THE UNIVERSITY OF MELBOURNE

Melbourne Institute Working Paper Series

Working Paper No. 8/05

Experimental and Quasi-Experimental Methods of Microeconomic Program and Policy Evaluation

Jeff Borland, Yi-Ping Tseng and Roger Wilkins

MELBOURNE INSTITUTE
of Applied Economic and Social Research

Experimental and Quasi-Experimental Methods of Microeconomic Program and Policy Evaluation*

Jeff Borland [†‡], Yi-Ping Tseng [‡] and Roger Wilkins [‡]

[†] **Department of Economics, The University of Melbourne**

[‡] **Melbourne Institute of Applied Economic and Social Research, The University of Melbourne**

Melbourne Institute Working Paper No. 8/05

ISSN 1328-4991 (Print)
ISSN 1447-5863 (Online)
ISBN 0 7340 3185 8

June 2005

* This is a revised version of a paper prepared for Productivity Commission Workshop on 'Quantitative Tools for Microeconomic Policy Analysis'. Opinions expressed in this paper are solely those of the authors.

Melbourne Institute of Applied Economic and Social Research
The University of Melbourne
Victoria 3010 Australia
***Telephone* (03) 8344 2100**
***Fax* (03) 8344 2111**
***Email* melb-inst@unimelb.edu.au**
***WWW Address* http://www.melbourneinstitute.com**

Abstract

In this paper we review new empirical methods for evaluating microeconomic policies. Experimental and quasi-experimental evaluation measure the causal impact of a policy by comparing outcomes in the presence of the policy 'treatment' with outcomes in the absence of this treatment. For example, evaluation of a government program involves comparing outcomes associated with participation and non-participation in the program. We describe the motivation for the use of experimental and quasi-experimental methods, the types of policy effects that they can identify, and how they are implemented. Application of experimental and quasi-experimental methods is illustrated through a brief review of a variety of recent Australian studies that have evaluated microeconomic policies such as labour market programs, welfare payments policies, education policies, health policies and minimum wage laws.

1. Introduction

Recent developments in econometrics have provided a new powerful set of tools for empirical analysis and evaluation of microeconomic policies. In this paper the focus is on a tool known as experimental and quasi-experimental program evaluation. This methodology provides a variety of approaches for estimating the impact of a program or policy on participants or some other specified population. Possible examples are the effect of participation in a labour market program on subsequent employment experience of participants, or the effect of a minimum wage increase on young labour force participants.[1]

The paper provides an overview of the methodology of experimental and quasi-experimental evaluation and describes some applications that have been made to Australia. We have sought to write in a way that will make the review useful to a wide audience of policy-makers, motivated by the strong belief that the methodology does provide a powerful tool that has very wide relevance. More comprehensive or technical overviews of the methodology are available, ranging from papers that emphasise more intuitive descriptions of the methods (Meyer, 1995, Riddell, 1998, Schmidt, 1999, Blundell and Costas-Dias, 2000, Smith, 2001, and Smith and Sweetman, 2001) through to econometrically-oriented presentations (Heckman et al., 1999, Cobb-Clark and Crossley, 2003, and Imbens, 2004).

Section 2 defines the impact evaluation approach, and what is meant by experimental and quasi-experimental methods. Section 3 describes the main methods of estimating program and policy impacts. Section 4 reviews several Australian applications of experimental and quasi-experimental methods. Ideas on the way forward for program and policy evaluation in Australia are discussed in section 5.

2. Some background

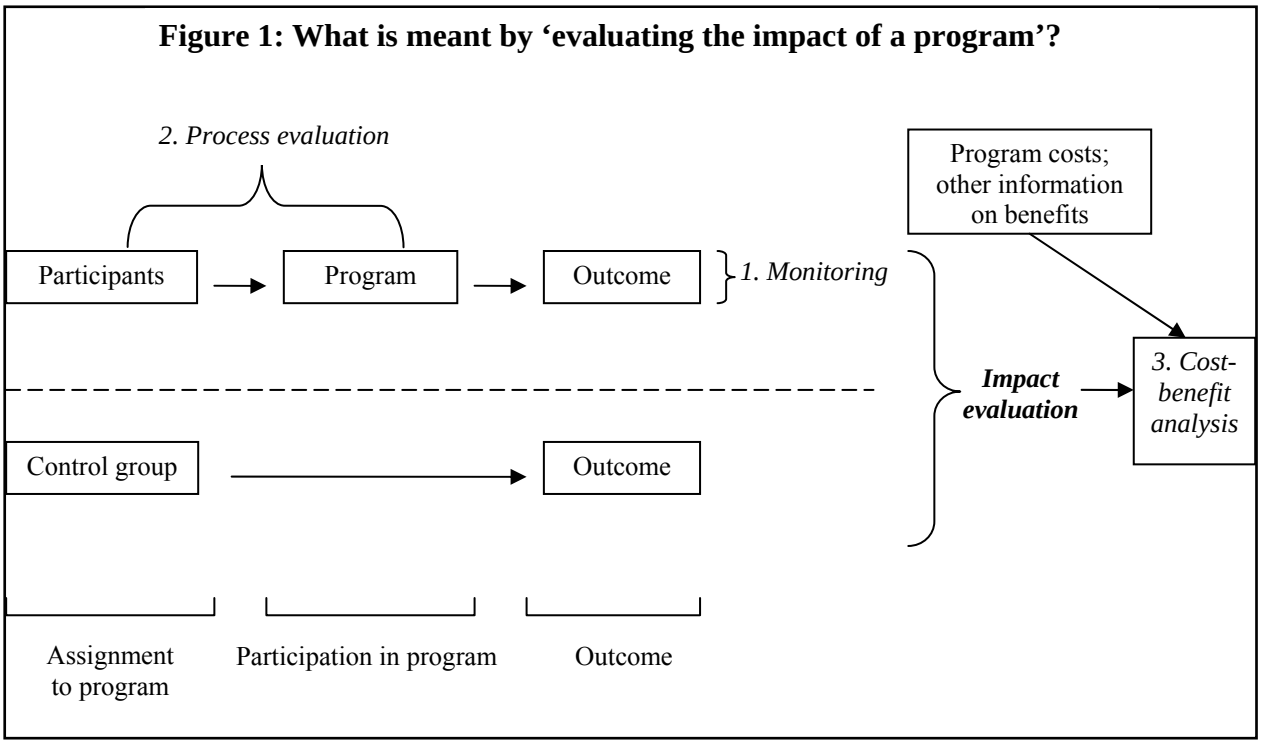
2.1. The meaning of the term 'evaluating the impact of a program'

The impact of a program is a measure of how outcomes for individuals are changed by program participation, or alternatively, the difference between what happens to a program participant compared to what would have happened had they not participated in the program.

[1] Other new tools for microeconomic analysis include experimental economics and simulation modelling that have been applied to test and predict the performance of new market designs (see, for example, Bardsley, 2003 and Stoneham et al., 2002) and behavioural micro-simulation modelling that has been applied to assess the effect of tax and transfer policies (see, for example, Creedy and Duncan, 2002).

Measuring the impact of a program or policy can be distinguished from at least three other approaches to describing or evaluating programs. Figure 1 summarises how a program impact measure is related to each of these other three approaches. The first approach, and one that is commonly applied, is to simply report outcomes for program participants. As an example, the Department of Employment and Workplace Relations provide regular reports on the proportion of labour market program participants in employment at specified time periods after completion of program participation (for example, Department of Employment and Workplace Relations, 2004). However, outcome monitoring involves only a report on employment of program participants. By contrast, an impact measure is obtained by identifying the difference between employment outcomes of participants and the outcomes that would have occurred for the same individuals in the absence of program participation. Significantly, Heckman et al., 2002 show that outcomes for labour market program participants are in general only weakly related to program impacts.

Figure 1: What is meant by 'evaluating the impact of a program'?



Second, a distinction can be made between impact evaluation of the outcomes of a program, and what is known as process evaluation that seeks to assess the operation of a program. The main objective of process evaluation is to "…provide feedback to managers on whether the program is being carried out as planned and in an efficient manner" (Riddell, 1998, p.3). It

involves describing features such as the number of program applicants and participants, the services provided and the program cost. Third, impact evaluation is a narrower exercise than cost-benefit evaluation that is intended to provide an overall measure of the net benefit to society from a program or policy. Obviously the impact of a program will be one part of the information required to measure its benefits, but the impact measure does not generally involve an attempt to *value* benefits, nor incorporate any information on program costs, both of which elements are a necessary part of cost-benefit analysis.

2.2. The approach to evaluating programs

Our objective is to measure the 'causal' impact of program participation on a specified outcome variable for an individual. To understand the approach, consider the following experiment often carried out in primary schools. We want to know the effect of sunlight on growth of plant life. We have two trays, each with cotton wool, the same amount of wheat seeds, and the same amount of water. One tray is put on a window ledge where it will receive some direct sunlight and the other tray is placed in a cupboard where it will receive no sunlight. This is an experiment that we can predict with a reasonable degree of confidence will give an estimate of the causal effect of sunlight on plant growth. This is because we have designed our experiment so that the only difference between the trays is the sunlight to which they are exposed.

