. This would include standard log earnings regressions that include years of schooling, or some such measure, as a covariate. While relatively low on our hierarchy of evidence, it is not that they do not contain useful information, but rather that estimates using this type of design should not be interpreted as causal. Appendix A provides a detailed discussion of the regression specifications that are most commonly used in estimating returns to education.

Case studies

Another approach is a case study design. This type of design may be quantitative, qualitative or both. The central characteristic of a case study is that it is narrow in scope. As discussed in the previous section, we note that this term is used very broadly and may apply to studies with a source of exogenous variation. For example, the US Department of Labor's Dislocated Worker Demonstration Project included a case study of a specific manufacturing firm in Buffalo. While the study includes an exogenous source of variation it is generally regarded as a case study because of its narrow focus on one firm in one specific geographic area (see Jones, 2011 for a review of this and other related studies). Thus for the purposes of our hierarchy we aim to distinguish between case studies with exogenous sources of variation and also case studies with comparison groups. Case studies with counterfactuals or comparison groups, and which involve good outcome measures, are generally of more value in terms of estimating impacts than those without such features. Similarly, among those without a counterfactual, studies with outcome measures are typically superior to those with only input and/or output measures.

Single case studies are designed to provide insight into a research question that has not been previously explored in much detail and/or is not feasible to explore on a larger scale. Outcomes from these studies often lay the groundwork for future research. Scientifically, single case studies are usually less credible and generalizable than other studies because the sample sizes tend to be small and the focus is very specific to one situation. The generalizability of a case study can be increased by using a multi-case approach in which one can replicate the findings from one study to another. Some case study designs aim to generate findings that are generalizable by starting from a theoretical framework that has been developed from earlier research. These studies have an explicit and detailed theory of how the program achieves its intended impact (known as "theory of change" in the literature). In addition, in order for the study to be generalizable, the sample must provide an adequate number of diverse perspectives. This means that study participants are generally included from multiple groups and/or settings. Of course, as with any research design, the generalizability of case studies is directly related to the academic rigour used while making methodological decisions such as: the selection of cases; sampling time (number of data collection points); and selection of more than one data source (interviews, documents, observation).

A specific type of case study design that is particularly relevant to adult learning is the Return on Training Investment approach developed by Kirkpatrick and Phillips (see Phillips, 1994) and popularized by practitioners such as Bailey (2007). This approach is used to evaluate the return on

investment of workplace training programs. The experiences of individual firms are analyzed in great detail, on a case by case basis. Since the data used in these studies were designed specifically for evaluating training programs, they often contain all of the details required to calculate returns on training investment. Craddock and Muttu (2010) provide an excellent review of this literature. See also Moy and McDonald, 2000. Generally speaking this approach can provide credible estimates under two conditions:

- First, the study is able to isolate effects of training from competing variables.
- Second, the study is able establish a causal chain of logic linking the training intervention to appropriate outcomes of interest.

Meeting these conditions depends on the quality of the logic model developed as well as the quality of the measures designed to test the model.

There are least four factors that make meeting these conditions challenging. The first is the same reason plaguing the econometric studies of training: training is a selective process. Second, the link between measured outcomes and actual financial returns is often not clear in the studies. Third, case studies are highly susceptible to "publication bias;" researchers may hand pick the major success stories to help them get published. More recent studies aim to address these factors by including multiple cases and selecting cases in advance of training being delivered (see CSTD, 2010 for an excellent example). Future research could create a set of guidelines for conducting studies using this type of design. This would make a significant contribution to the field.

Participant and expert surveys, interviews, and focus groups

Another type of study that does not usually use a comparison group is client/participant satisfaction or surveys or focus groups. This type of study can provide useful information, especially if well designed, at relatively low cost. Clear and unbiased (not "leading") questions about issues that respondents are knowledgeable about can elicit valuable information and can provide input into estimating an impact if the questions are phrased so that they measure not only the outcome, but the value added, or change in outcomes that the participant believes resulted from the treatment. This can sometimes be phrased in terms of improvements that could be made. As is well known in this area of research, care needs to be taken to minimize, or at least recognize, any biases that may exist. The economics literature is divided regarding how well program participants are aware of the value added of a program/treatment. On the one hand, there is concern that participants select non-randomly into programs based on expected future value. Those with a high expected benefit (high expected rate of return taking costs into account) are thought to be more likely to participate. This is one of the rationales for randomized assignment (see, e.g., Heckman, Lalonde, & Smith, 1999). On the other hand Heckman, Heinrich and Smith (2002) show that the JTPA participants were not able to estimate the impact of their training on their earnings with much accuracy. The key issue seems to be that participants are not especially knowledgeable about their counterfactual. They know their current earnings (sometimes with error), but not how much they would have earned had they not participated in the program. It remains an open question, and an area for interesting future research, to determine what types of impacts can be well estimated by participants and in what contexts.

