ACCOUNTING FOR HETEROGENEITY, DIVERSITY AND GENERAL EQUILIBRIUM IN EVALUATING SOCIAL PROGRAMMES*

James Heckman

This paper considers the merits and limitations of alternative criteria proposed to evaluate social programmes in the modern welfare state and the information required to implement them when individual responses to programmes are heterogeneous. Participation in programmes is based, in part, on this heterogeneity and the programmes have general equilibrium impacts on the economy. Particular attention is devoted to the voting criterion used in modern positive political economy and the ranking of alternative social states based on anonymity postulates. Evidence on heterogeneity is presented. Conflicts among the criteria are revealed in two empirical studies. Partial equilibrium approaches are shown to be misleading.

Diversity in the interests of its constituent groups is an essential feature of the modern welfare state. The demand for publicly documented ‘objective’ evaluations of social programmes arises from a demand from rival groups in the political process. Both the efficiency and distributional consequences of social programmes attract attention. Participants in the political debate consider the consequences of government activity for themselves as well as its consequences for other groups. Since participants differ in their goals and interests, a variety of evaluative criteria are proposed for considering the suitability of government programmes. No single criterion can lay claim to primacy.¹

Traditionally, economists have used cost–benefit analysis, consumer surplus measures or social welfare functions to rank alternative policies. More recently, positive political economists have judged alternative policies by a voting criterion – whether or not the proposed policies would survive an electoral test. Social psychologists have used questionnaires to elicit preferences of participants in programmes as well as those of members of the larger society. Micro econometricians have used choice behaviour to infer preferences of participants in voluntary programmes.

This paper examines alternative criteria and the data required to implement them. In particular, attention is focused on the criteria advocated in modern welfare economics (Sen, 2000) and the voting criterion proposed in modern political economy. Some criteria are much more demanding than other criteria. In the empirical analysis presented in this paper, the alternative

* A longer version of this paper was originally prepared for an AEI conference ‘The Role of Inequality in Tax Policy’ January 21–2, 1999 in Washington, DC. I am grateful to Christopher Taber for his help in the conducting the tax simulations, and to Jeffrey Smith for his help in analysing the job training data. This paper draws on joint work with Lance Lochner, Christopher Taber and Jeffrey Smith as noted in the text. I am grateful for comments received from Jennifer Boobar, Lars Hansen, Kevin Hassett, Louis Kaplow, Steve Machin, Robert Pollak, John Roemer, Michael Rothschild and Peyton Young. This research was supported by NSF-SBR-93-21-048, NIH-HD34958-04, NSF 97-09-873 and a grant from the Russell Sage Foundation.

¹ As discussed by Porter (1995), the very definition of ‘objective’ standards is often the topic of intense political debates. See also the discussion in Young (1994).
criteria are in conflict with each other. In the empirical studies presented in this paper, one cannot resolve the conflict among the criteria by hoping that they all rank programmes in the same order.

In considering the implementation of these criteria, this paper draws on a body of research that documents considerable heterogeneity in individual responses to government programmes, in preferences for goods and in many other outcome measures. Even conditioning very finely on observables, responses to programmes are heterogeneous just as wage variability is large even conditioning on available observable measures (Heckman, 2001). This heterogeneity has profound implications for the operational status of many proposed evaluative criteria. In the presence of such heterogeneity, it becomes much more difficult to construct the counterfactual states needed to evaluate social policies.

A large body of empirical research in microeconometrics suggests that not only is the error term $U_i$, an important contributor to the variance of outcome $Y_i$ in a regression for person $i$, $Y_i = \mathbf{X}_i \beta + U_i$ where $\mathbf{X}_i$ is a vector of treatment inputs and control variables (eg education and training measures) but that the coefficients (responses to external variation in the $\mathbf{X}_i$) are variable across persons: $Y_i = \mathbf{X}_i \beta_i + U_i$. An emerging body of evidence further suggests that the outcome response (coefficients $\beta_i$) are correlated with the $\mathbf{X}_i$ (eg education, job training or other measures); see Heckman (2001), and Heckman and Vytlacil (2001). People select into government programmes based in part on their response to the programme (ie on $\beta_i$). This heterogeneity in responses to government interventions is a fundamental aspect of the diversity of a modern economy. Accounting for it affects our ability to evaluate government programmes using the different criteria.

Given that heterogeneity and diversity are central to the modern state, it is unfortunate that the methods most commonly used in evaluating its policies do not account for it. The textbook econometric policy evaluation model, due to Tinbergen (1956), Theil (1961) and Lucas (1987), constructs a social welfare function for a representative agent to evaluate the consequences of alternative social policies. In this approach to economic policy evaluation, the general equilibrium effects and efficiency aspects of a policy receive all of the attention. Heterogeneity across persons in preferences and policy outcomes are treated as second-order problems and estimates of policy effects are based on macro time series of per capita aggregates.

Standard cost–benefit analysis ignores both distributional and general-equilibrium aspects of a policy and enumerates aggregate costs and benefits of policies at fixed prices. A paraphrase of Gertrude Stein that ‘a dollar is a dollar is a dollar’ succinctly summarises the essential features of this approach (Harberger, 1971). Attempts to incorporate distributional ‘welfare weights’ into cost–benefit analysis (Harberger, 1978) have an ad hoc and unsystematic character about them. In practice, such analyses often reflect the personal preferences of the individuals conducting particular evaluations.

Access to microdata facilitates the estimation of the distributional consequences of alternative policies but these data are no panacea. Until recently,
the empirical micro literature has focused exclusively on estimating mean impacts of programmes for specific demographic groups and estimates heterogeneity in programme impacts only across demographic groups. It neglects heterogeneity in responses within narrowly defined demographic categories. Such variation is shown to be empirically important in the recent literature. Even when microdata are fully exploited to reveal heterogeneity in responses to the treatment received in programmes, they must be used in conjunction with aggregate time-series data to estimate the full general-equilibrium consequences of policies.

Unlike the macro general-equilibrium literature, the literature in modern welfare economics – see, for example, Sen (1973; 2000) – recognises the diversity of outcomes produced under alternative policies but adopts a rigid posture about how the alternatives should be evaluated, typically invoking ‘Veil of Ignorance’ or anonymity assumptions. In this approach, initial positions are treated as arbitrary and redistribution is assumed to be costless. The political feasibility of adopting a proposed reform is treated as a subsidiary empirical detail that should not intrude on an ‘ethically correct’ or ‘moral’ analysis. In this literature, it is not uncommon to have the work of contemporary philosophers, and in particular Rawls, cited as authoritative; see, for example, Roemer (1996) or Sen (2000). The philosophers cited rarely consider the incentive effects of their ‘moral’ positions and ignore the political feasibility of their criteria in a modern democratic welfare state where people vote on positions in partial knowledge of the consequences of proposed policies on their personal outcomes.

The anonymity postulates adopted in this literature do not describe actual social decision making in which individuals and groups evaluate policies by asking whether they (or groups they are concerned about) are better off compared to a benchmark position. Agents know, or forecast, their positions in the distributions of outcomes under alternative policies and base their evaluations of the policies on them. Policy makers implementing policies are acutely sensitive to these perceptions.

This paper improves on the criteria widely advocated in modern social welfare theory by incorporating the evaluation of position-dependent outcomes into it, linking the outcomes under one policy regime to those in another. Such position-dependent outcomes are of interest to the individuals affected by the policies, to their representatives, and to other participants in the democratic process. This paper also considers whether these criteria are empirically operational and what is required to make them so – a topic that is neglected in the current literature in welfare economics. Ironically, accounting for the heterogeneity in outcomes that is the central focus of this literature makes the task of constructing empirically grounded counterfactuals more difficult than if the heterogeneity is ignored.

2 Ronald Reagan’s devastating rhetorical question in the 1980 US presidential campaign ‘Are you better off today than you were four years ago?’ summarises the importance of the initial position in social choice mechanisms based on voting.

© Royal Economic Society 2001
To make this discussion specific, consider the evaluation of human capital policies for schooling and job training. Human capital is the largest form of investment in a modern economy (Jorgenson and Fraumeni, 1989). Human capital involves choices at the extensive margin (schooling) and at the intensive margin (hours of job training). Differences in individual abilities are documented to affect the outcomes of human capital decisions in important ways.

The recent literature in microeconometrics establishes the empirical importance of heterogeneity in the outcomes of human capital policies even after conditioning on detailed individual and group characteristics; see Heckman (2001) for a partial survey. Using data from a social experiment evaluating a prototypical job training programme, further evidence is presented on this question and evaluations are compared under the different criteria. Theoretically important distinctions among alternative evaluative criteria turn out to be empirically important as well and produce different assessments of the same policy. This paper also summarises the recent empirical literature that documents that responses to ‘treatment’ or policies are heterogeneous and that persons make choices based on these heterogeneous responses.

I present an approach to policy evaluation that unites the macro general-equilibrium approach with the approach taken in conventional welfare economics. Using an empirically based general-equilibrium model that combines micro- and macrodata, I examine the distributional consequences of various tax policies. I present evidence on the misleading nature of the evidence produced from the partial equilibrium microeconomic treatment effect literature, and the incomplete character of the representative agent calculations that ignore distributional considerations entirely.

The plan of this paper is as follows. Section 1 presents alternative criteria that have been proposed to evaluate social policies and considers their limitations. A position-dependent criterion is proposed to evaluate policies that extends methods currently used in applied welfare economics and links them up to modern work on positive political economy. Section 2 then considers the informational requirements of the various criteria. Not surprisingly, the more interesting evaluation criteria are also more demanding in their requirements. Section 3 considers the consequences of heterogeneity in responses to policies by agents for the likely empirical success of various approaches to policy evaluation.

Section 4 considers the evidence on heterogeneity in programme impacts across persons, using data from a prototypical job training programme. A variety of criteria are used to evaluate the same programme, including revealed preference and self-assessment data and second-order stochastic-dominance comparisons as suggested by modern welfare economics. There is a surprisingly wide discrepancy in the conclusions derived from these alternative evaluation criteria. I summarise the findings from the recent literature in microeconometrics that documents that not only are responses to treatment heterogeneous, but that people act on these responses.

Section 5 then considers extending the voting criterion to a general
equilibrium setting with heterogeneous agents using an empirically based
dynamic overlapping-generations general-equilibrium model fit on both mi-
cro-and macrodata. The source model extends the analysis of Auerbach and
Kotlikoff (1987) on intergenerational accounting to include human capital
formation and heterogeneity in human ability. Estimates of alternative policy
criteria generated from the general-equilibrium framework are contrasted and
they are compared with those obtained from widely used partial equilibrium
approaches.

1. Alternative Criteria for Evaluating Social Programmes

This section considers alternative criteria that have been set forth in the
literature to assess the desirability of alternative policies. Define the outcome
for person $i$ in the presence of policy $j$ to be $Y_{ji}$ and let the personal
preferences of person $i$ for outcome vector $Y$ be denoted $U_i(Y)$. A policy
effects a redistribution from taxpayers to beneficiaries, and $Y_{ji}$ represents the
flow of resources to $i$ under policy $j$. Persons can be both beneficiaries and tax
payers. All policies considered in this paper are assumed to be feasible.

In the simplest case, $Y_{ji}$ is net income after tax and transfers, but it may also
be a vector of incomes and benefits, including provisions of in-kind services.
Many criteria have been proposed to evaluate policies. Let ‘0’ denote the no-
policy state and initially abstract from uncertainty. The standard model of
welfare economics postulates a social welfare function $W$ that is defined over
the utilities of the $N$ members of society:

$$W(j) = W[U_1(Y_{j1}), \ldots, U_N(Y_{jN})]$$ (1)

In the standard macroeconomic policy evaluation problem, (1) is collapsed
further to consider the welfare of a single person, the representative agent.
Policy choice based on a social welfare function picks policy $j$ with the highest
value for $W(j)$. A leading special case is the Benthamite social welfare func-
tion:

$$B(j) = \sum_{i=1}^{N} U_i(Y_{ji})$$ (2)

Criteria (1) and (2) implicitly assume that social preferences are defined in
terms of the private preferences of citizens as expressed in terms of their own
consumption. (This principle is called welfarism. See Sen (1979).) They could
be extended to allow for interdependence across persons so that the utility of
person $i$ under policy $j$ is $U_i(Y_{j1}, \ldots, Y_{jN})$ for all $i$.

Conventional cost–benefit analysis assumes that $Y_{ji}$ is scalar income and
orders different policies by their contribution to aggregate income:

$$CB(j) = \sum_{i=1}^{N} Y_{ji}$$ (3)

Analysts who adopt criterion (3) implicitly assume either that outputs can be
costlessly redistributed among persons via a social welfare function, or else
accept GNP as their measure of value for a policy. These measures are sometimes supplemented to generate various consumer surplus measures. While these criteria are traditional, they are not universally accepted and do not answer all of the interesting questions of political economy or ‘social justice’ that arise in the political arena of the welfare state. In a democratic society based on majority voting, politicians and advocacy groups are interested in knowing the proportion of people who benefit from policy $j$ as compared to policy $k$:

$$PB(j|j, k) = \frac{1}{N} \sum_{i=1}^{N} I[U_i(Y_{ji}) \geq U_i(Y_{ki})]$$

where 'I' is the indicator function: $I(A) = 1$ if $A$ is true; $I(A) = 0$ otherwise. In the median voter model, a necessary condition for $j$ to be preferred to $k$ is that $PB(j|j, k) \geq \frac{1}{2}$.