A controlled experiment represents the ideal or benchmark method for measuring the impact of a program. We would like to observe the same individual both as a program participant and non-participant, or possibly two individuals who are identical in all respects except that one participates in a program and the other does not. A comparison of outcomes for the individuals between participation and non-participation would then measure the causal impact of the program.

Unfortunately, for programs which involve human beings, a controlled experiment is likely to be all but impossible. Consider the example of a labour market program where the outcome of interest is employment. It is never possible to observe simultaneously the same individual as both program participant and non-participant, while it is nigh impossible to find two individuals who are identical in all relevant characteristics, including not just readily observable features such as education, gender and age, but also characteristics such as intelligence, motivation, and labour market history.

The impossibility of measuring the causal impact of a program via a controlled comparison involving the same individual or between two identical individuals is the motivation for application of experimental and quasi-experimental methods.

In an experimental approach, individuals in a population are randomly assigned between participation and non-participation in a program, and the outcome of interest is compared between those groups. Random assignment should generate groups of participants and non-participants where each group has the same average characteristics. The comparison between the two groups can therefore be thought of as a comparison between two individuals who have the same characteristics, except for whether they are assigned to participate in the program. Comparison of outcomes for the two groups will consequently provide an estimate of the causal impact of program participation.

In a quasi-experimental approach, outcomes are compared for groups of program participants and non-participants who have not been deliberately randomly assigned. In some circumstances, although random assignment to program participation has not been an explicit feature of program implementation, treatment assignment processes have meant that it is possible to treat program participants and non-participants 'as if' they had been randomly assigned. Such a situation is known as a 'natural experiment.' The key requirement for a natural experiment is that treatment assignment be based on characteristics that are not correlated with the outcome of interest. Some well-known international examples of natural experiments include: assignment of eligibility for military service of United States citizens in Vietnam through a lottery on date of birth where the lottery constitutes a source of randomness (Angrist, 1990); inter-state differences in the timing or incidence of policies between adjacent states in the United States where the introduction of the policy can be considered random with respect to the outcome variable of interest (for example, Card and Krueger, 1994 utilise inter-state differences in timing of minimum wage increases, while Gruber, 1994 utilises inter-state differences in maternity leave provisions); restrictions on class size in schools in Israel as a source of random variation in class size (Angrist and Lavy, 1999); and birth of twins as a source of exogenous variation in family size (Bronars and Grogger, 1993).

While some programs can be treated as natural experiments, in many situations this is not possible. Often, program participants and non-participants differ in characteristics which affect outcomes, leading to 'selection effects', that is, a situation where treatment assignment is correlated with the outcome variable of interest. For example, an explicit feature of

program implementation may be to select the most disadvantaged persons into treatment. In the case of a labour market program where participation is voluntary, it might be expected that unemployed persons who think they are most likely to benefit from the program, or who have most motivation, will participate. The existence of selection effects in assignment to participation will mean that a simple comparison of outcomes between participants and non-participants will confound the effects of program participation and those other characteristics, such as personal motivation. In such cases, the task for quasi-experimental methods is to control for selection effects in order to isolate the program impact.

Selection effects can be distinguished by whether the characteristics which affect selection can be observed by the researcher. The case where these characteristics are observed is referred to as 'selection on observables'. In this situation program participants and non-participants differ on average in their outcomes in the absence of treatment only because of observable differences in characteristics. Effectively this means that there is random assignment to program participation between individuals with the same observable characteristics. Hence, a quasi-experimental approach involves controlling for these differences in characteristics to obtain valid estimates of program impact.

When 'selection on unobservables' exists, comparing outcomes of participants and non-participants controlling for observed characteristics does not produce a valid estimate of the program impact. In this situation, methods are required that attempt to control for unobservable differences in characteristics, such as the 'difference-in-differences' estimator. However, situations in which there is selection on unobservables generally require stronger assumptions for a quasi-experimental estimator to provide a valid estimate of a program effect than situations where selection effects are restricted to observable characteristics.

3. Experimental and quasi-experimental methods

3.1. Some notation

The objective is to measure the impact of an individual's participation in a program on an outcome variable. Let Y_1 denote the outcome that an individual receives if she participates in the program, and Y_0 the outcome where she does not participate (with Y_1 not observable for the individual if Y_0 is observed, and Y_0 not observable if Y_1 is observed). Hence, for individual i, the impact of program participation, Δ_i, is given by:

$$\Delta_i = Y_{1i} - Y_{0i} \quad (1)$$

Define an indicator variable D that equals 1 for individuals who participate in the program and 0 for individuals who do not participate. A vector, X, denotes variables that affect both whether an individual participates in a program and the outcome they achieve. A vector of variables known as 'instruments' that affect whether an individual participates in a program, but do not affect the outcome achieved, is denoted Z.

Consider the example of a new program in primary schools, 'A Book a Day', which is intended to increase students' reading and writing skills. Suppose that the program is only introduced in primary schools in a set of randomly chosen geographic regions in Australia, and within each school is implemented in only one 'prep' grade that is chosen by the school. In this case Y_1 and Y_0 might denote scores on a reading skill test at the end of the prep year where a student does and does not participate in the program. The vector X could include variables such as age, reading ability prior to the program, number of older siblings, and parental income. Each of these variables is a potential explanatory factor for the reading score, Y_1. As well, it is conceivable that these variables might affect a school's choice of which students to assign to the program. The vector Z could include geographic region. By definition, region of residence is a determinant of program participation. Where region has no independent effect on reading scores, it would constitute an instrument for program participation.

3.2. Impact measures

A variety of measures of program impact can be estimated. Each measure is relevant to a different policy question that might be asked about a program. Choosing the correct program impact measure for the specific policy question of interest is important where the impact of a

program differs between individual participants, since in this circumstance, the different program impact measures produce different impact estimates.

i. Average Treatment Effect (ATE)

The ATE is the mean impact of program participation for everyone in a population:

$$E(\Delta) = E(Y_1 - Y_0) \tag{2}$$

Knowing the ATE would be relevant where the policy question under consideration is whether the program should be mandatory. For the example of the 'A Book a Day' program, the question would be whether all prep students in Australia should participate in the program.

ii. Average Effect of Treatment on the Treated (AETT)

The AETT is the mean impact of program participation for current program participants:

$$E(\Delta \mid D = 1) = E(Y_1 - Y_0 \mid D = 1) \tag{3}$$

Knowing the AETT is relevant where the policy question of interest is whether the program should be continued as it currently operates. For the 'A Book a Day' program, the question would be whether the program should be maintained for the subset of students currently participating.

iii. Marginal Average Treatment Effect (MATE)

The MATE estimates the effect of program participation on a subset of the population at some margin of participation. This margin is typically conceived in terms of expanding or contracting the program. Consider, for example, the MATE in the case of program expansion:

$$E(\Delta \mid S_M \subset N) = E(Y_1 - Y_0 \mid S_M \subset N) \tag{4}$$

where S_M is a subset of program non-participants who participate when the program is expanded, and N is the group of non-participants. Knowing the MATE is relevant where the policy question of interest is: Should this program be expanded to include some extra participants? For the 'A Book a Day' program, the MATE would be relevant is the question is whether the program should be expanded to include students in one prep grade in all primary schools in Australia.

iv. Local Average Treatment Effect (LATE)

The LATE estimates the average impact on individuals who change their participation status as a result of a change in a policy instrument. An instrument refers to a change in policy on assignment that is correlated with participation but not with outcomes. For example, in the case where a policy change causes some non-participants to participate, the LATE refers to the effect of program participation on this subset of the population:

$$E(\Delta \mid S_L \subset N) = E(Y_1 - Y_0 \mid S_L \subset N) \tag{5}$$

where S_L denotes a group of individuals who switch from not participating to participating in a program due to a change in some aspect of assignment policy. Since different policy changes on assignment will in general cause different groups of individuals to be induced to participate, where program effects differ between individuals, the LATE effect may differ across different policy changes (Imbens and Angrist, 1994).

Knowing the LATE is relevant where the policy question of interest is of the form: What will be the impact of the program on an extra group of participants who are induced to participate by a change in assignment policy? For the 'A Book a Day' program, this question might be: What would be the effect of providing a $1000 grant to any other school that chose to implement the program in one prep grade? The difference between LATE and MATE is subtle, but conceptually important. The key difference is that for MATE the variable causing the change in participation need not be an instrument.

Where the impact of program participation is identical for all members of a population (that is, $\Delta_i = \bar{\Delta}$), each of the alternative measures of the average program impact will be equivalent. For example, suppose participation in the 'A Book a Day' program increases reading performance of each prep student in Australia by 10 points (on some arbitrary scale). In this case, it does not matter whether the measure of impact is estimated for the whole population of prep students, or for a subset, since in all cases it will still be found that the average effect of program participation is to increase average reading performance by 10 points. By contrast, where there is heterogeneity in program impacts between individuals, the alternative measures of average program impact will in general differ. For example, suppose that there are differences in the extent of improvement in reading performance of prep students due to participation in the 'A Book a Day' program, and assume that schools assign to program participation the grade that they expect will benefit most. In this case it might be

found that the AETT measure would show the program impact to increase reading performance by 15 points, whereas the ATE measure would show an increase of only 5 points.