Similar to surveys of participants, surveys of or interviews with experts can provide some measures of impacts. But, again, there is a need to be wary of potential biases, which may result from reputational or ideological sources as much, or more in some contexts, than financial ones. Also, as with participant surveys, the nature of the queries matters and obtaining information of the value added is necessary.

Conclusion

In the first section of the report we propose a cost–benefit approach as a useful overarching framework for comprehensively taking account of the full range of benefits and costs associated with adult learning interventions. We argue that even if a full cost–benefit analysis is not conducted, or even if no aspect of one is formally undertaken, the conceptual framework is extremely useful in decision-making. It imposes an intellectual discipline that allows analysts to explicitly recognize and value all the (social, financial, aesthetic, and other) costs and benefits of a program. We also argue that a cost–benefit analysis is a moot point if evidence of a causal benefit cannot be found. In the second part of this report we provide a strategy for determining whether a study provides a credible estimate of a casual impact. We present a hierarchy of evidence as systematic way of ranking research designs aim to estimate causal impacts according to their scientific validity.

Consistent with established practice, our proposed hierarchy of evidence evaluates studies primarily on their ability to deliver a convincing counterfactual and thus produce valid estimates of a program's causal impact — the difference between the observed outcome after the adult education/training program and the outcome that would have been observed had the program not been undertaken/provided. It is generally accepted that the best estimates of counterfactuals have some source of "exogenous variation" in the selection into treatment. An exogenous source of variation is the defining characteristic of our upper tier. A second category of studies do not have an exogenous source of variation but they do have a well-defined comparison group(s) usually based on observable characteristics. In this case, the comparison group is usually comprised of individuals who "look like" those who receive the program based on observable characteristics, and are sometimes the treated individuals themselves prior to the treatment (as in a before-after comparison). Causality can be inferred if one believes that selection on the observed characteristics employed in the study is credible, or at least plausible.

A third category of studies do not have a comparison group. This category includes a range of designs including some types of cases, client satisfaction surveys and expert opinion. In general, evidence for any type of study without a comparison group should be considered suggestive in terms of casual impact. However, this type of evidence may certainly make a significant contribution to our knowledge base, especially in areas where knowledge is sparse.

It is worth restating that while a hierarchy of knowledge is useful in that it formalizes ideas and provides a common basis for discussion, it is not a replacement for judgment. Poorly executed studies in the highest-ranking category, or ones based on a misunderstanding of the institutional context or using inappropriate data, may be far less credible than well executed studies with a research design ranked lower in the hierarchy.

In the companion report State of Knowledge Review of the Wider Benefits of Adult Learning, we apply this hierarchy to provide an analysis of the quality of existing empirical evidence on the impacts associated with adult learning.

Appendix A: OLS regression and earnings equations

Overview of the Mincer Earnings Equation

Undeniably, the workhorse of statistical/econometric analysis of adult education data is the earnings equation or variations on it where the dependent variable is a measure of wages, earnings, employment or some other outcome, but we focus on the first two here since they are the most common. It is sufficiently common that it is worth reviewing some of its key properties. Most commonly, it is specified as:

$$\ln(w) = b_0 + b_1 Exp + b_2 Educ + b_3 X + \varepsilon.$$

In this specification ln(w) is the natural logarithm of some measure of earnings from employment. Exp is a set (formally a vector) of variables measuring labour market experience. The vector is sometimes equivalently written using summation notation where:

$$b_1 Exp = \sum_{i=1}^{p} b_1^i Exp_i = b_1^1 Exp_1 + b_1^2 Exp_2 + \cdots + b_1^p Exp_p$$

and both *b* and *Exp* are vectors and p measures the number of terms employed and may equal one. *Educ* is similarly a vector of measures of education, and *X* represents a set of other covariates that may be included in the regression including such variables as gender, province of residence, or industry; similar to *Exp*, both *Educ* and *X* are vectors. The b terms are coefficients to be estimated, and ε is an error term that is set to have zero correlation with the variables in the regression as the "identifying assumption" used to estimate the coefficients.