Persons preoccupied with ‘social justice’ are often concerned about the plight of the poor as measured in some base state $k$. For them, the gain from policy $j$ is measured in terms of the income or utility gains of the poor. In this case, interest centres on the gains to specific types of persons, eg, the gains to persons with outcomes in the base state $k$ less than $y$: $\Delta_{ji} = Y_{ji} - Y_{ki}, Y_{ki} \leq y$, or their distribution

$$F(\Delta_{jk}|Y_k = y_k, y_k \leq y)$$

or the utility equivalents of these variables; see, for example Sen (2000). Within a targeted subpopulation, there is sometimes interest in knowing the proportion of people who are at specified values of the base state $k$:

$$Pr(\Delta_{jk} > 0|Y_k \leq y)$$

In addition, measures (2) and (3) are often defined only for a target population and not for the full taxpayer population.\(^3\)

Uncertainty introduces important additional considerations. Participants in society typically do not know the consequences of each policy for each person, or for themselves, and do not know possible states not yet experienced. A fundamental limitation in applying the criteria just exposited is that, \textit{ex ante}, these consequences are not known and, \textit{ex post}, one may not observe all potential outcomes for all persons. If some potential states are not experienced, the best that agents can do is to guess about them. Even if, \textit{ex post}, agents know their outcome in a benchmark state, they may not know it \textit{ex ante},

\(^3\) The existence of merit goods like education or health implies that specific components of the vector $Y_j$ are of interest to certain groups. Many policies are paternalistic in nature and implicitly assume that some people make the wrong choices. ‘Social’ values are placed on specific outcomes, often stated in terms of thresholds. Thus one group may care about another group in terms of whether it satisfies an absolute threshold requirement:

$$Y_{ji} \geq y \quad \text{for } i \in S$$

where $S$ is a target set of persons toward which the policy is directed, or in terms of a relative requirement compared to a base state $k$: $Y_{ji} \geq Y_{ki}$ for $i \in S$.  

\(\copyright\) Royal Economic Society 2001
and they may always be uncertain about what they would have experienced in an alternative state.

In the received literature on welfare economics and social choice, one form of decision-making under uncertainty plays a central role. The ‘Veil of Ignorance’ of Vickery (1945; 1960) and Harsanyi (1955; 1975) postulates that decision makers are uncertain about their positions in the distribution of outcomes under each policy, or else they should act as if they are uncertain, and they should use expected utility criteria (Vickery–Harsanyi) or a maximin strategy (Rawls, 1971) to evaluate welfare under alternative policies. This form of ignorance is sometimes justified as capturing how an ‘objectively detached’ observer should evaluate alternative policies even if actual participants in the political process use other criteria (Roemer, 1996). An approach based on the Veil of Ignorance is widely used in applied work in evaluating different income distributions (Sen, 1973). It is an empirically more tractable approach than a criterion based on voting because it only requires information about the marginal distributions of outcomes produced under different policies, and not the joint distributions, although constructing marginal distributions of outcomes for counterfactual states not previously experienced is still a difficult task. The empirical literature on evaluating income inequality uses this criterion to compare the consequences of growing wage inequality in the past two decades; see, for example Karoly (1992). Individual outcomes under alternative policies are either assumed to be independent or else any dependence is assumed to be irrelevant for assessing alternative policies. This type of analysis is intrinsically static, whereas actual policy comparisons are made in real time: a current base state is compared to a future potential state.

An empirically more accurate description of preferences in a democratic welfare state recognises that persons or groups of persons act in their own self-interest, or in the interests of certain other groups (eg the poor, the less able). They have at least partial knowledge about how they (or the groups they are interested in) will fare under different policies, and they act on those perceptions, but only imperfectly anticipate their outcomes under different policy regimes. Assuming that positions in outcome distributions are independent across alternative policies, and that agents cannot anticipate where they will end up in each outcome distribution under each policy, justifies the Harsanyi–Vickery criterion on the basis of the self interest of participants. However, even if outcomes in alternative policy regimes are unknown (and hence represent a random draw from the outcome distribution that is independent of the place of the individual or the group in the current distribution), the outcomes under the current policy are known. In fact, the outcomes in different regimes are likely to be dependent so that persons who benefit under one policy may also benefit under another. Thus, for a variety of actual social choice mechanisms, both the initial and final positions of each agent are relevant for evaluation of social policy.4 Politicians, policy makers and participants in the

4 This theme is developed in Heckman et al. (1997), Heckman and Smith (1998), Coate (1998) and Besley and Coate (1999).
welfare state are more likely to be interested in how policies affect the fortunes of specific groups measured from a benchmark state than in some abstract measure of ‘social justice’. Even if one thought it to be ‘morally appropriate’ to ignore such considerations, they remain important in ascertaining whether policies arrived at on the basis of moral principles are politically feasible.

Agents may not possess perfect foresight so that the simple voting criterion may not accurately predict choices and requires modification. Let $I_i$ denote the information set available to agent $i$. He (she) evaluates policy $j$ against $k$ using that information. Let $F(y_j, y_k|\mathcal{J}_i)$ be the distribution of outcomes $(Y_j, Y_k)$ as perceived by agent $i$. Under an expected utility criterion, person $i$ prefers policy $j$ over $k$ if

$$E[U_i(Y_j)|\mathcal{J}_i] > E[U_i(Y_k)|\mathcal{J}_i]$$

Letting $\theta_i$ parameterise heterogeneity in preferences, so $U_i(Y_j) = U(Y_j; \theta_i)$, the proportion of people who prefer $j$ is

$$PB(j|j, k) = \sum_{i=1}^{N} I\{E[U(Y_j; \theta_i)|\mathcal{J}] > E[U(Y_k; \theta_i)|\mathcal{J}]\}$$

(7)

where the expectations are computed against the appropriate marginal distributions. The voting criterion previously discussed is the special case where $\mathcal{J}_i = (Y_{ji}, Y_{ki})$, so there is no uncertainty about $Y_j$ and $Y_k$, and

$$PB(j|j, k) = \sum_{i=1}^{N} I[U(y_j; \theta_i) > U(y_k; \theta_i)]$$

(8)

Expression (8) is a version of (4) when outcomes are perfectly predictable and when preference heterogeneity can be indexed by vector $\theta$.

Adding uncertainty to the analysis makes it fruitful to distinguish between ex ante and ex post evaluations. Ex post, part of the uncertainty about policy outcomes is resolved although individuals do not, in general, have full information about what their potential outcomes would have been in policy regimes they have not experienced and may have only incomplete information about the policy they have experienced (e.g. the policy may have long-run consequences extending after the point of evaluation). It is useful to index the information set $\mathcal{J}_i$ by $t$, $(\mathcal{J}_i,t)$, to recognise that information about the outcomes of policies may accrue over time. Ex ante and ex post assessments of a voluntary programme need not agree. Ex post assessments of a programme through surveys administered to persons who have completed it (Katz et al., 1975) may disagree with ex ante assessments of the programme. Both may reflect honest valuations of the programme but they are reported when agents

5 I abstract from the problem that politicians are more likely to be interested in voter perceptions of benefits in different policy states than in actual (post-electoral) realisations.

6 I do not claim that persons would necessarily vote ‘honestly’, although in a binary choice setting they do and there is no scope for strategic manipulation of votes. See Moulin (1983). $PB$ is simply a measure of relative satisfaction and need not describe a voting outcome where other factors come into play.

© Royal Economic Society 2001
have different information about it or have their preferences altered by participating in the programme. Before participating in a programme, persons may be uncertain of the consequences of participation in it. A person who has completed programme \( j \) may know \( Y_j \) but can only guess at the alternative outcome \( Y_k \) which they have not experienced. In this case, \textit{ex post} ‘satisfaction’ for agent \( i \) is synonymous with the inequality

\[
U_i(Y_{ji}) > \mathbb{E}[U_i(Y_{ki})|\mathcal{F}_t] \tag{9}
\]

where \( t \) is the post-programme period in which the evaluation is made. In addition, survey questionnaires about ‘client’ satisfaction with a programme may capture subjective elements of programme experience not captured by ‘objective’ measures of outcomes that usually exclude psychic costs and benefits.

### 2. The Data Needed to Evaluate the Welfare State

To implement criteria (1) and (2), it is necessary to know the distribution of outcomes across the entire population within each policy state and to know the utility functions of individuals. In the case where \( Y \) refers to scalar income, criterion (3) only requires knowledge of GNP (the sum of the programme \( j \) net output). If interest centres solely on the distributions of outcomes of direct programme participants, the measures can be defined solely for populations with \( D_j = 1 \). Criteria (4), (5), (6) and (8) require knowledge of outcomes and preferences \textit{across} policy states. Criterion (7) requires knowledge of the joint distribution of information and preferences across persons. Table 1 summarises the criteria and the data needed to implement them for the entire population.\(^7\) The cost–benefit criterion is the least demanding; the voting criterion is the most demanding in that it requires information about the \textit{joint} distributions of outcomes across alternative policy states.

Three distinct types of information are required to implement all of these criteria:

(a) private preferences, including preferences toward the consumption and well being of others;

(b) social preferences, as exemplified by social welfare function \( W \); and

(c) distributions of outcomes in alternative states, and for some criteria, such as the voting criterion, \textit{joint} distributions of outcomes \textit{across} policy states.

The reasons for the popularity of cost–benefit analysis are evident from the table, because it requires so little information relative to the other evaluative criteria. An important practical problem rarely raised in the literature on ‘social justice’ is that many proposed evaluative criteria are not empirically operational with current levels of knowledge.

There is a vast literature on the estimation of individual preferences defined

\(^7\) A parallel set of requirements can be formulated for a population of participants only \((D = 1)\). See Heckman and Smith (1998).

© Royal Economic Society 2001
## Table 1

**Data Requirements to Implement Criterion on the General Population**

Programme \( j \) compared to programme \( k \)

<table>
<thead>
<tr>
<th>Cost–benefit</th>
<th>Benthamite criterion</th>
<th>General social welfare function with interdependent preferences*</th>
<th>Selfish voting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Criterion</td>
<td>( U_i \geq U_k )</td>
<td>( W(j) &gt; W(k) )</td>
<td>( \sum_{i=1}^{N} 1[U(y_i, \theta_i) \geq U(y_k, \theta_i)] )</td>
</tr>
<tr>
<td>( \text{E}(Y_j) - \text{E}(Y_k) \geq 0 )</td>
<td>( U_l = \sum_{i=1}^{N} \text{E}[U(Y_i, \theta_i)] = \sum_{i=1}^{N} \int U(y_i, \theta_i) dF(y_i) )</td>
<td>( W(l) = W[U_l(Y_{1,l}, \ldots, Y_{N_1,l}, \ldots, \ldots, Y_{N_l,l})] )</td>
<td>( \int dF(y_j, y_k) \geq 0 )</td>
</tr>
<tr>
<td>Require</td>
<td>( \text{U}(Y_l, \theta_l) ) and distribution of ( Y_l, l = j, k ).</td>
<td>Each ( U_i(Y_{1,i}, \ldots, Y_{N_l,i}) ) for all ( i ), outcomes for each person*</td>
<td>Need ( U(Y, \theta), F(y_j, y_k)**</td>
</tr>
<tr>
<td></td>
<td>( \text{E}(Y_j), \text{E}(Y_k) )</td>
<td>Enough to get joint distribution of ( (Y_i, \theta) ).</td>
<td>Enough to get joint distribution of ( (Y_j, Y_k, \theta) ).</td>
</tr>
</tbody>
</table>

* In special cases, summary statistics of the distribution of \( Y \) may suffice
** For altruistic voting, \( U \) depends on \( Y_{1,j}, \ldots, Y_{N_l} \) or various sub-aggregators.
over goods and leisure although the literature on the determination of altruistic preferences is much smaller. Within the framework of the microeconomic treatment effect literature, the decisions of the agents to self select into a programme reveal their preferences for it. Much of the standard literature on estimating consumer preferences abstracts from heterogeneity. However, a growing body of evidence summarised in Browning et al. (1999) demonstrates that heterogeneity in marginal rates of substitution across goods at a point in time, and for the same good over time, is substantial. This heterogeneity is large across demographic and income groups and is large even within narrowly defined demographic categories; see, for example Heckman (1974a).

There are surprisingly few estimates of social welfare function (1) – Maital (1973); Saez (1998) and Gabaix (1998) are exceptions – despite the widespread use of the social welfare function in theoretical public economics. The paucity of estimates of it suggests that the social welfare function is an empirically empty concept. If a stable social welfare function characterised social decision making, it should be possible to estimate a stable demand for public goods using conventional methods in consumer demand theory. Yet the evidence for a stable demand function for public goods is scarce.

The rest of this paper focuses on the problem of determining outcomes under different policies. A full implementation of all of these criteria requires knowledge of \( W \) and \( U \). Their determination raises additional important questions that are not discussed here.

Income shocks, wages and the like vary widely across consumers. The evidence speaks strongly against the representative agent model even for outcomes. The focus of the empirical analysis of this paper is on estimating the distributions of outcomes across policy states as a first step toward empirically implementing the full criteria. This more modest objective can fit into the framework of section 1 by assuming that utilities are linear in their arguments and identical across persons. Even this more modest goal is a major challenge, as we shall see.