Most available evidence does suggest substantial heterogeneity in program impacts across participants (see for example, Heckman, 2001a and 2001b). It is therefore always important to be cognisant of the policy question that is of interest, and to choose the impact measure that is relevant to that question.

3.3. The evaluation problem

The ideal controlled experiment, observing the same individual as both program participant and non-participant, or two individuals who are identical except for whether they participate in the program, cannot happen. The 'evaluation problem' arises because it is only ever possible to observe an individual as either a program participant or a program non-participant. Solving the evaluation problem is the motivation for application of experimental and quasi-experimental methods.

Consider again the example of the 'A Book a Day' program. For students who participate in the program the researcher observes Y_1, and for students who do not participate Y_0 is observed. Suppose the researcher was interested in estimating the AETT. This requires information on $E(Y_1|D=1)$ and $E(Y_0|D=1)$. The 'evaluation problem' is that $E(Y_0|D=1)$ is not observed.

3.4. Experiments

Experiments involve random assignment of a population between program participation and non-participation. Random assignment implies that program participation (D=1) is independent of other variables (X) that will affect outcomes. Hence for a sufficiently large population the samples of participants and non-participants can be thought of as two individuals with the same 'average' characteristics who differ only in whether they participate in the program. More formally, D is independent of the non-participation outcome Y_0, which implies that $E(Y_0|D=1) = E(Y_0|D=0)$. A comparison of average outcomes for participants ($E(Y_1|D=1)$) with average outcomes for non-participants ($E(Y_0|D=0)$) therefore provides a valid estimate of the 'causal' impact of program participation.

The main advantages of experiments are their simplicity (especially for explaining to policy-makers) and transparency of methodology. However, experiments also have disadvantages. They are not able to identify some types of program impact measures, are generally costly, and are quite difficult to implement properly. Problems of implementation include possible disruption to a program, randomisation bias, program drop-out, substitution effects, and non-cooperation of program administrators. (See Burtless, 1995 and Heckman and Smith, 1995, for more detailed discussion of these issues.)

3.5. Quasi-experimental methods

Quasi-experimental methods seek to solve the 'evaluation problem' in the absence of a randomised experiment by using data on program non-participants, or on participants at a different time, as the basis for estimating outcomes that would have occurred for participants in the absence of program participation.

i. Cross-sectional methods

Matching

The matching method estimates the program impact by comparing outcomes for program participants and non-participants in the time period(s) after the program commences. That is, it uses data on outcomes of non-participants in the period after program commencement to estimate non-participation outcomes for the group of participants. The term 'matching' is used because the comparison is made conditional on observable covariates, X, that affect both the outcome and whether individuals are assigned to the program. For example, the matching method would estimate the AETT as:

$$E(\Delta \mid D=1,\ X) = E(Y_1 \mid D=1,\ X) - E(Y_0 \mid D=0,\ X) \tag{6}$$

For the matching method to provide valid estimates of program impact it is therefore necessary that $E(Y_0 \mid D=1, X) = E(Y_0 \mid D=0, X)$. That is, conditional on the observable covariates, outcomes for the non-participants in the time period(s) after the program commences must be the same outcomes that would have occurred for participants had they not participated in the program.

More generally, a matching estimator will provide a measure of the causal impact of program participation where the following two assumptions hold:

(a) Conditional Independence Assumption (CIA): $Y_0 \perp D \mid X$; and

(b) Common Support Assumption: $\Pr(D=1|X)<1$.

The CIA requires that whether an individual participates in the program is unrelated to what their outcome would have been in the absence of program participation. Put another way, after conditioning on all covariates, assignment between program participation and non-participation is effectively random. The Common Support Assumption requires that for each program participant, there is some individual with the same (or sufficiently similar) characteristics who does not participate, and hence who can be used as the matched comparison observation.

The matching method will provide valid estimates of the causal impact of a program where there is 'selection on observables', that is, where all differences in characteristics of program participants and non-participants that affect outcomes are observable to the researcher. The selection on observables condition requires that either the basis for assignment between program participation and non-participation is a known function of observable characteristics, or the researcher can match program participants and non-participants using a sufficiently rich set of covariates to give a high degree of confidence that there are no differences in characteristics unobservable to the researchers that will affect outcomes. For example, Heckman et al. (1999) argue that for labour market programs, it is particularly important to match on the basis of local labour market region, and labour market history of program participants and non-participants.

The most basic method of matching is exact matching. With this approach program participants and non-participants are classified into 'cells' based on their characteristics. For example, where the classification is done according to gender and age (with two age groups, 16 to 34 years and 35 to 54 years), there would be four possible cells to which individuals could be assigned. The difference in average outcomes between program participants and non-participants in each cell is calculated, and the overall impact is equal to the weighted average of those cell-level effects using the fraction of program participants in each cell as weights. The main problem with exact matching approach is that, where there is a relatively large number of covariates, there will be many cells into which program participants can be classified so that, even with a large number of observations on program non-participants, the common support assumption may be violated.

The main matching approach that has been applied in the existing research is known as propensity score matching. Propensity score matching has the major advantage of overcoming

the 'curse of dimensionality' that limits the applicability of the exact matching approach. It involves matching program participants and non-participants according to an index score or predicted probability of program participation, $P(X)$. The index score is derived from an empirical model for the determinants of program participation including all matching variables as covariates. The motivation for propensity score matching is that, where the non-participation outcome is independent of assignment to treatment conditional on a set of matching covariates, the same independence condition will hold conditional on a propensity score derived from the same set of covariates. Formally, $Y_0 \perp D|X$ implies $Y_0 \perp D|P(X)$ (Rosenbaum and Rubin, 1983). Underlying this result is the idea that, although specific characteristics may differ between any single program participant and non-participant with the same $P(X)$, these differences should balance out for a sufficiently large number of observations with the same $P(X)$.

A variety of methods for matching using the propensity score can be applied. The simplest method is 'nearest neighbour', where each program participant is matched with the non-participant with the closest propensity score. The difference in outcomes between each matched pair is calculated, and the overall impact is equal to the average effect across all matched pairs. More advanced methods include kernel weighting and local linear regression (see Heckman et al., 1999 for further discussion).

For the example of the 'A Book a Day' program a matching estimator could be implemented by comparing outcomes within each school for each student in the prep grade that participates in the program with students in the other prep grades who do not participate. Variables such as age, family income and reading test score at the start of the school year could be used for matching. For the matching approach to provide valid estimates of the program impact it is necessary to believe that, conditional on this set of covariates, there are no other differences between students who participate and do not participate that will affect their reading test scores at the end of the prep year. A weighted average of the estimates of program participation across all schools included in the program will provide an estimate of the AETT.

Regression

The regression method involves a simple OLS regression of the outcome variable on the indicator for program participation (and possibly other covariates):

$$Y_i = \alpha + \beta D_i + \delta X_i + \varepsilon_i \tag{7}$$

This approach will provide a valid or unbiased estimate of the effect of program participation on the outcome variable provided there are no omitted explanatory variables from the regression model that differ in their effect on the outcome variable between program participants and non-participants. Formally, it is required that $E(\varepsilon|D=1,X)=E(\varepsilon|D=0,X)=0$. The condition on the error term in the regression model is equivalent to $E(Y_0|D=1,X)=E(Y_0|D=0,X)$. This latter condition is the same 'selection on observables' requirement for validity of the matching estimator.

The regression method has three main shortcomings compared to the matching approach. First, regression imposes a linear functional form, whereas matching, as a non-parametric estimator, does not have this restriction. Where the linear assumption does not hold, regression analysis will not provide valid estimates of the program impact. Second, where the program impact is heterogeneous between participants, the regression method produces an impact estimate that is a weighted average across participants, with weights determined by observable characteristics of participants. Due to the definition of the weights, there is no basis for believing that this regression impact estimate will correspond to the types of program impacts likely to be of interest to policy-makers, such as AET or AETT. Third, the regression method does not impose any 'common support' condition. Hence, it is possible that program impact estimates are derived from comparisons of outcomes for program participants and non-participants who differ significantly in their observable characteristics. Matching methods, by contrast, while not solving the common support problem, make explicit the common support from which the treatment effect is identified, thereby facilitating appropriate interpretation of estimates of program impacts.

Regression discontinuity

A regression discontinuity method estimates the program impact by comparing outcomes for program participants and non-participants who are respectively 'just above' and 'just below' the threshold level of some characteristic that defines eligibility for participation.