Although the terminology is not universally agreed upon, if w is measured in $/hour or something very similar it is called a wage. Many economists like to use this measure since it is an estimate of the value of a worker's time and therefore accords with many economic theories, which focus on how skills and knowledge affect the value of time in the labour market. Sometimes w is a measure of employment income over an extended period comprising work and non-work time, such as a week, month or year. In this case it is common to refer to the dependent variable as earnings. (Warning: The literature is non-standard. For example, some researchers using census data, which measures annual employment earnings, will divide annual earnings by the number of weeks worked in the year and call it weekly earnings even though there is no consistency across workers in the number of hours worked per week.) Earnings do not measure the price of the worker's time in the labour market, but the product of that price times the number of hours worked in the period in question. If a training program both increases the hourly wage and reduces unemployment (or under-employment), as is common in practice, then the coefficient in a regression with earnings as the dependent (left hand side) variable will be larger than that using wages.

A closer look at the left side of the question (*w*)

Wages, earnings, earnings from self-employment, and income

An important issue in measuring wages and/or earnings is the treatment of income from net self-employment and unpaid employment, as opposed to paid employment. Some researchers focus on paid employment, some include paid employment plus positive self-employment income, and use the positive sum of paid plus self-employment earnings. One way or another, *w* cannot be negative since the natural logarithm (discussed later) is not defined for negative numbers. Unpaid employment is rarely addressed, and those, for example, doing volunteer work or working in a family owned business are commonly excluded from these analyses since it is difficult to ascertain the value of their work. A small number of studies focus explicitly on these later issues. Readers and policymakers need to be aware of how *w* is being measured to ensure that they are interpreting the results correctly. Typically, although not universally, these three factors are positively correlated and summing paid, self-employed, and (very rarely) unpaid work makes the size of the coefficient on education (discussed later) larger. Finally, note that the term "income" is usually employed to capture the broader concept of employment plus non-employment income. The latter are variously measured, but in any particular study may include government benefits such as social assistance or employment insurance payments, as well as investment income such as interest, capital gains and dividends, pension benefits, lottery winnings, and the like.

There is no "correct" measure of *w*. Rather, a researcher/evaluator must normally make do with the data that are available. When choice is an option, it is best to choose the measure (if only one may be used – alternatively multiple models may be estimated) that best matches the policy or economic question being addressed. If a training program is expected to affect both wages and employment, then earnings may be a good summary measure. However, if the higher hourly wage allows workers to cut back their hours of work as they maximize their quality of life, then (if feasible) looking at wages and hours separately may be worthwhile. One note to be aware of is that self-employment may be negative reflecting physical and financial capital investments in addition to human capital, and the increase in wages or earnings associated increased self-employment earnings reflects the joint return on human, physical and financial capital.

Logged earnings

A common question, even among experienced researchers, involves the reasons and implications of using the natural logarithm of earnings. The *reason* for the transformation is to make the error term approximately normally distributed so that the estimates of the standard errors (where the standard errors are the square root of the variances) are correct. Earnings are approximately log-normally distributed, so taking their logarithm makes the ln(w) approximately normal (not that we care about the dependent variable directly), and this transforms the error term so that it is approximately normally distributed (and we do care about the shape of the error term since if it is not what we expect, then the standard errors are incorrect and any t-tests, or other such tests, will be incorrect and we won't know if the coefficient in question is statistically significantly different from zero nor the results from any other statistical test). Some researchers, both to deal with this problem and to allow negative

values for self-employment income, use quantile regression (the generalization of median regression sometimes called least absolute deviations regression) instead since it is not bothered by these issues. However, other concerns need to be addressed in interpreting quantile regressions, but these go beyond the scope of this guide.

The *implications* of taking the logarithm of the left hand side variable speaks to interpretation. If the dependent variable were *w*, then b_1 has the interpretation of a slope coefficient. That is, it measures how much *w* increases (or decreases) when education is increased (or decreased) by some amount (e.g., the amount represented by an adult education program under study). On the other hand, if ln(w) is employed, then the coefficient has the interpretation of a percent change in w for a one unit change in the education measure. If the education measure is a zero/one (sometimes called a dummy or indicator variable), then this last is approximately true as long as the coefficient is not too far from zero.