2.1. The Policy Evaluation Problem

The policy evaluation problem in its most general form can be written as estimating a vector of outcomes for each person in each policy state. Consider policies \( j \) and \( k \). The potential outcomes are

\[
(Y_{ji}, Y_{ki}) \quad i = 1, \ldots, I
\]

(10)

Macroeconomic approaches focus exclusively on mean outcomes or some other low dimensional representation of the aggregate (eg geometric means). There are two important cases of this macro problem: (a) the case where \( j \) and \( k \) have been experienced in the past and (b) where one of \( j \) or \( k \), or possibly both, have never been observed.\(^8\) The first case requires that we somehow ‘adjust’ the data on \( j \) and \( k \) to account for changes in the conditioning

\(^8\) These distinctions go back to Marschak (1953). See Heckman (2001) for a discussion.
variables between the observation period and the period for which the policy is proposed to be implemented. Such adjustments are sometimes controversial. If the environment is stationary, no adjustment is required. With panel data on persons, one might build up the joint distribution of policy outcomes by observing the same people under different regimes.

The second case is the classical macroeconomic general-equilibrium policy-evaluation problem considered by Knight (1921), Tinbergen (1956), Marschak (1953), Theil (1961), Lucas and Sargent (1981) and Lucas (1987) forecasts and evaluates the impacts of policies that have never been implemented. To do this requires knowledge of policy-invariant structural parameters and a basis for making proposed new policies comparable to old ones.

An entire literature on structural estimation in econometrics has emerged to solve this problem. By focusing on the ‘representative consumer’, this literature simplifies a hard problem by ignoring the issue of individual heterogeneity in outcomes within each regime. If outcomes were indeed identical across persons, or if the representative consumer were a ‘reasonably good’ representation, from knowledge of aggregate means, one could implement all of the criteria in Table 1 using only data on mean outcomes under each regime provided that the appropriate preferences were known. This is a consequence of the implicit assumption of the representative consumer model that the joint distribution of (10) is degenerate. Notice that this task may still be quite difficult if some regimes are not observed.

One commonly used approach in the microeconometric literature explicitly allows for heterogeneity in outcomes. It considers evaluation of a programme in which participation is voluntary although it may not have been intended to be so. Accordingly, it is not well suited to evaluating programmes with universal participation such as a social security programme.

Persons are offered a service through a programme and may select into the programme to receive it. A distinction is made between direct participation in the programme and indirect participation. The latter occurs when people do not receive the services of the programme but pay taxes or suffer the market consequences of changed supplies of goods and factors as a consequence of the programme. Eligibility for the programme may be restricted to subsets of persons in the larger society. Participation in the programme is thus equated with direct receipt of the service, and payments of taxes and general equilibrium effects of the programme are typically ignored.

In this formulation of the evaluation problem, the no-treatment outcome distribution for a given programme is used to approximate the distribution of outcomes in the no-programme state. That is, the outcomes of the ‘untreated’ within the framework of an existing programme are used to approximate

---

9 As summarised in Browning et al. (1999), there is an emerging literature in macroeconomics that recognises the evidence of microheterogeneity and its consequences for model construction and policy evaluation.

10 Many ‘mandatory’ programmes allow persons to attrite from them or fail to comply with programme requirements.
outcome distributions when there is no programme. This approximation rests on two separate arguments:

(a) that general-equilibrium effects inclusive of taxes and spillover effects on factor and output markets can be ignored; and

(b) that the problem of selection bias that arises from using self-selected samples of participants and nonparticipants to estimate population distributions can be ignored or surmounted.11

The treatment effect approach also converts the evaluation problem into a comparison between an existing programme \(j\) and a benchmark no-programme state.12

More precisely, let \(j\) be the policy regime to be evaluated. Eligible person \(i\) in regime \(j\) has two potential outcomes: \((Y_{0j}, Y_{1j})\), where the superscripts denote non-direct participation (‘0’) and direct participation (‘1’). Ineligible persons have only one option: \(Y_{0j}\). These outcomes are defined at the equilibrium level of participation under programme \(j\). All feedback effects are incorporated in the definitions of the potential outcomes.

Let subscript ‘0’ denote a policy regime without the programme. Let \(D_{ji} = 1\) if person \(i\) participates in programme \(j\). A crucial identifying assumption that is implicitly invoked in the microeconomic evaluation literature is

\[ Y_{0j} = Y_{0i} \]  
(A-1)

ie that the no programme outcome for \(i\) is the same as the no treatment outcome.

Letting \(F(a|b)\) denote the conditional distribution of \(a\) given \(b\), the assumption implies that \(F(y_{0j}|D_j = 0, X) = F(y_0|D_j = 0, X)\) for \(y_{0j} = y_0\) given conditioning variables \(X\). The outcome of nonparticipants in policy regime \(j\) is the same in the no policy state ‘0’ or in the state where policy \(j\) is operative. This assumption is consistent with a programme that has ‘negligible’ general equilibrium effects and where the same structure of tax revenue collection is used in regimes \(j\) and ‘0’.

From data on individual programme participation decisions, it is possible to infer the implicit relative private valuations of the various programmes made by persons eligible for them. These evaluations constitute all of the data needed for a libertarian programme evaluation, but more than these are required to evaluate programmes from the vantage point of others, or to implement the criteria in Table 1. For certain decision rules, it is possible to use the data from self-selected samples to bound or estimate the joint distributions required to implement criteria (4) or (7), as is demonstrated below.

---

11 As noted below, evidence from self-selection decisions can be used to evaluate private preferences for the programme so that, in principle, the ‘problem’ of self selection can be used as a source of information about private valuations. See, for example, Heckman (1974a,b), Heckman and Honoré (1990) and Heckman and Smith (1998) where this is done.

12 In the case of multiple observed treatments, comparisons can be made among observed outcomes as well as against a benchmark no-programme state. In fact, the ‘no programme’ benchmark may well be a benchmark with an alternative in place but then it is implicitly assumed that (A-1) (stated below) applies to ‘k’ where ‘k’ replaces ‘0’.

© Royal Economic Society 2001
Consider now how access to microdata and social experiments enables one to answer the evaluation questions posed in section 1.

3. Using Micro Data to Operationalise the Criteria

Microdata would appear to be a fruitful source of information on heterogeneity in individual responses to treatment. This section considers the information on outcomes from observational data on individuals. Even abstracting from the problem that the analysis of these data typically ignores general-equilibrium effects, and from the problems of determining both social and private preferences, the information produced from microdata is surprisingly limited unless a strong form of homogeneity is invoked. This homogeneity assumption is implicitly invoked in most micro studies so there is a closer kinship between micro and representative agent approaches than might be first thought. The micro studies condition more finely. Both macro and micro studies ignore well-documented sources of heterogeneity among agents in responses to programmes.

Consider the analysis of programme \( j \) and suppose that assumption (A-1) is invoked. Within the framework of the ‘treatment effect’ literature, observe one of \((Y^0_{ij}, Y^1_{ij})\) for person \( i \). To simplify the notation, the \( j \) subscript is dropped in this section. At a point in time, one cannot observe a person simultaneously in the treated and untreated state. In general, one cannot form the gain of moving from ‘0’ to ‘1’ and \( \Delta_i = Y^1_i - Y^0_i \) for anyone. The evaluation problem is reformulated to the population level. The goal becomes to estimate some features of the distribution of \( \Delta \). To clarify this approach, let \( D_i = 1 \) if person \( i \) is a direct participant, and \( D_i = 0 \) if person \( i \) is not a direct participant. Observe that \( Y_i = D_i Y^1_i + (1 - D_i) Y^0_i \) for each person.

The potential outcomes for person \( i \) can be written as

\[
Y^0_i = \mu_0 + \varepsilon_{0i} \\
Y^1_i = \mu_1 + \varepsilon_{1i}
\]

where \( \mathbb{E}(\varepsilon_0) = \mathbb{E}(\varepsilon_1) = 0 \). The means can be written in terms of observed characteristics \( X[\mu_0(X); \mu_1(X)] \) but for simplicity of notation this dependence is kept implicit. Thus one can write

\[
Y_i = \mu_0 + (\mu_1 - \mu_0 + \varepsilon_{1i} - \varepsilon_{0i}) D_i + \varepsilon_{0i}
\]

Much of the evaluation literature assumes that the parameters of interest are means; see, however, Heckman et al. (1997) or Heckman and Smith (1998). Two means receive the most attention. The first is \( \mathbb{E}(Y^1 - Y^0) \) the average treatment effect (ATE) that records the average gain of moving a randomly selected person from ‘0’ to ‘1’. A second mean is \( \mathbb{E}(Y^1 - Y^0 | D = 1) \) the effect of treatment on the treated (TT). The two means are the same under one of the following conditions:

\[ (C-1): \quad \varepsilon_{1i} = \varepsilon_{0i} \text{ so } \Delta_i = \Delta \quad \text{(No response heterogeneity given } X) \]
(C-2): \( E(\varepsilon_{1i} - \varepsilon_{0i} \mid D_i = 1) = 0 \) (Agents do not enter the programme based on gains from it)

Under (C-1), outcome responses are identical among persons with given observed characteristics \( X \). Under (C-2), outcomes may differ among persons with identical \( X \) characteristics but, \textit{ex ante}, there is no perceived heterogeneity. (Persons place themselves at the mean of the response distribution for ‘0’ and ‘1’ in making their participation decisions.)

To understand these distinctions, it is useful to consider three regression models. Recall that, for simplicity, the conditioning on \( X \) is kept implicit.\(^{13}\)

Write the traditional textbook model as

\[
Y_i = \alpha_0 + \alpha_1 D_i + \varepsilon_i \quad E(\varepsilon_i) = 0
\]

(A)

In this framework, \( \alpha_1 \) is a common coefficient for each \( i \). It embodies assumption (C-1) where \( \varepsilon_i = \varepsilon_{1i} = \varepsilon_{0i} \) and \( \alpha_1 = \Delta = \mu_1 - \mu_0 \). There is no idiosyncratic response to treatment among persons with the same observed characteristics \( X \). This is the textbook model of econometric policy evaluation and the textbook model of econometrics. Selection or simultaneity bias is said to arise if \( E(\varepsilon_i \mid D_i = 1) \neq 0 \).

In contrast, consider a second model with outcome specific errors and

\[
Y_i = \alpha_0 + \alpha_{1i} D_i + \varepsilon_{0i}
\]

where \( \alpha_{1i} = Y_{1i} - Y_{0i}, E(\alpha_{1i}) = \mu_1 - \mu_0 \) and \( \varepsilon_{1i} - \varepsilon_{0i} \) satisfies

\[
E(\varepsilon_{1i} - \varepsilon_{0i} \mid D_i = 1) = 0
\]

(B)

In this framework, responses are different across persons (\( \alpha_1 \) has an \( i \) subscript) but conditional on \( X \), persons do not participate in the programme based on these differential responses.\(^{14}\) Again selection bias is said to arise if \( E(\varepsilon_{0i} \mid D_i = 1) \neq 0 \).

If persons participate in the programme based on these differential responses, we obtain

\[
Y_i = \alpha_0 + \alpha_{1i} D_i + \varepsilon_{0i} \quad E(\varepsilon_{1i} - \varepsilon_{0i} \mid D_i = 1) \neq 0
\]

(C)

Again, selection bias for \( E(Y_{1i} - Y_{0i} \mid D = 1) \) is said to arise if \( E(\varepsilon_{0i} \mid D_i = 1) \neq 0 \).

Under both models (A) and (B), the parameters \( E(Y_{1i} - Y_{0i}) \) and \( E(Y_{1} - Y_{0} \mid D = 1) \), are the same. Under Model (C), they are not. These distinctions, first introduced in Heckman and Robb (1985, 1986), have important consequences for what can be learned from micro evaluations.

Model (A) is the dominant paradigm in the applied literature. If it is true, and if assumption (A-1) is also true, one can go from a regression estimate of equation (A) to answer all of the policy questions posed in section 1 comparing the policy being evaluated with a benchmark no policy state. The distribu-

\(^{13}\) These models were first introduced in Heckman and Robb (1985, 1986).

\(^{14}\) Another way to say this is that \( \Pr(D_i = 1 \mid Z_i, \varepsilon_{1i} - \varepsilon_{0i}) = \Pr(D_i = 1 \mid Z_i) \). This is a cross-section 'noncausality' condition.
tion of gains, $\Delta$, across and within policy regimes is degenerate. Everyone either benefits or loses from the policy being evaluated. In this case, the inferences obtained from the representative agent paradigm, the inferences obtained from cost–benefit analysis, and the inferences obtained from the treatment effect literature are the same.

Model (B) captures heterogeneity in response to treatment but assumes that persons do not act on it. Individuals act under a partial ‘Veil of Ignorance’. They may forecast components of policy states conditional on $X$ but not on the basis of increments in the idiosyncratic components, $\varepsilon_{1i} - \varepsilon_{0i}$. Now the representative agent paradigm should be adjusted to account for ex post variation in individual responses to the programme. The cost–benefit approach is robust to this form of heterogeneity because it considers only mean outcomes. The treatment effect approach requires estimation of the variances of outcomes. See, for example, Heckman et al. (1997). If outcomes are heterogeneous in the sense of model (B), conventional instrumental variable and matching methods can be used to secure estimates of mean parameters. As long as means are the focus of attention, estimation of model (B) raises only well-known and easily solved heteroscedasticity problems. However, apart from the studies by Heckman and Smith (1993, 1998) and Heckman et al. (1997), there are few studies that estimate the distributions of programme impacts.