In the simplest case in which the regression discontinuity method can be applied, assignment to program participation is a deterministic but discontinuous function of some observable characteristic. Suppose for example that only some children in each prep grade at the selected primary schools will participate in the 'A Book a Day' program, and that eligibility for participation is determined by whether a child has parents with annual family income above or below $40,000. In this case, a regression discontinuity estimator would make comparisons

between children with a family income just below $40,000 (participants) and children with a family income just above $40,000 (non-participants). The motivation for the regression discontinuity estimator is that children who are 'close' on the selection variable should have similar characteristics, such that the CIA will hold. Of course, in this case the common support assumption cannot hold, and the program impact that is estimated is specific to those program participants with family incomes in the specified range.

Instrumental variables

The method of instrumental variables (IV) seeks to identify an 'instrument', which is a variable that affects program participation but has no effect on the outcome variable of interest, other than through its effect on program participation. Where such an instrument can be found, then even where an outcome may be affected by characteristics of participants and non-participants that are unobservable to a researcher (selection on unobservables), it is possible to obtain valid estimates of program impact.

In the 'A Book a Day' program, suppose that it is known that reading test scores at the end of the prep grade depend on the amount of reading done at home during the year and that this is likely to be negatively correlated with selection into the program. Where a variable measuring reading at home is unavailable to researchers, then a comparison of reading test outcomes for participants and non-participants will give a program impact estimate that is biased downwards compared to the true causal effect. One possible approach to overcoming this problem is to use an IV approach. Recall that it has been assumed that assignment of schools to the program is random between geographic regions. That is, where a student lives should be a significant determinant of their probability of program participation, but because of the feature of random assignment, region of residence should have no direct effect on reading test scores. In this case, geographic region is an 'instrument' for program participation. The comparison between students in different regions is of two groups who should have the same characteristics. This comparison will therefore give a valid estimate of the program impact.

The instrumental variable method is implemented through a two-stage process. In the first stage, the endogenous program participation variable is regressed on exogenous covariates and the instrument. In the second stage, the outcome variable is regressed on the exogenous determinants of the outcome, and the predicted values of the endogenous variable from the first stage. In the 'A Book a Day' example, the IV estimator would be implemented by estimating a first-stage regression of program participation on, amongst other variables,

geographic region, and in the second stage, by regressing reading test score on the predicted probability of program participation and other explanatory variables for reading score.

Where program impacts are the same for all individuals the IV estimate of program impact will equal the ATE and AETT, but where program impacts are heterogeneous the IV estimate is equal to the LATE. Furthermore, different instruments will in general produce different LATE estimates. This is because the IV approach estimates the average effect of program participation for a group whose status is changed from non-participation to participation by the instrument, and different instruments will in general cause different groups to switch status.

Choosing an appropriate instrument is the most important step in application of the IV method. Existing studies have generally used instruments that are derived from variation in policy or program jurisdiction or implementation (such as inter-state differences in policy regimes), from deliberate randomisation in the operation or implementation of a policy or program (for example, the draft lottery), or from economic theory of the determinants of program participation. Where changes in government policy, or differences in policy regimes across geographic regions, are used as the basis for an instrument, it is important to evaluate whether those differences are exogenous to the outcome variable of interest, or might in fact be related to the outcome that would have occurred in the absence of the policy intervention (see Besley and Case, 2000). For example, where two states have different laws on compulsory seat belts, this may not be exogenous, but might instead have arisen due to a high road death toll in the state that has introduced that law.

ii. Before/after methods

Before/after methods estimate the program impact by comparing outcomes for participants after their participation in the program with outcomes for the same group or a matched control group in the period prior to participation. This comparison can be made conditional on covariates that affect the outcome and vary between the 'before' and 'after' time periods.

Let A denote a time period after program participation, and B denote a time period prior to participation. Then, for example, the before/after method estimates the AETT as:

$$E(\Delta \mid D=1,\ X)\ = E(Y_1^A \mid D=1, X) - E(Y_0^B \mid D=1, X) \tag{8}$$

For the before/after method to provide valid estimates of the program impact it is necessary that $E(Y_0^B \mid D=1, X) = E(Y_0^A \mid D=1, X)$. That is, the outcome for participants before program

participation must be the same, conditional on the covariates, as the outcome for that group would have been after the program is implemented had they not participated in the program.

The before/after estimator with longitudinal data can be implemented through a regression model:

$$Y_{it} = \alpha + \beta D_i + \chi_i + \delta X_{it} + \varepsilon_{it};\ t = \{A,\ B\} \tag{9}$$

where χ_i represents time-invariant individual characteristics unobservable to a researcher. If the idiosyncratic error component ε_{it} is random noise, such that there are no variables unobservable to the researcher that have a systematic effect on the outcome Y, the parameter β will provide a valid estimate of the AETT. Note that variables that affect the outcome that are unobservable to a researcher that are fixed across time will be 'differenced-out' using the before/after estimator. The before/after model can be generalised to include multiple before and after time periods incorporating time trend variables (for example, Ashenfelter, 1978).

A type of before/after approach may still be implemented in the event that no data on outcomes for program participants are available in the pre-program time period, if data on another cross-section sample of outcomes from a population that is representative of program participants is available. In this case, a before/after estimator could be implemented through matching. Each program participant would be matched to control group observation(s) from the pre-program period.

The difficulty with the before/after method is controlling for all factors apart from program participation that will cause a change across time in the outcome. For example, it may be problematic to distinguish the program participation effect from the influence of macroeconomic factors or life-cycle effects, and it has often been observed that program participation is based on a transitory shock to an outcome variable (Ashenfelter's dip).

In the example of the 'A Book a Day' program, the before/after estimator would be implemented by comparing reading test scores of prep students who participate in the program in time periods before and after program participation. In a regression model using longitudinal data the comparison could seek to control for other factors that might cause test performance to change between the before and after periods such as the time of day at which the test was taken, and amount of reading practice at home. However, it may not be possible to estimate an effect of participation in the 'A Book a Day' program that adequately controls for life-cycle improvement in reading skills.

iii. Difference-in-differences method

The difference-in-differences method estimates the program impact as equal to the change in outcomes for program participants between time periods before and after program participation differenced from the change in outcomes for program non-participants between the same time periods. The method can be implemented conditional on covariates that are likely to cause different outcomes across time or between program participants and non-participants.

The difference-in-differences estimator of the AETT is:

$$\begin{aligned} E(\Delta \mid D=1,\ X) \ = & \left[E(Y_1^A \mid D=1,\ X) - E(Y_0^B \mid D=1,\ X) \right] \\ & - \left[E(Y_0^A \mid D=0,\ X) - E(Y_0^B \mid D=0,\ X) \right] \end{aligned} \tag{10}$$

The difference-in-differences estimator will produce valid estimates of the program impact where (Blundell et al., 1998):

(a) Any changes in those characteristics that are unobservable to a researcher between time periods prior to and after program implementation are the same for both program participants and non-participants; and

(b) The effect of changes in observable characteristics on the outcome variable between the time periods prior to and after program implementation is the same for program participants and non-participants.

Compared to the cross-section matching estimator the advantage of the difference-in-differences approach is that it can control for differences in unobservable characteristics between program participants and non-participants that are fixed across time (that is, a specific form of selection on unobservables). Compared to the before/after estimator, the advantage of the difference-in-differences estimator is that it can control for the influence on the outcome variable of unobservable factors that vary across time, such as life-cycle effects.

The difference-in-differences estimator has been applied most often in cases where program participants and non-participants are distinguished by being in different policy jurisdictions (for example, residing in different geographic regions), or eligibility for a policy or program is determined by some observable characteristics (such as age). For example, adjacent states may adjust minimum wages or policies on worker entitlements at different times.

The difference-in-differences approach can be implemented using a regression model. For example (Meyer, 1995):

(11) $Y_{it} = \alpha + \beta D_{it} + \gamma P_i + \theta T_t + \delta X_{it} + \varepsilon_{it}$ (Repeated cross-section)

(12) $Y_{it} = \alpha + \beta D_{it} + \theta T_t + \delta X_{it} + u_i + \varepsilon_{it}$; (Longitudinal)

where P_i is an indicator for being a program participant, and u_i represents time-invariant individual characteristics unobservable to a researcher. If the assumptions required for the difference-in-differences estimator to produce valid estimates hold, then β is a valid estimate of the AETT. The difference-in-differences approach can also be implemented using a matching approach (Blundell and Costa Dias, 2000). Program participants in the time period after implementation of the program are matched to three groups: 'participants' and 'non-participants' in the pre-program time period, and non-participants in the time period after program implementation.

3.6. Choosing the estimator

The main objective guiding a researcher's choice of estimator is to choose a method that is most likely to provide valid estimates of the program impact. Making this judgement should involve taking into account a variety of factors:

- The type of data available;
- The details of implementation and operation of the program;
- Economic theory about determinants of the outcome variable of interest and how the program is likely to affect behaviour; and
- The relative strengths and weaknesses of the different types of estimators.

The type of data available may limit the set of estimators that can be applied. For example, where data on outcomes for program participants and non-participants in the time period in which the program was implemented is available, but no data on outcomes in the time period prior to implementation of the program, then before/after or difference-in-difference estimators cannot be applied.