Right hand side variables

Labour market experience (Exp)

In the above earnings equation, *Exp* measures labour market experience so as not to confound the value of this experience with that from education since they represent two different types of human capital. Normally, a measure of actual experience is not available and "potential" (sometimes called "Mincer" experience) is employed instead. Potential experience is normally defined to equal age minus years of school minus five and is best thought of as maximum labour market experience given age and education. It fits the data quite well for prime age males, but less well for females. Sometimes age is used instead of experience. This has two closely interwoven implications. First, the interpretation of the coefficient differs from that on experience in the obvious way, but second the interpretation of the coefficient on education changes in two ways: i) the coefficient is mechanically larger when potential experience is used instead of age since potential experience is a function of education (and inasmuch as real experience is similar to potential experience the same effect is observed),[21] and ii) the interpretation of the coefficient on education controlling for experience differs from that controlling for age. Experience varies across individuals according to how many years of education they have. It, therefore, takes into account the opportunity cost of the person's time in school (at the price of the earnings they would have had working as estimated by the coefficient on education). This is why the coefficient on education in a regression controlling for experience is commonly called a "return to education" or a "return to schooling." It is the increase (or decrease) in earnings/wages associated with increased (or decreased) schooling taking into account the opportunity cost of the lost value of work not done while in school. This measure of the return to education does not control for direct costs of learning, but since most observers believe that the opportunity cost to training is the largest cost it gets a large part of the way there.

[21] A trick to employ if running a regression controlling for age where the coefficient on education is only marginally statistically significant is to switch the specification and control for experience instead. This will normally increase the statistical significance a bit. Readers should be as aware of this "trick" as researchers.

Can results be interpreted as causal?

It is sometimes argued that causality can be assumed in a well specified earnings equation estimated using only correlational data since the findings are consistent with theory. This is appealing, but is not always credible. A first question is: exactly what theory it is consistent with? All of causal human capital, non-causal signalling, and non-causal sample selection (especially the so-called Roy model of selection) models have effectively the same predictions. Unless one is willing to reject signalling and sample selection models as untenable, then it is difficult to say that the results observed are causal since they agree with the human capital model. To do so is a bit like saying that one has evidence in the data supporting the human capital model, and it is credible because the researcher believes the human capital model. An alternative appeal to causality is one interpretation of some particular previous US research is that the measurement error associated with education more or less offsets the error resulting from sample selection into formal education. While this may be true in some contexts, it is not likely to extend to all contexts. For example, Canada's public post-secondary system likely has quite different selection issues compared to the mixed public and private US system (especially, our community college system is massive by comparison), and measurement error can vary dramatically across contexts. In particular, measurement error for training looking at participants in a particular program is quite different from that in general purpose surveys such as the US Current Population Survey, which serves the same purpose as the Canadian Labour Force Survey, but is designed very differently.

Specifying the education variable (Educ)

The *Educ* variable can be any of a wide variety of measures of educational participation, attainment or achievement. Participation refers to attendance, without making any claim regarding satisfactory completion; it also encompasses adult learning programs without an assessment component. Attainment implies completion – of a formal degree, or an informal program – but it normally implies some type of assessment, although the assessment may be informal as in some apprenticeships. Achievement refers to a result from some type of test or measurement of skills. For youth, the OECD's Programme for International Student Assessment (PISA) is an example of achievement, as is the OECD's Programme for the International Assessment of Adult Competencies (PIAAC) and various language benchmarks. Sometimes the *Educ* variable is a simple zero/one indicator for participation, completion or even the simple offer of a training program. In this special, but very common, case it is sometimes referred to as a treatment effect.

Appreciable recent evidence that a crucial explanation for the wide variation in the return to apparently similar measures of education is that when achievement is measured the programs do not, in fact, have at all similar achievement outcomes. This applies to formal, apparently homogeneous degrees such the attainment of a high school diploma, and also varies as a function of, for example, field of study at the post-secondary level (although even here, electrical engineering from one university need not have the same return as the same degree from a different university), and implementations of adult training programs at different sites. When achievement is measured it is found to vary dramatically across (and within) programs. Moreover, much recent work suggests that achievement explains most of the labour market benefits of education (though these last studies have looked almost

exclusively at formal education). Two types of benefits are observed to exist: that for the individual where earnings or some other outcome is measured compared to other individuals, and the rate of national economic growth (in practice the growth rate of national GDP per capita) that is found to increase more rapidly with achievement but almost not at all with attainment.