Model (C) captures a more fundamental form of heterogeneity in programme impacts. Agents know much more than the observing economist and they act on this information in deciding whether to participate in a programme. $E(Y_1 - Y_0) \neq E(Y_1 - Y_0|D = 1)$. Estimating the full parameters of the outcome distributions and their correlations over states is a frontier topic in econometrics with recent developments surveyed in Heckman and Vytlacil (2001) and Heckman (2001). In this case, standard instrumental variable methods break down; see Heckman (1997) or Heckman and Vytlacil (1998). Heckman et al. (1997) and Heckman and Smith (1993, 1998) present estimates of outcome distributions under Model (C). Heckman et al. (1998) present evidence that Model (C) describes the data for the prototypical training programme discussed in section 5 below. This supports earlier findings by Björklund and Moffitt (1987) who reach a similar conclusion for Swedish training programmes. While most of the thinking about programme evaluation is in terms of Model (A) or more recently, in terms of Model (B), considerable evidence supports Model (C) for many programmes. Heckman and Vytlacil (2001) and Carneiro et al. (2001) present evidence that Model (C) also describes the returns to education. Heckman (2001) shows that Model (C) describes the outcomes for migration and job training.

As noted by Heckman (1992), the enthusiasm for social experiments in the policy evaluation community is premised on the implicit acceptance of Model (A). Knowing the mean impact $\alpha_1$ is enough to answer all of the policy evaluation questions posed in section 1 provided the preference parameters are determined. The joint distribution of (10) is degenerate when $k$ is the benchmark no-programme state. Even if randomisation alters the composition...
of programme participants (ie there is ‘randomisation bias’), for any observed \( X \) in the experiment, one can obtain \( \alpha_1 \).

If Model (C) characterises the data, all that can be recovered from social experiments administered to people who apply and are accepted into the programme (the common point in the enrolment process where randomisation is administered) are \( F(y^1|D = 1) \) and \( F(y^0|D = 1) \). Without further assumptions, one cannot recover the joint distribution \( F(y^1, y^0|D = 1) \) either for persons who seek to participate in the programme or for the general population \( (F(y^1, y^0)) \). Section 3.2 discusses what can be learned in this case. First, however, section 3.1 considers what can be learned from participation decisions under Model (C).

3.1. Information From Revealed Preference

If agents act on the idiosyncratic gain from the programme, so Model (C) is the appropriate one, and the Veil of Ignorance does not describe private self-interested evaluations of alternative social states, it is possible to use information on choices to infer the implicit valuations they place on the gains from the programme being evaluated. If they do not participate on the basis of the gain, then clearly there is no information on the gain from participation decisions. Participation includes voluntary entry into a programme or attrition from it.\(^{15}\)

The prototypical framework for models of selection is the Roy (1951) model. In that framework

\[
D = I(Y^1 \geq Y^0)
\]

so participation depends only on the net gain from the programme: \( \Delta = Y^1 - Y^0 \). Using (11) and (12)

\[
\Pr(D = 1|X) = \Pr(Y^1 - Y^0 \geq 0|X) = \Pr(\varepsilon_1 - \varepsilon_0 \geq -[\mu_1(X) - \mu_0(X)])
\]

Under conditions specified in Heckman and Honoré (1990), the joint distribution \( F(y^0, y^1, D|X) \) can be identified nonparametrically.\(^{16}\) Thus, putting general equilibrium problems aside and assuming preference parameters are known, we can form all of the evaluation parameters presented in Table 1 if the Roy model describes the data.

The crucial feature of the Roy model is that the decision to participate in the programme is made solely in terms of potential outcomes. No new

\(^{15}\) Heckman (1974a,b) demonstrates how access to censored samples on hours of work, wages for workers, and employment choices identifies the joint distribution of the value of nonmarket time and potential market wages under a normality assumption. Heckman and Honoré (1990) consider nonparametric versions of this model without labour supply.

\(^{16}\) Heckman and Honoré (1990) demonstrate that if \( X \) is independent of \((\varepsilon_1, \varepsilon_0)\), \( \text{Var}(\varepsilon_1) < \infty \) and \((\varepsilon_1, \varepsilon_0)\) are normal, the full model \( F(y^0, y^1, D|X) \) is identified even if we only observe \( Y^0 \) or \( Y^1 \) for any person and there are no regressors and no exclusion restrictions. If, instead of assuming normality, it is assumed that the supports of \((\varepsilon_1(X), \varepsilon_0(X))\) overlap or contain the supports of \((\varepsilon_1, \varepsilon_0)\), the full model \((\mu_1(X), \mu_0(X))\) and the joint distribution of \((\varepsilon_1, \varepsilon_0)\) are nonparametrically identified up to location normalisations. Precise conditions are given in Theorem A-1 in Appendix A of Heckman and Smith (1998).

© Royal Economic Society 2001
unobservable variables enter the model that do not appear in the outcome equations. In this case, information about who participates also informs us about the distribution of the value of the programme to participants \( F(y^1 - y^0 | Y^1 > Y^0, X) \). Thus, we acquire the distribution of implicit values of the programme for participants, which is all that is required in a libertarian evaluation of the programme. However, in the general case, evaluation of the welfare state requires information about ‘objective’ outcomes and their distributions that are needed to make the interpersonal comparisons that are an essential feature of the welfare state. In the Roy model the ‘objective’ and ‘subjective’ evaluations coincide.

Observe that under the assumptions that make it valid, estimation of a Roy model on ordinary nonexperimental data produced by the self-selection decisions of participants is more informative than analysis of experimental data on persons who attempt to enter the programme. As noted by Heckman (1992) and Moffitt (1992), social experiments are typically conducted on persons who apply and are initially accepted into a programme. Information from such experiments does not provide information about the determinants of programme participation because it only produces samples conditional on \( D = 1 \). Nonexperimental data can be used to infer the preferences of agents who select into the programme and, in the context of the Roy model, are more informative.

3.2. The Problem of Recovering Joint Distributions

In the general case where textbook Model (A) does not apply, and responses to alternative programmes are heterogeneous among individuals, a difficult evaluation problem is encountered. Unless the Roy model is invoked, one cannot identify the joint distribution of \( (Y^0, Y^1) \). At best one can extract the marginal distributions of \( Y^0 \) and \( Y^1 \), even from ideal social experiments. In this case, implementation of the voting criterion is difficult.

To see the problem, suppose that there are data from an ideal social experiment with compulsory participation so that standard self-selection problems can be ignored. Suppose that there are \( N \) treated and \( N \) untreated

---

17 We could augment decision rule (14) to be \( D = I(Y^1 - Y^0 - k(Z) \geq 0) \). Provided that we measure \( Z \) and condition on it, and provided that \( (U_1 - U_0) \perp (X, Z) \), the model remains non-parametrically identified. The crucial property of the identification result is that no new unobservable enters the model through the participation equation. However, if we add \( Z \), subjective valuations of gain \( (Y^1 - Y_0 - k(Z)) \) no longer equal ‘objective’ measures \( (Y^1 - Y^0) \).

18 If the Roy model is extended to allow for variables other than \( Y^0, Y^1 \) (and the observed conditioning variables) to determine participation, then the decision rule is changed to \( D = I(IN > 0) \) where \( IN = \eta(Y^1, Y^0, V, X) \), and it is not possible to identify the joint distribution \( F(u_0, u_1) \) even if the unobservables \( V \), \( U_0 \) and \( U_1 \) are independent of \( X \). Heckman (1990) demonstrates that, in this more general case, provided that some structure is placed on \( \eta \), we can nonparametrically identify \( F(y^0, D|X) \) and \( F(y^1, D|X) \) but not the full joint distribution \( F(y^0, y^1, D|X) \). A generalisation of his proof is given in Theorem A-2 of Appendix A of Heckman and Smith (1998).

19 Heckman and Smith (1998) extend the Roy model to allow for uncertainty in the outcomes as perceived by agents. They show that even when \( Y^0 \) and \( Y^1 \) are independent or even negatively correlated in the population, purposive decision making produces positive dependence in the population.
persons and that the outcomes are continuously distributed. Rank the individuals in each treatment category in the order of their outcome values from the highest to the lowest. Define $Y_{j}^{(i)}$ as the $i$th highest-ranked person in the $j$ distribution. Ignoring ties, one obtains two data distributions:

- **Treatment Outcome**: $F(y^1)$
- **Non-treatment Outcome**: $F(y^0)$

$$Y^1 = \begin{pmatrix} Y^1_{(1)} \\ \vdots \\ Y^1_{(N)} \end{pmatrix} \quad Y^0 = \begin{pmatrix} Y^0_{(1)} \\ \vdots \\ Y^0_{(N)} \end{pmatrix}$$

The experiments determine the marginal data distributions $F(y^1)$ and $F(y^0)$ but, from the experiments, one cannot determine where person $i$ in the treatment distribution would appear in the non-treatment distribution. This poses no problem for the criteria based on the anonymity postulates but is a major limitation for determining the voting criterion, since the joint distribution is not determined from knowledge of the marginals.

Corresponding to the ranking of the treatment outcome distribution, there are $N!$ possible patterns of outcomes in the associated non-treatment outcome distribution. By considering all possible permutations, one can form a collection of possible impact distributions, i.e., alternative distributions of the gain:

$$\Delta = Y^1 - \Pi_l Y^0 \quad l = 1, \ldots, N!$$

where $\Pi_l$ is a particular $N \times N$ permutation matrix of $Y^0$ in the set of all $N!$ permutations associating the ranks in the $Y^1$ distribution with the ranks in the $Y^0$ distribution; and $\Delta$, $Y^1$, and $Y^0$ are $N \times 1$ vectors of impacts, treated and untreated outcomes. By considering all possible permutations, one can obtain all possible sortings of treatment outcomes, $Y^1$, and non-treatment outcomes, $Y^0$, using realised values from one distribution as counterfactuals for the other.

Model (A) assumes a constant treatment effect for all persons conditional on characteristics. This model admits only one permutation: $\Pi = I$ for each $X$. The best in one distribution is the best in the other distribution. In this common effect case, $Y^1$ and $Y^0$ differ by the same constant for all persons. A generalisation of that model preserves perfect dependence in the ranks between the two distributions but does not require the impact to be the same at all quantiles of the base state distribution (Heckman et al., 1997).

In place of ranks, it is easier in empirical work to analyse the percentiles of the $Y^1$ and $Y^0$ distributions, which have much better statistical properties. Equating percentiles across the two distributions, one can form the pairs across the distributions and obtain a deterministic gain function $\Delta(y^1, y^0)$. This presents the gain in going from benchmark state ‘0’ to outcome state ‘1’. For the case of absolutely continuous distributions with positive density at $y^0$, the gain function can be written as $\Delta(y^0) = F_{1}^{-1}[F_{0}(y^0)] - y^0$. One can test non-parametrically for the classical common effect model by determining if per-

---

20 These distributions can also be defined conditional on $X$.

21 See Heckman and Smith (1993), Heckman et al. (1997).
centiles are uniformly shifted at all points of the distribution. One can form
other pairings across percentiles by mapping percentiles from the \( Y^1 \) distribution into percentiles from the \( Y^0 \) distribution using the map \( T : q_1 \to q_0 \). The
data are consistent with all admissible transformations including \( q_0 = 100 - q_1 \), where the best in one distribution is mapped into the worst in the other. They cannot reject any of these models or more general models where \( \Pi_i \) is a Markov transition matrix and we consider all possible Markov matrices (Heckman and Smith, 1998).

Heckman and Smith (1998) and Heckman et al. (1997) consider empirical
evidence on the question of the constancy of the gross gain \( \Delta \) across base state
quantiles using earnings data from an experimental evaluation of a major US
job training programme (JTPA). For a population conditional on \( D = 1 \) Fig.
1 displays the estimate of earnings gains \( \Delta(y_0) \) for adult women assuming that
the best persons in the ‘1’ distribution are the best in the ‘0’ distribution
(\( \Pi = I \)). No conditioning on covariates is made, so the full sample is utilised.
Between the 25th and 85th percentiles, the assumption of a constant impact is
roughly correct. In this range, the data support Model (A). However, the data
are grossly at odds with this model at the highest and lowest percentiles.\(^{23}\)

---

Fig. 1. Treatment ± Control Differences at Percentiles of the 18-Month Earnings Distribution

Source: Heckman and Smith (1993) and Heckman et al. (1997).

---

\(^{22}\) In their data, randomisation occurs among people who are initially accepted into the programme
\( (D = 1) \).

\(^{23}\) Standard errors for the quantiles are obtained using methods described in CsoÈrgö (1993).
Heckman et al. (1997) and Heckman and Smith (1993, 1998) present a more extensive empirical analysis of data using different conditioning sets and reach essentially the same conclusion. Observe that even though the ranks are assumed to be perfectly dependent across the two distributions, there is substantial heterogeneity in the gains at different points of the base state distribution.

4. Evidence on Impact Heterogeneity and the Conflict among Self-Assessments, Revealed Preference Information and Other Criteria

This section addresses three questions using data from the JTPA study. Question (1) is: ‘What is the empirical evidence on heterogeneity in programme impacts among persons?’ The conventional approach implicitly assumes impact homogeneity conditional on observables. This assumption greatly simplifies the task of evaluating the welfare state. Using data on earnings from an experimental evaluation of a prototypical job training programme, I implement the criteria discussed in section 1 under the assumption that preferences are linear in outcomes to bound or identify the joint distribution of outcomes conditional on $D = 1$. There is considerable evidence of heterogeneity of programme impacts, so that conventional econometric methods based on common response models do not take one very far in constructing the evaluation criteria discussed in section 1.