In some cases program rules may suggest which estimator should be chosen. For example, where program participation status is determined according to whether an individual is above or below a cut-off value of some variable (for example, Job Seeker Classification Index for participation in Job Network), a regression discontinuity design may be optimal.

Alternatively, where a program is implemented for the whole of the population in one geographic region, but not in another region, a matching or difference-in-difference approach may be optimal.

Economic theory can suggest likely determinants of both the outcome of interest and program participation. For example, job search theory might assist in thinking about what variables would affect exit from unemployment, and hence would need to be controlled for in determining how participation in a labour market program has affected the rate of exit from unemployment. The availability of data on those variables might influence whether it would be appropriate to apply a matching or difference-in-difference estimator.

Where possible, using a variety of estimators may be a sensible strategy. Application of different estimators may allow the validity of those estimators to be assessed, and provide a check on the robustness of the estimated program effect. For example, application of matching and difference-in-difference estimators may provide a check that estimates from the matching approach are not biased by time-invariant differences in outcomes for program participants and non-participants.

A guide to the validity of an estimator can often be provided through a 'pre-program specification test'. For example, outcomes for program participants and non-participants can be evaluated in the pre-program period. In the period before operation of a program there should be no program effect – hence a finding of a zero effect in that period provides support for the hypothesis of no selection effects between groups of participants and non-participants (see Heckman and Hotz, 1989, and Imbens, 2004). The validity of an estimator may also be apparent where it is possible to use data on multiple groups of participants and/or non-participants and theory provides strong guidance on how impact estimates should vary for those alternative groups. For example, a welfare policy could impose reductions in payments for individuals who move to live in a region with a higher rate of unemployment than where they currently reside. The size of reduction in payments might vary inversely with the rate of unemployment in current region of residence. Hence it would be predicted that the effect of the policy on reducing geographic mobility would be larger in regions with lower rates of unemployment. Alternatively, events such as announcement of a policy change that is not implemented or reversal of a policy change may provide an opportunity to test estimator validity (Meyer, 1995).

3.7. Limitations of experimental and quasi-experimental methods

The main shortcomings of experimental and quasi-experimental methods that have been identified concern the 'generalisability' of findings derived using these methods, and whether the findings accurately identify all effects of the program.

Results from evaluations of specific programs provide evidence on the effects of those specific types of programs, but a question exists as to whether the findings can then be used to predict effects of other programs. For example, a quasi-experimental evaluation of an earnings credit scheme that reduces income tax rates on labour market earnings by 5 percentage points for a low-income group may provide a valid estimate of the effect of that program. But there is not any immediate way in which the finding could be used to predict the effect of a scheme that reduced tax rates by 10 percentage points.

The difficulty in generalising from findings by studies using experimental and quasi-experimental methods is that these findings cannot be related back to structural models of behaviour of program participants and non-participants. They provide evidence on the effect of a program on behaviour, but not on the underlying preferences or objectives that gave rise to the behaviour. For example, Heckman (2000, p.51) has argued that:

> "…the absence of explicit structural frameworks makes it difficult to cumulate knowledge across studies conducted within this framework. Many studies produced by this research program have a "stand alone" feature and neither inform nor are influenced by the general body of empirical knowledge in economics."

In some circumstances the robustness of experimental and quasi-experimental estimates of program impacts may be a concern. The validity of these methods depends on an assumption known as 'the stable unit treatment value assumption' (SUTVA). This assumption implies that the effect of program participation on the outcome variable for an individual participant is a stable, and that outcomes for non-participants are not affected by the program. Provided these conditions hold, then a program effect estimated using a quasi-experimental method can be interpreted as the unique measure of the causal effect of participation on the participants' outcomes. But where SUTVA does not hold, then the quasi-experimental estimate cannot be interpreted in that way. For example, suppose that the 'A Book a Day' program involves diversion of resources from prep grades not participating in the program to the grade that participates. In this case, it would be expected that the quasi-experimental estimate of the impact of program participation would be positive. But it would also be expected that the

existence of the program will have a negative effect on the reading test score of non-participants. Hence, the quasi-experimental estimate of the overall effect on reading scores of prep grade students due to implementation of the program does not equal the estimated impact for program participants. One way of dealing with this problem is to use variation in the scale of implementation of a program across geographic regions, or across time, to measure the overall effect of a program on society (for example, Forslund and Krueger, 1994). An alternative approach is the development of general equilibrium models that can be applied to simulate the overall effects on society of policy changes (for example, Davidson and Woodbury, 1993, and Heckman et al., 1998).

4. Some Australian applications

4.1. Active labour market programs

i. Effect of intensive review process for very long-term unemployed – Random experiment/Matching – Breunig et al. (2003)

This study examines effects of intensive interviews and follow-up contact for persons currently unemployed who had been in receipt of income support payments for more than five years. A variety of outcome measures relating to employment, training, and social participation, are evaluated.

The intensive review program was implemented as a random experiment. However, the availability of outcome data only for participants who completed all stages of the program – which is a non-random sample of the original group of participants – meant that it was not possible to simply compare outcomes between participants and non-participants. Instead it was necessary to use matching to select a comparable group from the original group of non-participants. As well, due to problems with the implementation of the experiment – different selection rules for participant and non-participant groups, and differences in interview methods between treatment and control groups – it was necessary to restrict the sample of program participants. Ultimately, the policy effect that is estimated is the effect of full participation in the program for that segment of the group of participants aged under 50 who have a recorded phone number, and who complete all stages of the program.

The main findings from the study are that participation in the intensive review process spend on average less time working, but more time in study or training. There is no effect on job search or participation in voluntary activities. These findings are summarised in Table 1,

which presents the estimated effects of the program on both the incidence and average levels of work, job search, study/training, and voluntary work. The minimal scale of the intensive review intervention, and the severe disadvantage of participants, make it unsurprising that the program should have minimal effects (see Heckman et al., 1999, and Heckman, 1999).

Table 1: Effects of trial for very long-term unemployed – Results from intervention

	Average	Incidence
1. Weekly hours working		
Treatment	3.64	0.299
Control	5.88	0.349
Impact estimate	-2.24	-0.05
	(0.75)	(0.038)
2. Weekly hours looking for work		
Treatment	7.04	0.751
Control	7.56	0.755
Impact estimate	-0.52	-0.004
	(0.76)	(0.036)
3. Weekly hours studying or training		
Treatment	2.72	0.176
Control	1.57	0.123
Impact estimate	1.15	0.053
	(0.55)	(0.030)
4. Weekly hours voluntary work		
Treatment	1.73	0.236
Control	1.24	0.222
Impact estimate	0.49	0.014
	(0.406)	(0.035)

Note: Standard errors are in parentheses.
Source: Breunig et al. (2003, Table 3).

ii. Effect of Job Seeker Diary – Matching – Borland and Tseng (2004)

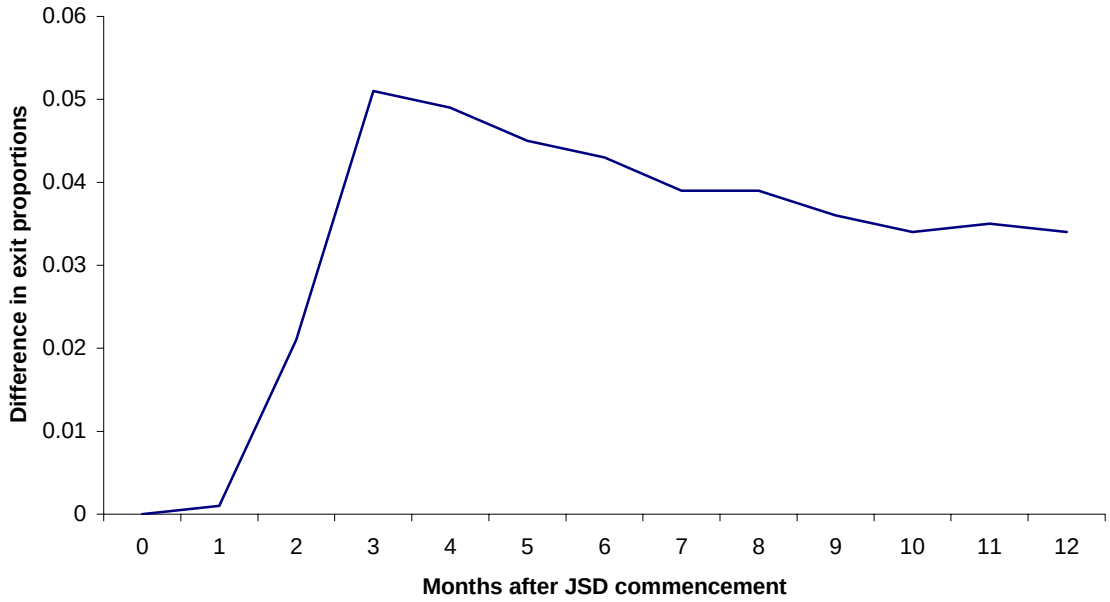
This study examines effects of participation in the Job Seeker Diary (JSD), a work search verification program that requires unemployment payment recipients to complete a fortnightly diary in which details of a specified minimum number of job applications must be recorded. Participation in the JSD occurs primarily at the start of new unemployment payment spells, and has a maximum duration of 6 fortnights.