Including and interpreting control variables (X)

In considering the *X* variables, econometricians classify them as being in one of two groups. The first, termed exogenous variables, include background factors such as age, sex and parents' education. Coefficients in these regressions have "normal" interpretations and reflect the association between the variable in question and the dependent variable controlling for the other regressors. In contrast, some regressors are endogenous, and are sometimes referred to a control variables. These variables, such as industry, hours of work and the like, may have a direct relationship with the outcome, but they are also the result of other background processes (especially worker self-selection), and their coefficients normally have no clear interpretation since they are the result of a mixture of underlying processes. Rather, they are included in the regression to assist in properly estimating the coefficient on the variable in question; they are control variables rather than having coefficients that are of interest in their own right.[22]

[22] A useful and easily digestible discussion of this issue can be found in the undergraduate econometrics textbook by Stock and Watson (3rd edition).

Appendix B: Various types of impacts that can be estimated

Take the example of a study that randomly assigned workers who, on average, would have the same earnings under the training and no-training scenario, as depicted in Table 2. The identification problem is solved in this case. The fact that we only observe worker A under the training scenario and worker B under the no-training scenario is irrelevant since both are identical to each other under the same scenario. They have been assigned to separate scenarios by luck, not by their characteristics. As a result, the difference in the observed outcomes ($65,000 less $60,000) represents the true impact of training.

Table 2 Comparing earnings of trainees and non-trainees

	Earnings without training	Earnings with training	Impact of training
A (Trainee)	~~$60,000~~	$65,000	~~$5,000~~
B (Non-trainee)	$60,000	~~$65,000~~	~~$5,000~~

For policy purposes, it is extremely useful to reflect upon the impact estimated in the example above. The experiment began by offering the opportunity to receive training to individuals. Only a subset of those individuals agreed to participate in the experiment. Specifically, only those who stand to gain from the training should agree to participate. This group can be called the "treated." In the real world, they are analogous to individuals who would seek treatment (training) if it were offered. From this group (the treated), the experiment randomly assigns them into a treatment and control group.[23] The difference in outcomes is thus interpreted as the impact of training. But the impact is only valid for individuals who sought treatment.

Does this matter? The answer depends on how the government is planning on rolling out the program. If the program will be offered to individuals, then the treatment effect on the treated is precisely what the government should care about. This is because the "treated" are representative of those who would seek treatment once the program is offered to the broader population (not just experimental participants). If, however, the government plans on imposing training on individuals, then understanding the impact on those who did not seek treatment would also be useful. However, imposing a training program upon participants is probably quite challenging and not likely to happen unless it is tied to an incentive program (e.g., social assistance or employment insurance receipt). Even

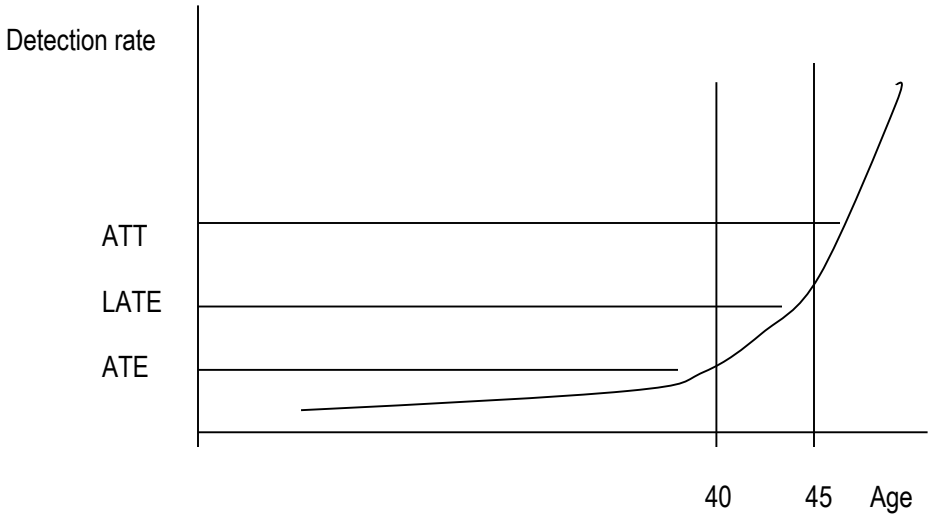
[23] We acknowledge that the terminology can be somewhat confusing here. More specifically, it is tempting to think of the treatment group as being the treated. What an experiment does is recruit a group of people who are interested in the opportunity to receive a treatment. In the real world, these people would line up to receive the treatment if it were available. In the experiment, researchers cannot give the treatment to everyone. Some people who seek treatment will be given the treatment, and others not. The treatment group is thus nothing but a randomly chosen subset of the group of individuals who are interested in receiving the treatment.