Given the evidence on impact heterogeneity, question (2) is: ‘How sensitive are the estimates of the proportion of people who gain from the programme – the “voting criterion” – to alternative assumptions about the dependence between $Y^0$ and $Y^1$?’ The estimates are very sensitive to alternative assumptions. At the same time, for adult women, the estimated percentage that benefit from the programme exceeds 50% in every case considered but one, and is close to 100% in some cases.

Some of the estimates used to answer question (2) assume that $Y^0$ and $Y^1$ are positively dependent given $D = 1$. Under purposive selection based on outcomes in the treated and untreated states, such dependence among participants arises even if $Y^1$ and $Y^0$ are independent or negatively correlated in the population as a whole (Heckman and Smith, 1998). An alternative to imposing a particular decision rule is to infer it from self-assessments of the programme. These assessments are all that are required for a libertarian evaluation of the welfare state. I examine the implicit value placed on the programme by addressing two questions: (3a) ‘Are persons who applied to the programme and were accepted into it but then randomised out of it placed in an inferior position relative to those accepted applicants who were not randomised out?’ I measure ex ante rational regret using second-order stochastic dominance, which is an appropriate measure under the assumption that individuals are completely uncertain of both $Y^1$ and $Y^0$ before going into the programme. I also consider ex post evaluations of participants by asking: (3b) ‘How “satisfied” are participants with their experience in the programme?’ Self-assessments of programmes are widely used in evaluation research – see,
for example, Katz et al. (1975) – but the meaning to be placed on them is not clear. Do they reflect an evaluation of the experience of the programme (its process) or an evaluation of the benefits of the programme? The evidence presented here suggests that respondents report a net benefit inclusive of their costs of participating in the programme. Groups for whom the programme has a negative average impact as estimated by the ‘objective’ experimental data express as much (or more) enthusiasm for the programme as groups with positive average impacts. A third source of revealed preference evaluations uses the revealed choices of attriters from the programme. Econometric models of self-selection since Heckman (1974a,b) have used revealed choice behaviour to infer the evaluations people place on programmes either by selecting into them or dropping out of them. The third part of the third question is thus (3c): ‘What implicit valuation of the programme do attriters place on it?’ This question is not examined in this paper; Heckman and Smith (1998) present evidence on it.

4.1. Evidence on Heterogeneity

Heckman and Smith (1993, 1998) apply the nonparametric Frechet Bounds of classical probability theory to the data from the JTPA programme to establish that the variance of the gain $\Delta$ is positive for a variety of conditioning sets. Their estimate for the JTPA data is reported in the first row of Table 4. The $675 lower bound on the standard deviation is to be compared with a $400 gain and mean $7,200 base income for women. Heckman and Smith report a variety of other estimates that support the conclusion that even within narrowly defined conditioning sets, the variance in outcomes is substantial for women and for other demographic groups.24

Using the sample data from the JTPA experiment – see Orr et al. (1995) and discussion in Heckman and Smith (1998) – we may pair percentiles of the $Y^1$ and $Y^0$ distributions for any choice of Kendall rank correlation $\tau$ between $-1.0$ and 1.0. The case of $\tau = 1.0$ corresponds to the case of perfect positive dependence, in positions across potential outcome distributions where $\Pi = 1$ and $q_1 = q_0$. The case where $\tau = -1.0$ corresponds to the case of perfect negative dependence, where $q_1 = 100 - q_0$. The first and last rows of Table 2 display estimates of quantiles of the impact distribution and other features of the impact distribution for these two cases.

Heckman et al. (1997) show how to obtain random samples of permutations conditional on values of $\tau$ between $-1.0$ and 1.0. Table 2 displays two sets of estimates of these distributions from their work. The first set assumes positive but not perfect dependence between the percentiles of $Y^1$ and $Y^0$, with

24 The classical solution to bounding a joint distribution from its marginals uses the Frechet-Hoeffding Bounds. For the population conditional on $D = 1$ they are

$$\max[F(y^1|D = 1) + F(y^0|D = 1) - 1, 0] \leq F(y^1, y^0|D = 1) \leq \min[F(y^1|D = 1), F(y^0|D = 1)].$$

These do not directly apply to the distribution of the gain $\Delta = Y^1 - Y^0$ but can be used to bound the variance of the gain. See Heckman and Smith (1993; 1998) for details.
Estimates based on a random sample of 50 percentile permutations with this value of $\hat{\sigma}$ appear in the second column of Table 2. These results show that even a modest departure from perfect positive dependence substantially widens the distribution of impacts. More striking still are the results in the third column of Table 2, which correspond to the case where $\hat{\sigma} = 0.0$. This value of $\hat{\sigma}$ is implied by independence between the percentiles of $Y^1$ and $Y^0$.

Here (as in the case with $\hat{\sigma} = 0.0$) the distribution of estimated impacts is implausibly wide with large positive values in each distribution often matched with zero or small positive values in the other. However, the conclusion that a majority of adult female participants benefit from the JTPA programme is robust to the choice of $\hat{\sigma}$.

Even though many joint distributions of outcomes are consistent with the marginals produced from a social experiment, one model is not: common effect model equation (A). Heckman et al. (1997) test and reject the assumption that $\Delta(= \alpha_1)$ is a common coefficient, using a variety of conditioning sets. Heterogeneity is a central feature of the data, even within narrowly defined

\[ \tau = 0.95. \] Estimates based on a random sample of 50 percentile permutations with this value of $\tau$ appear in the second column of Table 2. These results show that even a modest departure from perfect positive dependence substantially widens the distribution of impacts. More striking still are the results in the third column of Table 2, which correspond to the case where $\tau = 0.0$. This value of $\tau$ is implied by independence between the percentiles of $Y^1$ and $Y^0$. Here (as in the case with $\tau = -1.0$) the distribution of estimated impacts is implausibly wide with large positive values in each distribution often matched with zero or small positive values in the other. However, the conclusion that a majority of adult female participants benefit from the JTPA programme is robust to the choice of $\tau$.

Even though many joint distributions of outcomes are consistent with the marginals produced from a social experiment, one model is not: common effect model equation (A). Heckman et al. (1997) test and reject the assumption that $\Delta(= \alpha_1)$ is a common coefficient, using a variety of conditioning sets. Heterogeneity is a central feature of the data, even within narrowly defined

Table 2

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Perfect positive dependence ($\tau = 1.0$)</th>
<th>Positive dependence ($\tau = 0.95$)</th>
<th>Independence of $Y^1$ and $Y^0$ ($\tau = 0.0$)</th>
<th>Perfect negative dependence ($\tau = -1.0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5th percentile</td>
<td>0.00</td>
<td>0.00</td>
<td>-18,098.50</td>
<td>-22,350.00</td>
</tr>
<tr>
<td></td>
<td>(47.50)</td>
<td>(360.18)</td>
<td>(630.73)</td>
<td>(547.17)</td>
</tr>
<tr>
<td>25th percentile</td>
<td>572.00</td>
<td>125.50</td>
<td>-6,043.00</td>
<td>-11,755.00</td>
</tr>
<tr>
<td></td>
<td>(232.90)</td>
<td>(124.60)</td>
<td>(300.47)</td>
<td>(411.83)</td>
</tr>
<tr>
<td>50th percentile</td>
<td>864.00</td>
<td>616.00</td>
<td>0.00</td>
<td>580.00</td>
</tr>
<tr>
<td></td>
<td>(269.26)</td>
<td>(280.19)</td>
<td>(163.17)</td>
<td>(389.51)</td>
</tr>
<tr>
<td>75th percentile</td>
<td>966.00</td>
<td>867.00</td>
<td>7,388.50</td>
<td>12,791.00</td>
</tr>
<tr>
<td></td>
<td>(305.74)</td>
<td>(272.60)</td>
<td>(263.25)</td>
<td>(253.18)</td>
</tr>
<tr>
<td>95th percentile</td>
<td>2,003.00</td>
<td>1,415.50</td>
<td>19,413.25</td>
<td>23,351.00</td>
</tr>
<tr>
<td></td>
<td>(543.03)</td>
<td>(391.51)</td>
<td>(423.63)</td>
<td>(341.41)</td>
</tr>
<tr>
<td>Percent positive</td>
<td>100.00</td>
<td>96.00</td>
<td>54.00</td>
<td>52.00</td>
</tr>
<tr>
<td></td>
<td>(1.60)</td>
<td>(3.88)</td>
<td>(1.11)</td>
<td>(0.81)</td>
</tr>
<tr>
<td>Impact std dev</td>
<td>1,857.75</td>
<td>6,005.96</td>
<td>12,879.21</td>
<td>16,432.43</td>
</tr>
<tr>
<td></td>
<td>(480.17)</td>
<td>(776.14)</td>
<td>(259.24)</td>
<td>(265.88)</td>
</tr>
<tr>
<td>Outcome</td>
<td>0.9903</td>
<td>0.7885</td>
<td>-0.0147</td>
<td>-0.6592</td>
</tr>
<tr>
<td>Correlation</td>
<td>(0.0048)</td>
<td>(0.0402)</td>
<td>(0.0106)</td>
<td>(0.0184)</td>
</tr>
</tbody>
</table>

* $\tau$ is the rank correlation coefficient.

Source: Heckman et al. (1997)

25 Heckman et al. (1997) present methods for allowing for mass points of zero earnings in the population, and present some evidence derived from such methods. Their qualitative conclusions on variability are similar to mine.
demographic categories. Evaluative criteria based on homogeneity ignore a central feature of the data.

4.2. *Assuming the Gain is Independent of the Base*

Heckman *et al.* (1997) and Heckman and Smith (1998), estimate Model (B) assuming that $\Delta$ is not known at the time decisions are made and that $Y^0$ is independent of $\Delta$. This is one version of a ‘Veil of Ignorance’ assumption. They find that $\text{Var}(\Delta) > 0$. Using both normal models and nonparametric models, they estimate that 43% of the women benefit from the programme. This estimate is suspect, however, because of the evidence reported in Heckman *et al.* (1998) in favour of Model (C), which is evidence against this version of the ‘Veil of Ignorance’ model.

4.3. *Testing for Ex Ante Stochastic Rationality of Participants*

If individuals choose whether to participate in a programme based on the gross gains from it, if they possess concave utility functions (not necessarily the same across persons) and if they know the marginal distribution of outcomes in the participation and non-participation states, but not necessarily their place in the distribution, then a second-order stochastic dominance criterion orders the outcomes. For non-negative $y^1$, $y^0$, this form of rationality implies that

$$\int_0^\alpha F_1(y^1)dy^1 < \int_0^\alpha F_0(y^0)dy^0$$  \hspace{1cm} (15)

Draws from the $Y^1$ distribution produce higher expected utility than draws from the $Y^0$ distribution. When applied to persons randomised out of a programme, the difference between the two integrals is a measure of regret among persons randomised out from the programme and forced into the no-treatment state. This condition may fail for many reasons: persons may possess more information about their potential outcomes than just the marginal distributions, or persons may participate in the programme on a principle other than expected utility formulated in terms of gross outcomes. Condition (15) is a sufficient condition. Agents might still prefer distribution 1 even if it is not satisfied. Thus failure to reject (15) is informative; rejection is not.

Heckman and Smith (1993, 1998) test this condition using the JTPA data. The data for adult males and females provide strong evidence of rational behaviour in the sense of (15), suggesting that personal objectives and programme objectives are aligned for adult men and women. Results for youth are mixed.

4.4. *Evidence from Self-assessments of Programme Participants*

Self-assessments of programme participants are an alternative to comparisons of observed outcomes as a measure of programme impact. Unlike the *ex ante* measures based on second-order stochastic dominance, these measures are
statements about *ex post* evaluations. There is no reason why the two types of measures should agree if people revise their assessments based on what they learn about a programme by participating in it. This section considers the strengths and limitations of self-reported assessments of satisfaction with the programme as an evaluation criterion, and reports on self-evaluations by participants in the JTPA experimental treatment group. It also considers what can be learned from self-assessment data regarding the heterogeneity of individual treatment effects and the rationality of programme participants.

Using participant assessments to evaluate a programme has two advantages relative to the approaches previously discussed. First, participants have information not available to external programme evaluators. They typically know more about certain components of the cost of programme participation than do evaluators. Most evaluations, including the National JTPA Study, do not even attempt to value participant time, transportation, child care or other costs in evaluating programme effectiveness, unless they are paid by the programme through subsidies. Participants are likely to include such information in arriving at their self-assessments of the programme. Second, participant evaluations provide information about the values placed on outcomes by participants relative to their perceived cost. They have the potential of providing a more inclusive measure of the programme’s effects than would be obtained from looking only at gross outcomes – one that includes ‘client satisfaction’. To some parties in the welfare state, ‘customer satisfaction’ is an important aspect of a programme. An input-based measure of programme evaluation and not an output-based measure may be appropriate if outputs are hard to measure.

However, participant self-assessments may not be informative on the outcomes of interest to other parties in the welfare state. In evaluations of medical interventions, for example, treatment effects may not be observed by participants or may be difficult for them to assess compared to what observing scientists might report. Participant assessments of the counterfactual state may be faulty because participants’ judgements are based on inputs or on outcome levels rather than gains over alternative levels. Persons who choose to go into the programme may rationalise their participation in it by responding to questions in a certain way. In addition, self-assessments, like all utility-based measures, are difficult to compare across individuals.