A matching method is used to evaluate the impact of JSD participation on time spent on unemployment payments. The specific policy effect estimated is the average effect of

commencing JSD in the first fortnight of an unemployment payment spell compared to not commencing JSD participation in that fortnight or never commencing JSD participation. The sample for the study is unemployment spells of persons aged 18 to 49 years commencing in the 1997-98 financial year, excluding where possible those spells that would not have been eligible for JSD participation. The JSD program began in 1996; however the sample period examined is the earliest phase of operation of the JSD for which it is possible to identify JSD participants in the Family and Community Services administrative data set used in the study. For this sample period there are 57,779 new spells on unemployment payments (excluding ineligible spells), of which 73.4 per cent have at least one fortnight of JSD participation.

The main motivation for use of a matching method is the existence of a natural experiment for assignment between JSD participation and non-participation. During the initial phase of its operation a critical determinant of assignment to the JSD was an industrial relations dispute that meant some Centrelink offices were not assigning payment recipients to the JSD. The source of the industrial relations dispute does not appear to have been related to perceptions of the efficacy of the JSD. Furthermore, the dispute can be shown to have affected the geographic pattern of JSD participation, but in a way that is uncorrelated with local labour market conditions. Hence the industrial relations dispute introduces a source of randomness in assignment to JSD participation. Matching between JSD participants and non-participants is undertaken using a relatively rich set of covariates, most notably, income support payment history as a proxy for labour market history.

Figure 2: NSA/YA(o) payment recipients - Difference in proportion of treatment and matched control groups exiting payments by month after JSD commencement (New spells commencing July 1997 to June 1998)



Source: Borland and Tseng (2003)

Some results from the study are shown in Figure 2 and Table 2. It is found that JSD participation has a large and significant effect on exit from payments. For example, in the twelve months after commencing an unemployment payment spell, JSD participants spend on average about one fortnight less on payments than non-participants. The effect of JSD participation appears to occur primarily during the period of program participation, and qualitative evidence indicates that the effect is mainly due to increased intensity of job search. There is a high degree of heterogeneity in the program impact. About one-half to two-thirds of participants are found to benefit from JSD participation, and the effect is largest where labour demand conditions are most favourable – where a participant does not have an extensive history of payment receipt or resides in a local labour market with low rate of unemployment.

Table 2: Effects of JSD – Unemployment payment recipients aged 18 to 49 years with at least one fortnight on JSD – 'Basic' matching method – July 1997 to June 1998

	Treatment	Control	Difference	p-value
% Off payments				
By 3 months after spell commencement	36.6	31.5	+5.1	0.000
By 6 months after spell commencement	58.7	54.4	+4.3	0.000
% On payments				
At 6 months after spell commencement	49.1	53.7	-4.6	0.000
At 12 months after spell commencement	35.1	39.4	-4.3	0.000
Time on payments				
First 6 months after spell commencement	7.887	8.296	-0.409	0.000
First 12 months after spell commencement	12.958	13.888	-0.930	0.000
Number of observations				
Observations matched	39280	15643		
Total no. of observations	39287	15645		

Source: Borland and Tseng (2003, Table 7)

iii. Other studies

A variety of other studies have used experimental or quasi-experimental approaches to evaluate labour market programs in Australia. Barrett and Cobb-Clark (2001) examine a random experiment designed to test the effect of compulsion compared to voluntary participation in inducing Parenting Payment Single recipients to attend an interview at Centrelink. Richardson (2002, 2003) and Borland and Tseng (2004b) have used difference-in-difference and matching techniques to examine the effect of the Mutual Obligation Initiative on time on unemployment payments for payment recipients aged 18 to 24 years. Borland and Tseng (2004a) have evaluated the impact of the pilot phase of the 'Work for the Dole' initiative on time on payments using an exact matching approach motivated by a natural experiment in the initial assignment of WfD projects. Borland and Wilkins (2003) use a before/after method to examine the impact on exit from payments of the 9 and 12-month Intensive reviews. Findings from experimental and quasi-experimental studies of Australian labour market programs are summarised in Borland and Tseng (2004c).

An alternative approach for estimating the impact of program participation on labour market outcomes is through estimation of hazard function models for the determinants of exit from unemployment or receipt of unemployment/welfare payments. In this approach the program impact is identified as a time-varying covariate for duration – see for example Abbring and Van Den Berg, 2003, and Van Den Berg et al., 2004. In Australia, a hazard model approach to estimating program impacts has been applied to analyse effects of the Working Nation labour market programs – see Stromback and Dockery, 2000.

4.2. Education policy

i. Effect of change in years of compulsory schooling – Difference-in-difference – Ryan (2001)

During the mid-1980s the 'Early Years of School' policy in South Australia changed its junior school progression arrangements. The changes meant that an identifiable subset of students had an additional year of schooling for their age/completed grade compared with their predecessors and other students in their 'cohort'. One important consequence of the policy change was that some students who left school at the earliest possible time could now do so at a lower grade of schooling than previously.

This study examines how the 'Early Years of School' policy affected retention rates to year 12. The question is of interest as a way of testing the 'signalling' theory of education. Signalling theory implies that 'high ability' individuals acquire extra education only to signal their ability compared to 'low ability' individuals who leave school at earlier grades. Hence, the capacity of 'low ability' individuals to leave school at a lower grade than previously, would be predicted to also lower the leaving grade of some 'high ability' individuals, who can now signal their higher ability with one less year of education.

A difference-in-difference method, comparing the change in year 12 retention rates before and after the policy change, between South Australia and other states where the policy was not implemented, is used. Both a simple comparison of means, and regression approach that seeks to control for other influences on retention rates, are applied. The motivation for validity of the difference-in-difference method is that introduction of the policy change appears not to have been related to prior movements in the retention rate in South Australia, and that retention rates in South Australia and other states seem to have followed similar paths prior to the policy change.

It is found that the 'Early Years of School' Policy had a significant effect on year 12 retention rates. Table 3 presents some main results from the study. Depending on the choice of before and after time periods, and the method applied, the effect is to lower the retention rate in South Australia relative to other states by 7 to 14 percentage points. While this finding is consistent with signalling theory, it is noted that other factors – such as peer effects on school attainment – may also explain the change in retention rates.

Table 3: Estimated effect of the implementation of the *Early Years of School* policy in South Australia on adjusted retention, 1991-93 to 1997-99

	Change in retention rate
A. Difference-in-difference estimates	
SA	-9.7
Rest of Australia	0.2
Difference-in-differences	-9.9
	(-4.7)
B. Regression estimates	
With time dummy variables	-13.0
	(-6.3)
With economic variables	-13.9
	(-6.1)

Note: Standard errors are in parentheses.
Source: Ryan (2001, Table 3.3).

ii. Effect of introduction of Austudy – IV/Difference-in-difference – Dearden and Heath (1996)

Significant changes to income support payments to students were made by the Commonwealth government with the introduction of the Austudy payment in 1987. This involved large increases in the level of income support to students, especially secondary-school students. Austudy payments are means-tested on parents' income and the amount of payment depends on a child's adjusted family income, a measure that takes into account the number of dependent children under 16 years, and the number of 'eligible siblings' aged over 16 in full-time education. Receipt of Austudy allowance disqualifies a parent from receiving other payments for a child. Another important aspect of the policy change was equalising payments for youth in full-time education with unemployment payments, and introduction of means-testing on family income for unemployment payments.

This study examines the effect of introduction of Austudy on secondary school participation. A sample of individuals aged 16 to 18 years in 1989 to 1993 who were or could have been in their final two years of schooling and who were living at home is considered. One empirical approach is to estimate the effect of Austudy using an instrumental variable method. The number of siblings aged over 15 in an individual's household is used as an instrument in the first-stage regression for determinants of Austudy receipt. The predicted probability of Austudy receipt is included as an explanatory variable in the second-stage regression for determinants of secondary-school participation. The second approach is a difference-in-difference method. Differences in participation for groups affected and no affected by the policy change are compared between a post-policy change period (1989-93) and a pre-policy change period (1984-86). The treatment group is individuals estimated to have an above-median probability of receiving Austudy; and control group is individuals with below-median probability. For both time periods the total sample consists of individuals aged 16 to 18 years who were or could have been in their final two years of schooling.

Results from both empirical methods suggest that the Austudy scheme increased year 11 and 12 participation rates by 3 to 4 percentage points. This represents about 15 to 20 per cent of the total increase in participation rates that occurred between 1984-86 and 1989-93.

iii. Other studies

In a series of other papers Ryan (2003, 2004) has used the 'Early Years of School' policy as a natural experiment to study the impact of extra junior-level schooling on literacy and numeracy, and on post-education labour market performance.