in such cases, training would only be mandatory for individuals who want to receive social benefits — some may opt out at this point.[24,25]

There is another dimension of treatment effects on the treated worthy of discussion. Suppose only recently displaced workers were offered the training opportunity. From the experiment, we have no way of knowing the impact of training on a broader population. However, governments can expand the program to other groups. For example, the government may want to know the impact on currently employed individuals with low skills. Or they may want to know the impact on all individuals. Only by expanding the experiment can these impacts be known.

An overly simplified example in the health field illustrates this point. Consider a mammography-screening program to detect breast cancer. Let's assume age is the only factor (in reality, there are many more). The curve in Figure 1 below depicts the true impact of the program by the age of women. Not surprisingly, breast cancer rates rise with age. As a result, we expect detection rates to rise as well with age (we assume 100% accuracy in detection). The impact corresponds to the detection rate and is measured along the vertical axis. Age is denoted along the horizontal axis. The policy issue is that early detection is important, but it may be costly and not worth it if the detection rate is low.

Figure 1 The true impact of a mammography-screening program by the age of women



Source: Smith and Sweetman (2001).

24 Weaker conditions have convinced some social assistance receipts to leave the system. When some provincial governments imposed that recipients pick up their cheques in person in the 1990s, welfare rolls declined suddenly (National Council of Welfare, 1997).

25 Some experiments may bypass recruitment altogether, resulting in impacts that are valid for the entire population of interest. Suppose a training grant offered to recently displaced workers was being evaluated. Researchers could begin by randomly offering the grant to a subset of recently displaced workers based on Record of Employment (ROE) slips. Linking these data to tax data would provide a convenient way to estimate the impact of offering the grant on the full population of displaced workers.

The program is initially offered to women above the age of 45. In fact, for simplicity, let's assume that the program is imposed on those women. The detection rate among this group is denoted by the horizontal line ATT. This stands for the Average Treatment effect on the Treated. For this age group, the detection rate is quite high, and the program is deemed a success. As a result, policymakers consider expanding the program. A first step is to offer the program to women between the ages of 40 and 45. Since breast cancer rates are lower among younger women, the impact among this group is lower — LATE (Local Average Treatment Effect, referring to the average impact of the program on individuals who are incited to try the program as a result of expansion). If the program is made available to all women, the impact would be even lower — ATE (Average Treatment Effect, referring to the average impact of the program on all women).

Which parameter should matter to a policy analyst: ATT, LATE, or ATE? The answer depends on the policy question. If an existing program is being evaluated, the average treatment effect on the treated (ATT) should matter. This gives us an accurate idea of the impact of the program on those who are currently receiving it. If the program is a candidate for expansion, ATT should not be viewed necessarily as the likely impact expected among newly eligible individuals. LATE is more appropriate in this case. Finally, if the policy analyst is considering making the program available to everyone, then ATE is the most useful parameter.

In our example, we have assumed that program take-up is 100%. In reality, some women may choose not to undergo mammography screening even if it were free. When evaluating the impact of offering free mammography screening, it is important to include all women who are eligible to receive the screening. In doing so, researchers isolate the "intention to treat" effect. In other words, the program was offered to a certain group of women and policy analysts are likely to be interested in the effects of offering the program (i.e., their only available policy action, unless they can make the program mandatory) to those women. If researchers only look at women who chose to participate in the program, they would violate the intention to treat principle. More concretely, they would provide policy analysts with an estimate of the impact of the program on those who took it up, ignoring the overall effect on the target (intended) group of interest. In a worst case scenario, mammography may do wonders to diagnose breast cancer among a very small percentage of women who decide to participate. This is likely due to selection effects. Tracking the overall impact of offering the program on all eligible women would reveal much smaller effects and be more useful for policy.