Table 3 reports JTPA participant responses to a question about whether the programme made them better off. Assuming people answer honestly, and are reporting a gross impact, the self-assessment data contradict the hypothesis of impact homogeneity. For all four demographic groups, 65–70% of self-reported participants give a positive self-assessment, not the 100% or 0% predicted if impacts were homogeneous. The entries in the third row of

26 The exact wording of the survey question is ‘Do you think that the training or other assistance you got from the programme helped you get a job or perform better on the job?’. The question is asked only of treatment group members who report receiving JTPA services.

27 However, if respondents are reporting a perceived net impact, the evidence reported in Table 3 does not necessarily contradict an assumption of gross impact homogeneity if there is heterogeneity in costs across participants.
Table 3 reveal that the fractions reporting a positive impact are far lower than those obtained from the analyses based on outcome data. This evidence is consistent with the hypothesis that respondents are reporting net outcomes and that costs borne by participants are a substantial fraction of gross outcomes.

For adult women, consider how well the self-assessment data match up with the analyses considered in earlier sections. The self-assessment data are not consistent with the assumption of perfect positive dependence in outcomes across the two states. As shown in Fig. 1, for adult women, the JTPA data indicate that perfect positive dependence in outcomes between the treated and untreated states implies a strictly positive impact of the programme for about 85% of participants – all except those with zero earnings in both states. This value far exceeds the overall self-reported effectiveness rate of 44% reported in row 3 of Table 3. The 44% rate lies below that found even for the case of perfect negative dependence. Overall, the self-reported impact data appear to be too negative when compared to our analyses of the experimental earnings data. This evidence is consistent with participants reporting a net measure while the experimental ‘treatment effect’ measures gross outcomes. The lower positive rating of the programme from self-assessment data than from gross outcome data is all the more striking given that the self-assessments are recorded only for people who report receiving training while the gross outcome data for participants include those who leave the programme, and the attriters have lower earnings than the non-attriters.

Heckman and Smith (1998) present additional evidence on the JTPA programme using the revealed preferences of programme dropouts. They document substantial heterogeneity in participant evaluations of the programme using this information.

Table 3

<table>
<thead>
<tr>
<th>National JTPA Study 18-Month Impact Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full sample percentages</strong></td>
</tr>
<tr>
<td><strong>Percent who self-report participating</strong></td>
</tr>
<tr>
<td>61.63</td>
</tr>
<tr>
<td><strong>Percent of self-reported participants with a positive self-assessment</strong></td>
</tr>
<tr>
<td><strong>Overall percent with positive self-assessments</strong></td>
</tr>
</tbody>
</table>

Source: Heckman and Smith (1998)
4.5. Summary of the Evidence on Impact Heterogeneity and Its Consequences

Table 4 presents a summary of the main findings of the analysis of the JTPA data.

1. Under a variety of different assumptions, there is evidence of substantial heterogeneity in net impacts, $\Delta$.

2. The analysis of self-assessments suggests that respondents are reporting different impacts from the ‘objective’ impacts determined from experimental data. This is a further source of heterogeneity and a source of disparity across studies. Nonetheless, this net measure is informative of how participants value the programme just after they have completed it.

3. Departures from high levels of positive dependence between $Y^0$ and $Y^1$ produce absurd ranges of impacts on gross outcomes. (The implicit correlations between $Y^0$ and $Y^1$ produced under different identifying assumptions are given in the last column of the table.)

4. The range of the estimated proportion of people benefiting from the programme in the sense of gross outcomes (the ‘voting criterion’) varies widely under different assumptions about the dependence in outcomes. Nonetheless, over the entire range of values, the majority of people benefit from the programme. The data from the self-report show a lower proportion benefiting – a phenomenon consistent with the hypothesis that net returns and not gross returns are being reported by participants.

This evidence speaks strongly against Model (A) and in favour of Models (B) and (C). Evidence reported by Heckman et al. (1998) suggests that Model (B) is rejected in favour of Model (C). Not only are responses to treatment heterogeneous, but people act on these responses (i.e. they select into the programme based on them). In actual behaviour, people are not participating in the programme under a veil of ignorance.

Carneiro et al. (2001), Heckman (2001) and Heckman and Vytlacil (2001) summarise the empirical literature on the relevance of Models (A), (B) and (C) for studies of the wage returns to unionism, job training, migration and education. Except for studies of unionism, the bulk of the literature rejects common effect Model (A), and favours (C) over (B). Thus the evidence reported here from the JTPA programme is consistent with a broader literature. Heterogeneity in outcomes on which people act is a central feature of microdata.

5. Accounting for General Equilibrium and Heterogeneity in Evaluating Human Capital Policies

A major limitation of the microeconomic treatment effect literature is its failure to consider the general equilibrium consequences of the social programmes being evaluated. Many human capital policies are large scale in

---

28 Positive dependence between $Y^0$ and $Y^1$ for persons with $D = 1$ is implied by economic models of self selection. This narrows the range of possible dependence across outcome states.

© Royal Economic Society 2001
Table 4  
**Summary of Empirical Evidence on Impact Heterogeneity, the Voting Criterion and the Dependence Between \( Y^1 \) and \( Y^0 \)**

*National JTPA Study 18-Month Experimental Impact Sample*

<table>
<thead>
<tr>
<th>( Y^1 )</th>
<th>( Y^0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description of analysis</td>
<td>Evidence of heterogeneity?</td>
</tr>
<tr>
<td>Frechet bounds</td>
<td>Statistical bounds on the joint distribution of outcomes, ( F(y_0, y_1</td>
</tr>
<tr>
<td>Perfect positive percentile dependence</td>
<td>Assumes ( q_1 = q_0 ) where ( q_1 ) is a percentile of ( Y^1 ) given ( D = 1 ) and ( q_0 ) is a percentile of ( Y^0 ) given ( D = 1 ). Conditional on ( D = 1 ), the counterfactual for each percentile in the ( Y_1 ) distribution is the same percentile in the ( Y_0 ) distribution.</td>
</tr>
<tr>
<td>Perfect negative percentile dependence</td>
<td>Assumes ( q_1 = 100 - q_0 ) where ( q_1 ) is a percentile of ( Y_1 ) given ( D = 1 ) and ( q_0 ) is a percentile of ( Y_0 ) given ( D = 1 ). Conditional on ( D = 1 ), the counterfactual for each ( q )th percentile of the ( Y_1 ) distribution is the ( 100 - q )th percentile in the ( Y_0 ) distribution.</td>
</tr>
<tr>
<td>Positive percentile dependence with rank correlation ( \tau = 0.95 )</td>
<td>Assumes that the percentiles of ( Y^1 ) and ( Y^0 ) given ( D = 1 ) have a rank correlation of 0.95. Estimates are based on a random sample of 50 such permutations.</td>
</tr>
<tr>
<td>Independence of percentiles of ( Y_1 ) and ( Y_0 ), which implies a percentile rank correlation ( \tau ) of 0.0</td>
<td>Assumes that the percentiles of ( Y^1 ) and ( Y^0 ) given ( D = 1 ) have a rank correlation of 0.0, which is implied by independence between them. Estimates are based on a random sample of 50 such permutations.</td>
</tr>
<tr>
<td>Random coefficient model</td>
<td>Assumes that ( \Delta \perp Y^0</td>
</tr>
<tr>
<td>Deconvolution</td>
<td>Assumes that ( \Delta \perp Y^0</td>
</tr>
<tr>
<td>Self-assessments</td>
<td>\textit{Ex post} self-evaluations by participants based on a survey question regarding whether the programme provided a benefit.</td>
</tr>
<tr>
<td>Dropouts</td>
<td>Attrition decisions after application and acceptance into the programme.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Standard deviation of impacts</th>
<th>Evidence on voting criterion</th>
<th>Dependence between ( Y^1 ) and ( Y^0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frechet Bounds</td>
<td>Bounded between $675) and $14,969</td>
<td>Product-moment correlation ( \rho ) between ( Y^1 ) and ( Y^0 ) bounded between (-0.760) and 0.998</td>
</tr>
</tbody>
</table>

Bounds do not apply to the indicator function \( I(Y^1 > Y^0) \) as this function is not super- or sub-additive. Thus, the voting criterion cannot be bounded.
Table 4
(Continued)

<table>
<thead>
<tr>
<th>Standard deviation of impacts</th>
<th>Evidence on voting criterion</th>
<th>Dependence between $Y^1$ and $Y^0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect positive percentile dependence</td>
<td>$1,857$</td>
<td>100% of participants benefit or are indifferent.</td>
</tr>
<tr>
<td>Perfect negative percentile dependence</td>
<td>$16,432$</td>
<td>52% positive</td>
</tr>
<tr>
<td>Positive percentile dependence with rank correlation $\tau = 0.95$</td>
<td>Average standard deviation of $1,857$ (with standard deviation of $480$)</td>
<td>Average of 93% positive (with standard deviation of 3.88)</td>
</tr>
<tr>
<td>Independence of percentiles of $Y^1$ and $Y^0$, which implies a percentile rank correlation $\tau$ of 0.0</td>
<td>Average standard deviation of $12,879$ (with standard deviation of $259$)</td>
<td>Average of 54% positive (with standard deviation of 1.11)</td>
</tr>
<tr>
<td>Random coefficient model</td>
<td>Standard deviation is $2,271$</td>
<td>If the random coefficient $\Delta$ is assumed to be normally distributed then 60.45% have positive impacts.</td>
</tr>
<tr>
<td>Deconvolution</td>
<td>Standard deviation is $1,675$</td>
<td>56.35% positive</td>
</tr>
<tr>
<td>Self-assessments</td>
<td>NA</td>
<td>Varies from a low of 39.49% positive for adult men to a high of 47.55% positive for female youth</td>
</tr>
<tr>
<td>Dropouts</td>
<td>NA</td>
<td>Dropping out ranges from a low of 34.83% for male youth to a high of 40.92% for adult males.</td>
</tr>
</tbody>
</table>

Note: NA denotes ‘not applicable’
Source: Heckman and Smith (1998)
nature, and expansion of the stock of skill affects skill prices. Rational agents
will act on that information. This feedback substantially alters the inferences
obtained from microeconomic evaluations.

In addition, many policies do not fall into the ‘treatment effect’ category at
all. A tax on labour earnings affects everyone, although not uniformly. There
are no natural comparison groups (or control groups) for policies with
universal coverage.

Conventional general-equilibrium analysis ignores heterogeneity among
agents and so is poorly suited to analyse distributional issues. An important
exception is the class of overlapping generations models which explicitly
consider inequality among generations. The goal of this section of the paper is
to summarise research by myself, Lochner and Taber (Heckman et al., 1998a,b,
1999) that extends the empirical overlapping generations model of Auerbach
and Kotlikoff (1987) by

(a) allowing for human capital, and
(b) introducing heterogeneity in ability within cohorts.

Ability is a major determinant of human capital investment, and adding it to a
model, along with human capital, enables one to develop a model of human
capital and wage formation that can explain rising wage inequality and
inequality within narrow schooling groups. This model provides a framework
which accounts for heterogeneity and diversity and which enables one to
answer the evaluation questions posed in section 1 for a dynamic economy. It
is a vehicle for examining the performance of micro evaluation methods
within a general equilibrium setting.

Heckman et al. (1998a) formulate and estimate dynamic general equili-
brium models with endogenous heterogeneous human capital accumulation.
Their model explains rising wage inequality in the US economy. This section
of the paper uses their model to study the impacts on skill formation of
proposals to switch from progressive taxes to flat income and consumption
taxes. First, a brief review of the main ingredients of their model:

5.1. **A Dynamic General Equilibrium Model of Human Capital Accumulation with
Heterogeneous Agents**

Heckman et al. (1998a) build on the model of Auerbach and Kotlikoff (1987)
in three ways:

1. They introduce skill formation and consider both schooling choices and
   investment in on-the-job training.
2. They allow for heterogeneity in ability, endowments and skills. Different
   schooling levels are associated with different skills and different post-school
   investment functions. Heckman et al. replace the Auerbach–Kotlikoff effi-
   ciency units assumption for labour services with a model of heterogeneous
   skills. Models with efficiency units for labour services do not explain rising
   wage inequality among skill groups.
3. Heckman et al. use micro data joined with macro time series evidence from the US economy to determine the parameters of the model, rather than picking parameters in an unsystematic fashion from the micro literature or ‘calibrating’ the model to aggregates, as is commonly done in the empirical general equilibrium literature.

The Heckman et al. (1998a) model has three sources of heterogeneity among persons:

(a) in age
(b) in ability to learn and in initial endowments of ability and human capital (indexed by \( \theta \) below), and
(c) in the economic histories experienced by cohorts.

In a transition period, different cohorts face different skill prices, make different investment decisions and, hence, accumulate different amounts of human capital and have different wage levels and trajectories. The Heckman et al. model extends the analysis of Davies and Whalley (1991) and Lord (1989) who introduce human capital into the Auerbach–Kotlikoff model but assume only one skill. Heckman et al. allow for multiple skills, incorporate both schooling and on-the-job training, and allow for rational expectations in calculating transition paths.