4.3. Other applications

i. Effect of changes to sole parent pension (1987) – Difference-in-difference – Doiron (2004)

Several major changes were made to the sole parent pension in 1987, including a change in the definition of a dependent child, an increase in the income test 'free area', the introduction of an earnings credit and the abolition of a separate income test for rent assistance. In addition, in 1989 the Jobs, Employment and Training (JET) scheme was introduced for sole parents.

The objective of this study is to measure the impact of these changes on labour force participation and employment of sole parents. A difference-in-difference matching method that compares outcomes before (1986) and after (1990) the policy changes, and matches

female sole parents to a comparison group of married mothers, is used. The motivation for the choice of married mothers as a comparison group is that they were subject to the same changes in other family payments as sole parents, and do not have such high participation rates that it is implausible that the responsiveness of their labour supply to incentives would be relatively low (as is argued to be the case for an alternative control group, single females). However, the potential problem is that there was a substantial long-run increase in labour force participation by married females during the 1980s. This raises the possibility that changes in participation of married and single mothers would not have been the same in the absence of the policy change.

The policy changes that occurred for sole parents in the late 1980s are found to have had a quite large significant positive effect on participation and employment. Table 4 reports some main results. Difference-in-difference matching estimates show an increase of 6.8 percentage points in the employment/population rate, and 8.6 percentage points in the rate of labour force participation.

Table 4: Effect of sole parent pension reforms – Difference-in-difference matching estimates

	Change between pre- and post-policy change		
	Treatment	**Control**	**Difference-in-difference**
Current:			
Employment	0.070	0.004	0.066
	(0.038)	(0.029)	(0.049)
Hours - Total	1.435	-0.125	1.559
	(1.357)	(1.065)	(1.724)
Unemployment	0.039	0.013	0.025
	(0.020)	(0.015)	(0.024)
Participation	0.109	0.017	0.091
	(0.038)	(0.026)	(0.048)
Previous year:			
Employment - Incidence	0.115	0.014	0.101
	(0.040)	(0.029)	(0.052)
Employment - Weeks	5.409	0.725	4.684
	(1.780)	(1.431)	(2.378)

Note: Matching method is 5 nearest neighbours from control group for each treatment observation. Standard errors are in parentheses.
Source: Doiron (2004, Table 4).

ii. Effect of changes in the minimum wage in Western Australia – Difference-in-difference – Leigh (2003)

How minimum wage increases affect employment is a subject that has excited considerable interest in recent years. This study uses a difference-in-difference method that compares the change in employment in Western Australia before and after six increases in the minimum wage in that state from 1994 to 2001, with the change in employment over the same periods in the rest of Australia. Unlike other states (except Victoria for a brief period in the early 1990s), in Western Australia under the Minimum Conditions of Employment Act 1993 there is a state-specific statutory minimum wage for non-federal non-award workers. This wage is set by the Minister through regulation, and was adjusted annually on six occasions between 1994 and 2001. For each minimum wage change, the employment/population rate from three months after the wage change is subtracted from the employment/population rate three months prior to the wage change. Hence the estimated policy effect is the effect of a minimum wage increase on the rate of employment of the Western Australian population.

Table 5: Employment to population ratios before and after minimum wage rises

	Change in E/P		
	Western Australia	**Rest of Australia**	**Difference-in-difference**
Change in minimum wage:			
August 1994	0.006	0.010	-0.037
(9.29%)	(0.007)	(0.002)	(0.085)
September 1995	-0.003	0.001	-0.005
(5.31%)	(0.007)	(0.002)	(0.007)
October 1996	-0.020	-0.014	-0.006
(4.69%)	(0.007)	(0.002)	(0.007)
December 1998	-0.015	-0.001	-0.014
(3.49%)	(0.007)	(0.002)	(0.007)
March 2000	-0.008	-0.004	-0.003
(6.14%)	(0.007)	(0.002)	(0.007)
March 2001	-0.032	-0.014	-0.018
(8.80%)	(0.007)	(0.002)	(0.007)

Note: Standard errors in parentheses.
Source: Leigh (2003, Table 2).

For the individual episodes of minimum wage increases, only two of the difference-in-difference estimates of changes in the employment are significant (see Table 5). However,

aggregating across all time periods, and taking into account the size of wage change, the estimated elasticity of the employment/population rate with respect to the minimum wage is estimated to be significant and negative, equal to -0.13 in aggregate and -0.39 when attention is restricted to workers aged 15 to 24 years.

iii. Effect of 'Lifetime Health Cover' policy on take-up of private health insurance – Regression discontinuity – Palangkaraya and Yong (2004)

'Lifetime Health Cover' was introduced by the Australian government in July 2000. This policy allows private health insurance funds to vary premiums according to a member's age at entry into a fund. More specifically, anyone who joins a health fund after reaching 30 years of age is required to pay a 2% surcharge per year over age 30 without private health insurance.

This study uses a regression discontinuity design to examine how the Lifetime Health Cover affected membership of private health funds in Australia. This is done by comparing the change in the incidence of membership for single individuals between a pre-policy change period (1995) and post-policy change period (2001), using as a treatment group individuals just above the threshold age (35-39 years) and as a comparison group individuals just below the threshold age (25-29 years). A linear probability model is used to implement the regression discontinuity method. Other covariates such as income, gender, and health status, are included together with an indicator for the treatment group. The policy effect estimated is therefore the effect of introduction of the Lifetime Cover policy on a subset of the population group affected by that policy.

The main finding is that the Lifetime Health Cover increased the incidence of membership by about 7 percentage points. Between income groups the estimated impact varies from zero for low-income individuals to 17 percentage points for high-income individuals. Overall, the Lifetime Health Cover is estimated to account for about 30 to 45 per cent of the overall change in the incidence of private health fund membership that has been attributed to the set of policy changes that occurred between 1995 and 2001. As well as Lifetime Health Cover, the other main changes were the introduction of a tax levy for high-income earners not purchasing private health insurance, and the private health insurance incentive subsidy. This is an interesting finding given that previous studies had tended to attribute a much larger share of the overall effect to the Lifetime Health Cover scheme.

iv. Other studies

Many of the applications described have been to programs that affect labour market outcomes. However, the potential scope of experimental and quasi-experimental methods encompasses any type of microeconomic program or policy. As one example, Loke (2002) has examined the effect of bicycle helmet laws on the incidence of cyclist fatalities. This study uses a difference-in-difference approach that compares time periods before and after the introduction of laws making it compulsory for cyclists to wear helmets, using fatalities to pedestrians as a control for cyclist fatalities.

5. The way forward

The application of experimental and quasi-experimental methods to policy evaluation is at a very early stage in Australia compared to countries such as the United States, Canada, the United Kingdom and other European countries, or compared to its application within international agencies such as the World Bank. In these countries the benefits of extensive program and policy evaluation can be seen by looking at the knowledge gained in recent years about areas of government activity such as active labour market programs (see Heckman et al., 1999), education policy (Angrist, 2003) and health policy (see Currie and Madrian, 1999).

What then explains the relative paucity of research on program evaluation in Australia? One explanation may be a lack of knowledge about the new methods of evaluation, and it is to be hoped that, as with other international technology transfer, there is simply a lag prior to Australia adopting best-practice. Partly it may also reflect that, as a small country, Australia does not have the resources to finance the same volume of policy-oriented research as countries such as the United States. Moreover, it may be that many of the lessons from international policy evaluation apply in Australia, so it is better to learn from that research rather than re-inventing the wheel. But it does also seem that by comparison with Europe and North America, there is less commitment by government in Australia to this type of research. There is minimal government funding for program evaluation (either in-house or externally), little effort to facilitate evaluation through the way in which policies are implemented, or by data collection and dissemination, and what evaluation occurs within government departments is often not of high quality.[2]

[2] There are, however, notable exceptions. The Commonwealth Department of Family and Community Services has established a very strong record of commissioning and sponsoring evaluation-oriented research, and in seeking to facilitate research through its construction and dissemination of administrative and general purpose data sets.

In an ideal world, what would be the future of program evaluation in Australia? Fundamentally, the creation of an ideal world would need a broad commitment from a variety of areas of government activity to the value of doing policy evaluation. In this world, government would seek to implement policies in a way that would allow policy evaluation to occur, it would invest in data collection for program evaluation and be willing to release that data externally, and it would seek to support research through funding to external researchers and sponsoring in-house research for public release. With support from government providing the basis for strong growth in program evaluation research, it would be hoped that, as in other countries, private foundations would begin to support program evaluation. Eventually, the market could become large enough to support private research agencies (such as Manpower Research Development Corporation and Abt Associates in the United States) that would have program evaluation as their main area of operation, providing an alternative source of supply of program evaluation expertise.

6. References

6.1. Review articles on experimental and quasi-experimental methodologies

Besley, T. and A. Case (2000), 'Unnatural experiments? Estimating the incidence of endogenous policies', *Economic Journal*, 110, 672-94.

Blundell, R. and M. Costa Dias (2000), 'Evaluation methods for non-experimental data', *Fiscal Studies*, 21, 427-68.

Burtless, G (1995), 'The case for randomised field trials in economic and policy research', *Journal of Economic Perspectives*, 9(2), 63-84.