References

Angrist, J.D. and Krueger, A.B. (1999). "Empirical Strategies in Labour Economics." Handbook of Labour Economics, Vol. 3, Chap. 23. Orley Ashenfelter and David Card eds. (New York: North Holland).

Aos, S., Lee, S., Drake, E., Pennucci, A., Klima, T., Miller, M., Anderson, L., Mayfield, J., and Burley, M. (2011). "Return on investment: Evidence-based options to improve statewide outcomes." Document No. 11-07-1201. Olympia: Washington State Institute for Public Policy.

Bailey, A. (2007). "Connecting the dots… linking training investment to business outcomes and the economy." Ottawa: Canadian Council on Learning. Work and Learning Knowledge Centre (WLKC).

Bleakley, H., Chin, A. (2004). "Language Skills and Earnings: Evidence from Childhood Immigrants". Review of Economics and Statistics 86, 481–496.

Bleakley, H., Chin, A. (2008). "What Holds Back the Second Generation? Transmission of Language Human Capital among Immigrants". Journal of Human Resources 43, 267–298.

Blomquist, G. C., Coomes, P. A., Jepsen, C., Koford, B., and Troske, K. (2009). "Estimating the Social Value of Higher Education: Willingness to Pay for Community and Technical Colleges." IZA Discussion Paper No. 4086.

Bloom, H. S. (1984). "Accounting for No-shows in Experimental Evaluation Designs." *Evaluation Review* 8: 225-46.

Bloom, H., L. Orr, S. Bell, G. Cave, F. Doolittle, W. Lin, and J. Bos (1997). "The Benefits and Costs of JTPA Title II-A Programs: Key Findings from the National Job Training Partnership Act Study." Journal of Human Resources. 32 (3): 549-576.

Boardman, A. E., Greenberg, D. H., Vining, A. R., and Weimer, D. L. (2006). Cost–benefit Analysis: Concepts and Practice, 3rd Edition. Upper Saddle River, New Jersey: Prentice Hall.

Brighton, B., Bhandari, M., Tornetta III, P., and Felson, D. T. (2003). "Hierarchy of Evidence: From Case Reports to Randomized Controlled Trials." *Clinical Orthopaedics and Related Research*, 413: 19–24.

Burgess, D. F. and Jenkins, G. P. eds. (2010). Discount Rates for the Evaluation of Public Private Partnerships (Kingston: McGill-Queen's University Press), 338 pp.

CSTD (Canadian Society for Training and Development). (2010). Investing in People Project. Available at: http://www.cstd.ca/ResearchandResources/InvestinginPeople/tabid/81/Default.aspx?PageContentID=66.

Card, D. (1999). "The Causal Effect of Education on Earnings." Handbook of Labour Economics, Vol. 3, Chap. 30. Orley Ashenfelter and David Card eds. (New York: North Holland).

Concato, J., Shah, N., and Horwitz, R. I. (2000). "Randomized Controlled Trials, Observational Studies, and the Hierarchy of Research Designs." *New England Journal of Medicine*, 342: 1887–1892.

Craddock, T. and Muttu, A. (2010). "Return on Investment in Literacy and Essential Skills Training: Canadian and International Experience." Ottawa: Human Resources and Skills Development Canada.

Daly, J., Willis, K., Small, R., Green, J., Welch, N., Kealy, M., and Hughes, E. (2007). "A Hierarchy of Evidence for Assessing Qualitative Health Research." *Journal of Clinical Epidemiology*, 60(1): 43–49.

Davies, P. (1999). "What is Evidence-Based Education?" *British Journal of Educational Studies,* 47: 108-121.

Evans, D. (2003). "Hierarchy of Evidence: A Framework for Ranking Evidence Evaluating Healthcare Interventions." *Journal of Clinical Nursing*, 12: 77–84.

Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., Vermeersch, C. M. J. (2011). *Impact Evaluation in Practice.* Washington DC: The World Bank.

Greenberg, D., & Knight, G. (2007). "Review of the DWP Cost Benefit Framework and how it has been applied" (Vol. 40). London: UK Department for Work and Pensions Corporate Document Services.

Green, D.A. and Riddell, C. (2013) "Ageing and literacy skills: Evidence from Canada, Norway and the United States" Labor Economics 22.Greenberg, D. and Knight, G. (2007). "Review of the DWP Cost Benefit Framework and how it has been applied." Department for Work and Pensions Working Paper No. 40, Corporate Document Services.