In the Heckman et al. model, individuals live for \( \bar{a} \) years and retire after \( a_R \leq \bar{a} \) years. In the first stage of the lifecycle, a prospective student chooses the schooling option that gives him (or her) the highest level of lifetime utility. Define \( K^S_{a,t} \) as the stock of physical capital held at time \( t \) by a person age \( a \); \( H^S_{a,t} \) is the stock of human capital at time \( t \) of type \( S \) at age \( a \) with schooling \( S \). The optimal lifecycle problem can be solved in two stages: first, condition on schooling \( S \) and solve for the optimal path of consumption \( (C^S_{a,t}) \) and post-school investment time \( (I^S_{a,t}) \) for each schooling level, second, select among schooling levels to maximise lifetime welfare.

Given \( S \), an individual age \( a \) at time \( t \) has the value function

\[
V_{a,t}(H^S_{a,t}, K^S_{a,t}, S) = \max_{C_{a,t}, I^S_{a,t}} \frac{(C^S_{a,t})^\gamma - 1}{\gamma} + \delta V_{a+1,t}^S(H^S_{a+1,t+1}, K^S_{a+1,t+1}, S) \tag{16}
\]

where \( \delta \) is a time preference discount factor. Heckman et al. follow Kotlikoff et al. (1997), by assuming that the tax schedule can be approximated by a progressive tax on labour income and a flat tax on capital income. This gives a dynamic budget constraint,

\[
K^S_{a+1,t+1} \leq K^S_{a,t}[1 + (1 - \tau_k) r_t] + R^S_t H^S_{a,t}(1 - I^S_{a,t}) - \tau_l R^S_t H^S_{a,t}(1 - I^S_{a,t}) - C^S_{a,t} \tag{17}
\]

where \( \tau_k \) is the proportional tax rate on capital, \( \tau_l \) is the progressive tax schedule on labour earnings, \( R^S_t \) is the price of human capital services of type \( S \) at time \( t \), and \( r_t \) is the net return on physical capital at time \( t \). Experiments with other progressive tax schedules produce results similar to the ones reported here. Heckman et al. abstract from labour supply. Estimates of
intertemporal substitution in labour supply estimated on annual data are small, so ignoring labour supply does not affect our analysis (Browning et al., 1999). This simplification makes the Heckman et al. model comparable to that of Davies and Whalley, who also ignore leisure.

On-the-job human capital for a person of schooling level $S$ accumulates through the human capital production function

$$H_{a+1,t+1}^S = A^S(\theta) I_{at}^{\alpha_S} H_{at}^{\beta_S} + (1 - \sigma^S) H_{at}^S$$

where the conditions $0 < \alpha_S < 1$ and $0 \leq \beta_S \leq 1$ guarantee that the problem is concave, and $\sigma^S$ is the rate of depreciation of skill-$S$ specific human capital. ‘$\theta$’ is an ability or heterogeneity factor, such that different people have different abilities to learn. This functional form is widely used in both the empirical literature and the literature on human capital accumulation. The $\alpha$ and $\beta$ are also permitted to be $S$-specific, which emphasises that schooling affects the process of learning on the job in a variety of different ways. This model distinguishes between human capital that is primarily useful in earning and human capital that is also useful in learning.

Notably absent from the model are short-run credit constraints that are often featured in the literature on schooling and human capital accumulation. This model is consistent with the evidence presented in Cameron and Heckman (1999, 2001) and Cameron and Taber (2000) that long-run family factors correlated with income (the $\theta$ operating through $A^S(\theta)$ and the initial condition for (18)) affect schooling, but that short-term credit constraints are not empirically important. Such long-run factors account for the empirically well-known correlation between schooling attainment and family income.

At the beginning of life, agents choose the value of $S$ that maximises lifetime utility as in a Roy model:

$$\hat{S} = \arg\max_S [PV^S(\theta) - D^S + \epsilon^S]$$

where $PV^S(\theta)$ is the tax-adjusted present value of lifetime earnings given schooling level $S$, tuition costs are $D^S$, and $\epsilon^S$ represents monetised nonpecuniary benefits of schooling level $S$, or else unobserved components of tuition subsidies (negative costs). All values and costs are discounted back to the beginning of life.

Tuition costs are permitted to change over time so that different cohorts face different schooling costs. The economy is assumed to be competitive so that the prices of skills and capital services are determined as derivatives of an aggregate production function. To compute service flow prices for capital and the different types of human capital, it is necessary to construct aggregates for

---

29 Because of the separation between consumption and investment, the decision to go to school can be formulated in terms of comparisons among present values of earnings.

© Royal Economic Society 2001
each of the factors over each of the ability types and over all cohorts to insert into an aggregate production function.

Post-school human capital of type $S$ is a perfect substitute for post-school human capital of the same schooling type, whatever the age or experience level of the agent, but it is not perfectly substitutable with human capital from other schooling levels. In this model, cohorts differ from each other only because they face different price paths and policy environments within their lifetimes.

The aggregate production function exhibits constant returns to scale. The equilibrium conditions require that marginal products equal pre-tax skill prices. In the two-skill economy Heckman, Lochner and Taber analyse, the production function at time $t$ is defined over the inputs $\overline{H}^1_t$, $\overline{H}^2_t$ and $\overline{K}_t$, where $\overline{H}^1_t$ and $\overline{H}^2_t$ are aggregates of utilised skills (high school and college, respectively) supplied to production, and $\overline{K}_t$ is the aggregate stock of capital. The technology is

$$F(\overline{H}^1_t, \overline{H}^2_t, \overline{K}_t) = a_3\left\{a_2[a_1(\overline{H}^1_t)^{\rho_1} + (1 - a_1)(\overline{H}^2_t)^{\rho_1}]^{\rho_2/\rho_1} + (1 - a_2)\overline{K}_t^{\rho_2}\right\}^{1/\rho_2}$$

Heckman et al. estimate that $\rho_2 = 0$ but $\rho_1 = 0.693$, which yields an elasticity of substitution between high school and college human capital of 1.441. Heckman et al. explore both open economy (world capital market) and closed economy versions of their model. The latter produces estimates of aggregates closer to data from the US economy and that version is used here.

Human capital accumulation functions (18) are estimated using micro data assuming that taxes are proportional. However, an extensive sensitivity analysis reveals that within the range of the data for the US economy, misspecification of the tax system does not affect parameter estimates if the model is recalibrated on aggregate data (Heckman et al., 1998a, Appendix B). Heckman et al. (1998a) also present an array of sensitivity checks to alternative specifications of their model and find that their estimates are robust to alternative identifying assumptions. The Heckman et al. model is now used to evaluate the effects of alternative tax policies on efficiency and distribution and the consequences of accounting for heterogeneity and general equilibrium in evaluating social policies are considered.

5.2. Tax Effects on Human Capital Accumulation

In the absence of labour supply and direct pecuniary or nonpecuniary costs of human capital investment, there is no effect of a proportional wage tax on human capital accumulation. Both marginal returns and costs are scaled down in the same proportion. When untaxed costs or returns to college are added to the model (ie non-pecuniary costs/benefits), proportional taxation is no longer neutral. An increase in the tax rate decreases college attendance if the net financial benefit before taxes is positive. Letting $S = 1$ denote college and $S = 0$ denote high schooling, a person goes to college if $PV^1 - D^1 - PV^0 > 0$ where $PV$ denotes discounted (to age 0) earnings. Progressivity reinforces this
effect. A progressive wage tax reduces the incentive to accumulate skills, since human capital promotes earnings growth and moves persons to higher tax brackets. As a result, marginal returns on future earnings are reduced more than marginal costs of schooling.

Heckman (1976) notes that, in a partial-equilibrium model, proportional taxation of interest income with full deductibility of all borrowing costs reduces the after-tax interest rate and, hence, promotes human capital accumulation. In a time-separable, representative-agent general-equilibrium model, the after-tax interest rate is unaffected by the tax policy in steady state as agents shift to human capital from physical capital (Trostel, 1993). In that framework, flat taxes with full deductibility have no effect on human capital investment. In a dynamic overlapping generations model with heterogeneous agents and endogenous skill formation and with progressive rates, taxes have ambiguous effects on human capital and both their quantitative and qualitative effects can only be resolved by empirical research. I use the empirically grounded model of Heckman et al. to study alternative proposals for tax reform, their consequences for inequality, and their ability to construct the policy counterfactuals discussed in section 1.

5.3. Analysing Two Tax Reforms

Following Kotlikoff et al. (1997), I assume that the US income tax can be captured by a progressive tax on labour income and a flat tax on capital income. Each earner has 1.22 children and is single. For each additional dollar beyond $9,660, there is an increase in itemised deductions of 7.55 cents. An individual with labour income $Y$ has taxable income $(Y - 9660)(1 - 0.0755)$.

Using the 1995 tax schedule, the taxes paid on income are computed and approximated by a second-order polynomial. A 0.15 flat tax rate on physical capital is assumed.

Consider two revenue-neutral tax reforms from this benchmark progressive schedule. The first reform (which I call ‘Flat Tax’) is a revenue-neutral flattening of the tax on labour earnings, holding the initial flat tax on capital income constant. The second reform (‘Flat Consumption Tax’) is a uniform flat tax on consumption. In both flat tax schemes, tuition is not treated as deductible. (The consequences of making it deductible are discussed below.)

For each tax, consider two models:

1. A partial-equilibrium model in which skill prices and interest rates are fixed
2. A closed-economy general-equilibrium model where skill prices and interest rates adjust

A tax policy with universal coverage does not produce natural ‘comparison’ or ‘control’ groups. For that reason, I do not consider estimates based on

---

30 This section is based in part on Heckman, Lochner, Taber (1998b).
methods from the ‘treatment effect’ literature because that approach is ineffective in this context.

Table 5 presents both partial-equilibrium and general-equilibrium results measured relative to a benchmark economy with Kotlikoff et al.’s tax schedule.

Table 5
Closed Economy Effects of Alternative Tax Proposals

<table>
<thead>
<tr>
<th>Percentage Difference from Progressive Case†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>After tax interest rate</td>
</tr>
<tr>
<td>Skill price college HC</td>
</tr>
<tr>
<td>Skill price HS HC</td>
</tr>
<tr>
<td>Stock of physical capital</td>
</tr>
<tr>
<td>Stock of college HC</td>
</tr>
<tr>
<td>Stock of HS HC</td>
</tr>
<tr>
<td>Stock of college HC per college graduate</td>
</tr>
<tr>
<td>Stock of HS HC per HS graduate</td>
</tr>
<tr>
<td>Aggregate output</td>
</tr>
<tr>
<td>Aggregate consumption</td>
</tr>
<tr>
<td>Mean wage college</td>
</tr>
<tr>
<td>Mean wage HS</td>
</tr>
<tr>
<td>Standard deviation log wage</td>
</tr>
<tr>
<td>Gini coefficient</td>
</tr>
<tr>
<td>College/HS wage premium at 10 yrs exp§</td>
</tr>
<tr>
<td>Fraction attending college</td>
</tr>
<tr>
<td>Type 1††: Fraction attending college</td>
</tr>
<tr>
<td>Type 2††: Fraction attending college</td>
</tr>
<tr>
<td>Type 3††: Fraction attending college</td>
</tr>
<tr>
<td>Type 4††: Fraction attending college</td>
</tr>
<tr>
<td>Type 1‡‡: College HC gain first 10 years**</td>
</tr>
<tr>
<td>Type 2‡‡: College HC gain first 10 years**</td>
</tr>
<tr>
<td>Type 3‡‡: College HC gain first 10 years**</td>
</tr>
<tr>
<td>Type 4‡‡: College HC gain first 10 years**</td>
</tr>
<tr>
<td>Type 1‡‡: HS HC gain first 10 years**</td>
</tr>
<tr>
<td>Type 2‡‡: HS HC gain first 10 years**</td>
</tr>
<tr>
<td>Type 3‡‡: HS HC gain first 10 years**</td>
</tr>
<tr>
<td>Type 4‡‡: HS HC gain first 10 years**</td>
</tr>
</tbody>
</table>

* General equilibrium (GE) effects allow skill prices to change, while partial equilibrium (PE) effects hold prices constant.
† In the progressive case, we allow for a progressive tax on labour earnings, but assume a flat tax on capital at 15%.
‡ In the flat tax regime, we hold the tax on capital fixed to the same level as the progressive tax, but the tax on labour income is flat and is calculated to balance the budget in the new GE steady state. This yields a tax rate on labour income of 7.7%. In the consumption regime, we tax only consumption at a 10.0% rate, again balancing the budget in steady states.
§ The college–high school wage premium measures the difference in log mean earnings between college graduates and high school graduates with ten years of experience.
** These rows present changes in the ratio of human capital at ten years of experience versus human capital on entering the labour force.
†† Types correspond to quartiles of the ability distribution with Type 1 being the lowest ability and Type 4 being the highest ability.
Source: Heckman et al. (1999)
I first discuss the partial-equilibrium effects of a move to a ‘Flat Tax’, which eliminates progressivity in wages and stimulates skill formation. College attendance rises dramatically as the higher earnings associated with college graduation are no longer taxed away at higher rates. The amount of post-school on-the-job training also increases for each skill group as measured by the stocks of human capital per worker of each skill. The aggregate stock of high school human capital declines while the aggregate stock of college human capital increases as a result of the rise in college enrolment. The college–high school wage differential increases slightly as does another widely used measure of inequality – the standard deviation of log wages. The effects of the reform on aggregates of consumption and output are modest at best. However, capital formation is greatly reduced as the tax code now favours human capital compared to the benchmark economy.