Cobb-Clark, D. and T. Crossley (2003), 'Econometrics for evaluations: An introduction to recent developments', *Economic Record*, 79, 491-511.

Heckman, J. (2000), 'Causal parameters and policy analysis in economics: A twentieth century retrospective', *Quarterly Journal of Economics*, ???, 45-97.

Heckman, J. (2001a), 'Micro data, heterogeneity, and the evaluation of public policy: Nobel Lecture', *Journal of Political Economy*, 109, 673-748.

Heckman, J. (2001b), 'Accounting for heterogeneity, diversity and general equilibrium in evaluating social programs', *Economic Journal*, 111, F654-F699.

Heckman, J., C. Heinrich and J. Smith (2002), 'The performance of performance standards', *Journal of Human Resources*, 37, 778-811.

Heckman, J. and J. Hotz (1989), 'Alternative methods for evaluating the impact of training programs', *Journal of the American Statistical Association*, 84, 862-74.

Heckman, J., R. Lalonde and J. Smith (1999), 'The economics and econometrics of active labour market programs', pages 1865-2097 in O. Ashenfelter and D. Card (eds) *Handbook of Labor Economics Volume 3A* (Amsterdam, Elsevier).

Heckman, J. and J. Smith (1995), 'Assessing the case for social experiments', *Journal of Economic Perspectives*, 9(2), 85-110.

Imbens, G. (2004), 'Nonparametric estimation of average treatment effects under exogeneity: A review', *Review of Economics and Statistics*, 86, 4-29.

Meyer, B. (1995), 'Natural and quasi-experiments in economics', *Journal of Business and Economics Statistics*, 13, 151-161.

Riddell, C. (1998), 'Quasi-experimental evaluation', Report prepared for Human Resources Development Canada, SP-AH053E-01-98.

Schmidt, C. (1999), 'Knowing what works: The case for rigorous program evaluation', Discussion paper No. 77, IZA.

Smith, J. (2001), 'A critical survey of empirical methods for evaluating active labor market policies', *Swedish Economic Review*, 136, 1-22.

Smith, J. and A. Sweetman (2001), 'Improving the evaluation of employment and training programs in Canada', Paper presented to Human Resources Development Canada Conference on Evaluation Methodologies.

6.2. Applications to Australia of experimental and quasi-experimental methods

Barrett, G. and D. Cobb-Clark (2001), 'The labour market plans of parenting payment recipients: Information from a randomised social experiment', *Australian Journal of Labour Economics*, 4, 192-205.

Borland, J. and Y. Tseng (2003), 'How do administrative arrangements affect exit from unemployment payments? The case of the Job Seeker Diary in Australia', Working Paper No. 27/03, Melbourne Institute, University of Melbourne.

Borland, J. and Y. Tseng (2004a), 'Does 'Work for the Dole' work?', Working Paper No. 14/04, Melbourne Institute, University of Melbourne.

Borland, J. and Y. Tseng (2004b), 'Effects of activity test arrangements on exit from payments: Mutual Obligation', mimeo, Department of Economics, University of Melbourne.

Borland, J. and Y. Tseng (2004c), 'Testing the 'Activity Test': What works and what doesn't', mimeo, Department of Economics, University of Melbourne.

Borland, J. and R. Wilkins (2003), 'Effect of activity test arrangements on exit from payments: The 9-month Intensive Review', Working Paper No. 25/03, Melbourne Institute, University of Melbourne.

Breunig, R., D. Cobb-Clark, Y. Dunlop and M. Terrill (2003), 'Assisting the long-term unemployed: Results from a randomised trial', *Economic Record*, 79, 84-102.

Commonwealth Department of Employment and Workplace Relations (2004), 'The sustainability of outcomes: Job search training, Intensive Assistance and Work for the Dole', mimeo, Evaluation and Progamme Performance Branch.

Dearden, L. and A. Heath (1996), 'Income support and staying in school: What can we learn from Australia's Austudy experiment?', *Fiscal Studies*, 17(4), 1-30.

Doiron, D. (2004), 'Welfare reform and the labour supply of lone parents in Australia: A natural experiment approach', *Economic Record*, 80, 157-76.

Leigh, A. (2003), 'Employment effects of minimum wages: Evidence from a quasi-experiment', *Australian Economic Review*, 36, 361-73.

Loke, P. (2002), 'A re-evaluation of the offsetting behaviour hypothesis: The case of Australian bicycle helmet laws', mimeo, Department of Economics, University of Melbourne.

Palangkaraya, A. and J. Yong (2004), 'How effective is 'Lifetime Health Cover' in raising private health insurance coverage in Australia? An assessment using regression discontinuity', mimeo, Melbourne Institute, University of Melbourne.

Richardson, L. (2002), 'Impact of the Mutual Obligation Initiative on the exit behaviour of unemployment benefit recipients: The threat of additional activities', *Economic Record*, 78, 406-21.

Richardson, L. (2003), 'The Mutual Obligation Initiative and the income support dynamics of young unemployment benefit recipients: An empirical analysis', Ph.D dissertation, Australian National University.

Ryan, C. (2001), 'Education: Tests of whether it enhances productivity or merely conveys information on individual productivity in the labour market', Ph.D dissertation, University of Melbourne.

Ryan, C. (2003), 'A 'causal' estimate of the effect of schooling on full-time employment among young Australians', Research Report No. 35, Longitudinal Surveys of Australian Youth, Australian Council for Educational Research.

Stromback, T. and M. Dockery (2000), 'Labour market programs, unemployment and employment hazards', Occasional Paper No. 6293.0.00.002, Australian Bureau of Statistics.

6.3. Other references

Abbring, J. and G. Van den Berg (2003), 'The nonparametric identification of treatment effects in duration models', *Econometrica*, 71, 1491-1517.

Angrist, J. (1990), 'Lifetime earnings and the Vietnam-era draft lottery: Evidence from Social Security Administration records', *American Economic Review*, 80, 313-36.

Angrist, J. (2003), 'Randomized trials and quasi-experiments in education research', *NBER Reporter*, Summer, 11-14.

Angrist, J. and V. Lavy (1999), 'Using Maimonides' rule to estimate the effect of class size on student achievement', *Quarterly Journal of Economics*, 114, 533-75.

Ashenfelter, O. (1978), 'Estimating the effect of training programs on earnings', *Review of Economics and Statistics*, 60, 47-57.

Bardsley, P. (2003), 'Missing environmental markets and the design of 'market based instruments'', Research paper No. 891, Department of Economics, University of Melbourne.

Blundell, R., A. Duncan and C. Meghir (1998), 'Estimating labour supply responses using tax policy reforms', *Econometrica*, 66, 827-61.

Bronars, S. and J. Grogger (1993), 'The socioeconomic consequences of teenage childbearing: Findings from a natural experiment', *Family Planning Perspectives*, 25, 156-62.

Creedy, J. and A. Duncan (2002), 'Behavioural microsimulation with labour supply responses', *Journal of Economic Surveys*, 16, 1-39.

Card, D. and A. Krueger (1994), 'Minimum wages and employment: A case of the fast-food industry', *American Economic Review*, 84, 772-93.

Currie, J. and B. Madrian (1999), 'Health, health insurance and the labor market', pages 3309-3416 in O. Ashenfelter and D. Card (eds.) *Handbook of Labor Economics Volume 3C* (Elsevier).

Davidson, C. and S. Woodbury (1993), 'The displacement effects of reemployment bonus programs', *Journal of Labor Economics*, 11, 575-605.

Forslund, A. and A. Krueger (1994), 'An evaluation of the Swedish active labour market policy', pages 267-98 in R. Freeman, B. Swedenborg and R. Topel (eds.) *The Welfare State in Transition* (University of Chicago Press).

Gruber, J. (1994), 'The incidence of mandated maternity benefits', *American Economic Review*, 84, 622-641.

Heckman, J. (1999), 'Policies to foster human capital', Working paper No. 7288, National Bureau of Economic Research.

Heckman, J., L. Lochner and C. Taber (1998), 'Explaining rising wage inequality: Explorations with a dynamic general equilibrium model of labor earnings with heterogeneous agents', *Review of Economic Dynamics*, 1, 1-58.

Imbens, G. and J. Angrist (1994), 'Identification and estimation of local average treatment effects', *Econometrica*, 62, 467-76.

Rosenbaum, P. and D. Rubin (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika*, 70, 41-55.

Stoneham, G., V. Chaudhri, A. Ha and L. Strappazzon (2002), 'Auctions for conservation contracts: An empirical examination of Victoria's BushTender Trial', Working Paper No. 2002-08, Melbourne Business School.

Stromback, T. and M. Dockery (2000), 'Labour market programs, unemployment and employment hazards: An application using the 1994-1997 Survey of Employment and Unemployment Patterns', Australian Bureau of Statistics, Catalogue No. 6293.0.00.002.

Van den Berg, G., B. Van der Klaauw and J. van Ours (2004), 'Punitive sanctions and the transition rate from welfare to work', *Journal of Labor Economics*, 22, 213-41.