Grogger, J., Karoly, L. A., and Klerman, J. A. (2002). "Consequences of welfare reform: A research synthesis." DRU-2676-DHHS. Santa Monica, CA: The Rand Corporation.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., and Schünemann, H. J. (2008). "GRADE: An Emerging Consensus on Rating Quality of Evidence and Strength of Recommendations." *British Medical Journal*. 336(7650): 924–926.

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

Heckman, J., Lalonde, R., and Smith, J. (1999). "The Economics and Econometrics of Active Labor Market Programs." Handbook of Labor Economics, Volume 3, Ashenfelter, A. and D. Card, eds., Amsterdam: Elsevier Science.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). "The rate of return to the HighScope Perry Preschool Program." *Journal of Public Economics*, 94(1), 114-128.

Horwitz, L. and Ferleger, L.A. (1988). *Statistics for Social Change*. Montreal: Black Rose Books.

Imai, K., King, G., and Nall, C. "The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation." Statistical Science 24 (2009): 29-53.

Imbens, G. W. and Wooldridge, J. M. (2009). "Recent Developments in the Econometrics of Program Evaluation." Journal of Economic Literature, 47(1): 5–86.

Jones, S. R. (2011). "About the Mowat Centre." Toronto: The Mowat Centre.

Khandker, S. R., Koolwal, G. B., Samad, H. A. (2010). *Handbook on Impact Evaluation: Quantitative Methods and Practices.* Washington DC: The World Bank.

Lipsey, M. W. (1997). "What Can You Build with Thousands of Bricks? Musings on the Cumulation of Knowledge in Program Evaluation." *New Directions for Evaluation,* 76: 7-23.

McConnell, S. and Glazerman, S. (June 2001). "National Job Corps Study: The Benefits and Costs of Job Corps." Submitted to U.S. Department of Labor. Mathematica Policy Research Inc.: Princeton.

Morris, P. et al. (2001) "How Welfare and Work Policies Affect Children: A Synthesis of Research", New York: Manpower Demonstration Research Corporation.

Moy, J., & McDonald, R. (2000). "Analysing enterprise returns on training." NCVER.

National Council of Welfare (1997). "Another Look at Welfare Reform." Minister of Public Works and Government Services, Canada.

Nutley, S. M., Davies, H. T. O., and Tilley, N. (2000). "Editorial: Getting Research into Practice." *Public Money and Management*, 20: 3-6.

Petrosino, A. and Lavenberg, J. (2007). "Systematic Reviews and Meta-Analyses: Best Evidence on 'What Works' for Criminal Justice Decision Makers." *Western Criminology Review*, 8(1): 1–15.

Phillips, J. J. (ed.). (1994). Measuring Return on Investment, vol. 1. Alexandria, VA: American Society for training and development.

Rogers, M. (2003). "A survey of economic growth." *Economic Record*, 79(244), 112-135.

Rossi, P., Freeman, H., and Lipsey, M. (2004). Evaluation A Systematic Approach Seventh Edition, Thousand Oaks, California: Sage.

Schocket, P. Z., Burghardt, J., and Glazerman, S. (June 2001). "National Job Corps Study: The Impacts of Job Corps on Participants' Employment and Related Outcomes." Submitted to U.S. Department of Labor. Mathematic Policy Research Inc.: Princeton.

Smith, J. and Sweetman, A. (2010). "Putting the evidence in evidence-based policy" in *Strengthening Evidence-based Policy in the Australian Federation*, Volume 1: Proceedings. Canberra: Australian Productivity Commission; 59-101.

Smith, J. and Sweetman, A. (2001). "Improving the Evaluation of Employment and Training Programs in Canada." Report for HRSDC.

Stock, J.H. and Watson, M.W. (2011). Introduction to Econometrics, 3rd Edition. Prentice Hall.

Vining, A. R. and Weimer, D. L. (2009). "General Issues Concerning the Application of Benefit-Cost Analysis to Social Policy," paper commissioned by the Benefit-Cost Analysis Center, University of Washington. As of December 1, 2010: http://evans.washington.edu/research/centers/benefit-cost-analysis/principles-andstandards-papers.

Weimer, D. L. and Vining, A. R. (eds.) (2009). Investing in the Disadvantaged: Assessing the Benefits and Costs of Social Policies. Washington, D.C.: Georgetown University Press, 17-29.