In general equilibrium, the effects of the reform on skill formation are, in general, qualitatively similar, but they are greatly diminished. The effects on aggregate consumption and output are weak, as they are in the partial-equilibrium case. Furthermore, the negative effects of the reform on physical capital are muted, since the return to capital increases. The rise in the after-tax interest rate chokes off skill investment. *Per capita* post-school on-the-job training accumulation still increases for both skill groups, although the increase is dampened compared to the partial-equilibrium case. Aggregate stocks of both high school and college human capital now rise, since college enrolment increases much less. The distinction between partial equilibrium and general equilibrium is especially striking for the fraction attending college. In general equilibrium, college attendance increases only for the most able, whereas in the partial-equilibrium case, it increases for all ability groups. Changes in skill prices and interest rates virtually offset the removal of the disincentives of progressive taxes on schooling enrolment. The college–high school wage differential (at 10 years of experience) now declines slightly, and the increase in the standard deviation of log wages is less. The Gini coefficient, which is the preferred measure in modern welfare economics based on the Harsanyi-Vickery criterion (Sen, 1973; 2000) is ordered in the same way.\textsuperscript{31} By this measure of welfare, flat tax reform is not to be preferred. In general equilibrium, the increase in the standard deviation is smaller, because skill prices adjust and because higher after-tax interest rates flatten wage profiles.

Figs 2\textsuperscript{a} and 2\textsuperscript{b} show how the reform affects the utility of the current generation. It lowers the overall utility of the least able and the least schooled, and raises the overall utility of the most able and the most skilled. This is a consequence of the pro-human-capital bias of the tax reform and the interaction between ability and human capital in producing human

\textsuperscript{31} Recall that second-order stochastic dominance and Gini ranking are equivalent. See Rothschild and Stiglitz (1970).
capital. On a strict voting criterion for those in the current generation, the reform would not pass (43% in favour; 57% against). Evaluated at the final steady state, the reform would not be favoured. (See the numbers in Table 6.) For this reform, the voting criterion and the social welfare ranking based on anonymity agree.

Next, consider a move to a flat consumption tax. This reform is more pro-capital and is less favourable to human capital. The partial equilibrium analysis reveals that it raises output, capital and consumption more than a ‘Flat Tax’ reform, and it reduces the aggregate stock of high skill human capital and the stock of human capital per worker for each skill group. The fraction attending college declines. The reform raises wage inequality as measured by the college–high school wage premium but lowers it as measured by the standard deviation of log wages, and in terms of the Gini coefficient.

In general equilibrium, this reform is slightly less favourable to human capital formation than the ‘Flat Tax’, since the after-tax rate of return on capital rises more. College attendance increases slightly, but the increase is concentrated among the least and most able persons. Wage inequality
Fig. 2. (Continued)
Source: Simulations based on Heckman et al. (1998a, b, 1999).

Table 6

<table>
<thead>
<tr>
<th></th>
<th>Movement to flat income tax</th>
<th>Movement to flat consumption tax</th>
</tr>
</thead>
<tbody>
<tr>
<td>% in favour in initial state</td>
<td>43%</td>
<td>52%</td>
</tr>
<tr>
<td>Final Steady State Utility Gain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School ability 1:</td>
<td>−0.61</td>
<td>0.27</td>
</tr>
<tr>
<td>High School ability 2:</td>
<td>−0.20</td>
<td>0.71</td>
</tr>
<tr>
<td>High School ability 3:</td>
<td>−0.13</td>
<td>0.78</td>
</tr>
<tr>
<td>High School ability 4:</td>
<td>0.09</td>
<td>0.93</td>
</tr>
<tr>
<td>College ability 1:</td>
<td>−0.53</td>
<td>0.35</td>
</tr>
<tr>
<td>College ability 2:</td>
<td>−0.32</td>
<td>0.58</td>
</tr>
<tr>
<td>College ability 3:</td>
<td>−0.18</td>
<td>0.72</td>
</tr>
<tr>
<td>College ability 4:</td>
<td>0.23</td>
<td>1.16</td>
</tr>
<tr>
<td>Ability 1</td>
<td>−0.57</td>
<td>0.30</td>
</tr>
<tr>
<td>Ability 2</td>
<td>−0.28</td>
<td>0.64</td>
</tr>
<tr>
<td>Ability 3</td>
<td>−0.03</td>
<td>0.85</td>
</tr>
<tr>
<td>Ability 4</td>
<td>0.11</td>
<td>1.05</td>
</tr>
</tbody>
</table>

Ability 1 is the lowest ability and Ability 4 is the highest as measured by the AFQT distribution quartiles. Simulations based on the closed economy estimates of Heckman et al. (1999).

© Royal Economic Society 2001
increases slightly by both conventional measures. Using Gini dominance, under the Harsanyi–Vickery criterion, the reform would not be favoured. Real wages rise for both skill groups, and the rise is greater than in the ‘Flat Tax’ reform. This is due to a larger increase in capital under proportional consumption taxation. Since capital is a direct complement with both forms of human capital, and there is no evidence of skill bias in this complementarity relationship, the increase in capital raises skill prices about equally for both skill groups. The greater increase in real wages in this case is not due to a larger increase in per capita human capital accumulation within skill groups.

Figs 3a and 3b and Table 6 reveal that across ability and schooling groups, the consumption tax reform produces more winners among contemporary voters than does the flat tax reform. On a voting criterion, the consumption reform would be favoured (52% in favour; 48% against). Comparing steady states, the reform would be passed by a substantial majority. The short-run

![Fig. 3. Changes in Utility from the Reform in the Current Generation: Consumption Tax.](image-url)
and long-run voting criteria and the Harsanyi–Vickery criterion are in conflict.\textsuperscript{32,33}

The lessons from partial-equilibrium analyses are substantially misleading guides in analysing the effects of tax policy on skill formation. Changes to proportional taxation are unlikely to have large effects on skill formation or output. A change to a flat consumption tax has the largest effect on output, consumption, and real wages, but it also slightly raises wage inequality. These

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig3}
\caption{(Continued)}
\label{fig:3}
\end{figure}

\begin{footnotesize}
\textsuperscript{32} When the same two policies are evaluated in the case with 30 years of skill biased technical change of the magnitude that is consistent with US data (Heckman et al., 1998\textsuperscript{a}), both tax reforms are favoured by all groups by the voting criterion, both those in the current generation and those in the steady state. The voting criterion and the Harsanyi–Vickrey criteria do not agree in the analysis of the consumption tax reform. The tax reform expedites the transition to the new economy.

\textsuperscript{33} When deductibility of tuition is introduced in both reforms, and revenue neutrality is preserved, there is virtually no effect of this changed tax treatment on skill formation (or anything else) in general equilibrium.
\end{footnotesize}
conclusions also hold for open economy simulations in which the interest rate is set in world markets (Heckman et al., 1999). They are robust to a variety of tax schedules and empirically grounded parameter estimates.

5.4. Tuition Subsidies

Heckman et al. (1998c, 1999) also consider the effect of evaluating the general equilibrium effect of tuition subsidies. Tuition subsidies promote schooling and depress the price of educated labour. Thus assumption (A-1) may be inappropriate because the price of uneducated labour ‘0’ is increased by the policy. They simulate the effects of a revenue-neutral $500 increase in tuition subsidy (financed by a proportional tax) on enrolment in college and wage inequality starting from the baseline economy. The partial-equilibrium model predicts an increase in college attendance of 5.3% in the new steady state. This is in the range of effects reported by Kane (1994) and Cameron and Heckman (1999). This analysis holds skill prices, and therefore college and high school wage rates, fixed – a typical assumption in microeconomic ‘treatment effect’ analyses of tuition policies.

When the policy is evaluated in a general-equilibrium model, the estimated effect falls to 0.49%. Because the college–high school wage ratio falls as more individuals attend college, the returns to college are less than when the wage ratio is held fixed. Rational agents understand this effect of the tuition policy on skill prices and adjust their college-going behaviour accordingly.34

Using the Gini coefficient as a measure of welfare, both partial-equilibrium and general-equilibrium simulations suggest that the reform is welfare improving. So on the basis of the Harsanyi–Vickery criterion the reform would be favoured. On the basis of the voting criterion applied to the current generation, it would not be favoured but it would be favoured by the long-run steady state voter. In general equilibrium, the overall variance of log wages is reduced. Again, the criteria are in conflict with each other.

5.5. General-equilibrium vs Partial-equilibrium Approaches

The sharp contrast between the general-equilibrium estimates of programme impacts and the estimates from partial-equilibrium approaches highlights the potential benefit of applying the general-equilibrium approach to conduct evaluations. Not only is the general-equilibrium approach appropriate for the

34 The steady state results are long-run effects. When Heckman et al. simulate the model with rational expectations, the short-run enrolment effects are also very small, as agents anticipate the effects of the policy on skill prices and calculate that there is little gain from attending college. When they simulate using myopic expectations, the short-run enrolment effects are much closer to the estimated partial-equilibrium effects. Of course, the steady state results are not affected and are large under either myopic or rational expectations. All of these results are qualitatively robust to the choice of different tax schedules. Progressive tax schedules choke off skill investment and lead to even lower enrolment responses in general equilibrium.

© Royal Economic Society 2001
evaluation of programmes with economy-wide effects, it also offers an economically interpretable evaluation of a policy.

Critics of the general-equilibrium approach often dismiss it because it is based on ‘questionable’ empirical foundations. They ignore, or trivialise, the use of economic theory to produce counterfactual policy states, and they assume that empirically credible general-equilibrium models are not possible to construct. Careless calibration exercises that are often used to produce empirical estimates for general equilibrium models have called the entire enterprise of using applied general equilibrium as a tool for policy evaluation into question. Reacting to this casual empiricism, many microeconomists dismiss the general equilibrium enterprise as an empirically unfounded exercise built on weak foundations.

These criticisms ignore the emerging field of empirically grounded general-equilibrium theory that unites microevidence, macro time series and general-equilibrium theory to produce credible parameter estimates of models to evaluate counterfactual states. Browning et al. (1999) summarise current developments in this field and present an agenda for research in unifying micro evidence with macro general-equilibrium models. The empirical analysis presented by Heckman et al. (1998a) uses micro data and macro data to estimate the dynamic general-equilibrium model that is used to present the counterfactual simulations reported in this paper.

There is, no doubt, much room for improvement in producing the empirical foundations used in general-equilibrium models. At the same time, there is room for substantial improvement in the micro ‘treatment effect’ literature that entirely ignores the general-equilibrium consequences of the policies it evaluates. Even if the estimated general-equilibrium effects presented here are scaled down by 50%, they are still substantial. An evaluation literature that ignores price adjustment, and the investment responses to price adjustment, is likely to err substantially in forecasting policy impacts of policies applied to the national level.

6. Summary and Conclusions

Diversity, heterogeneity and involuntary redistribution are the defining features of the modern welfare state. Disagreements over outcomes and how they should be evaluated gives rise to the demand for publicly justified evaluations of social programmes, and the interests of different groups lead to a variety of different criteria being used to assess the impact of public policy.

This paper critically examines the main criteria proposed in the modern literature in welfare economics, in the theory of policy evaluation in macroeconomics, and in cost–benefit analysis. Cost–benefit analysis and macro policy evaluation ignore distributional features of policies treating heterogeneity of programme impacts as either uninteresting or empirically irrelevant. Modern welfare economics explicitly recognises heterogeneity in outcomes but focuses on one measure of alternative policies because of its ‘ethical correctness’ and does not conduct general equilibrium analysis. The anonymity postulates

© Royal Economic Society 2001
proposed in the modern literature on welfare economics ignore self-interested agents who evaluate policies by comparing their initial positions under current policies with their positions under proposed alternative policies or else the postulates argue that such interests should be disregarded in evaluating social policy. A voting criterion that enumerates gainers or losers and quantifies the magnitudes of their losses comes closer to capturing the information useful in evaluating social policies in a modern democracy than does a criterion based on axioms of ‘correct’ behaviour that assume that positions in the distribution of outcomes are independent across policies or that any such dependence should be ignored and that persons (or their elected agents) ignore the initial position in evaluating policies.

This paper considers the information required to implement the various criteria and how different evaluation methods obtain them. Cost–benefit and partial-equilibrium ‘treatment effect’ approaches ignore general-equilibrium considerations which are also ignored in modern welfare economics. General equilibrium is centre stage in the macro policy evaluation model. Virtually all methods currently widely used to assess government policies ignore the heterogeneity in programme impacts that is a major source of demand for evaluations in the welfare state.

Using data from an influential social experiment, I demonstrate that heterogeneity in programme outcomes is an empirically important feature of the data even after conditioning on the observed characteristics. This evidence is consistent with the literature as summarised by Carneiro \textit{et al}. (2001), Heckman (2001) and Heckman and Vytlacil (2001). As a consequence of this heterogeneity, different evaluative criteria produce different assessments of the same policy, reflecting their different vantage points. Combining micro and macro data, I draw on my work with Lochner and Taber to develop an empirically based general-equilibrium model of human capital accumulation that can be used to analyse the consequences of heterogeneity and diversity for the evaluation of social programmes. The benefits of this approach are examined in evaluating several proposed tax reforms.

Accounting for heterogeneity, diversity and general equilibrium has important implications for the way social policies should be evaluated. An evaluation criterion that counts gainers and losers produces a very different assessment of the suitability of policies than the ‘ethically correct’ criteria based on anonymity postulates that are advocated in modern welfare economics, and one that is more closely attuned to the requirements of positive political economy.

\textit{University of Chicago and American Bar Foundation}

References


© Royal Economic Society 2001
