NBER WORKING PAPER SERIES

ACCOUNTING FOR HETEROGENEITY, DIVERSITY AND GENERAL EQUILIBRIUM IN EVALUATING SOCIAL PROGRAMS

James J. Heckman

Working Paper 7230
http://www.nber.org/papers/w7230

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 1999

This paper was prepared for an AEI conference, "The Role of Inequality in Tax Policy," January 21-22, 1999 in Washington, D.C. I am grateful to Christopher Taber for help in conducting the tax simulations, and to Jeffrey Smith for help in analyzing 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 Lars Hansen, Kevin Hassett, Louis Kaplow, and Michael Rothschild. This research was supported by NSF-SBR-93/21/048, NSF 97-09-873, and a grant from the Russell Sage Foundation. All opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

Accounting For Heterogeneity, Diversity and
General Equilibrium In Evaluating Social Programs
James J. Heckman
NBER Working Paper No. 7230
July 1999
JEL No. C31

ABSTRACT

This paper considers the problem of policy evaluation in a modern society with heterogeneous agents and diverse groups with conflicting interests. Several different approaches to the policy evaluation problem are compared including the approach adopted in modern welfare economics, the classical representative agent approach adopted in macroeconononomics and the microeconomic treatment effect approach. A new approach to the policy evaluation problem is developed and applied that combines and extends the best features of these earlier approaches.Evidence on the importance of heterogeneity is presented. Using an empirically based dynamic general equilibrium model of skill formation with heterogeneous agents, the benefits of the more comprehensive approach to policy evaluation are examined in the context of examining the impact of tax reform on skill formation and the political economy aspects of such reform. A parallel analysis of tution policy is presented.

James J. Heckman
Dept. of Economics
University of Chicago
1126 E 59th Street
Chicago, IL 60637
and NBER
jjh@uchicago.edu

Introduction

Coercive redistribution and diversity in the interests of its constituent groups are essential features of the modern welfare state. Disagreement over perceived consequences of social policy creates the demand for publically justified "objective" evaluations. If there were no coercion, redistribution and intervention would be voluntary activities and there would be no need for public justification of voluntary trades. The demand for publically documented objective evaluations of social programs arises in large part from a demand for information by rival parties in the democratic welfare state.[1] Since different outcomes are of interest to rival parties, a variety of criteria should be used when considering the full consequences of proposed policies. This paper examines these criteria and considers the information required to implement them.

Given that heterogeneity and diversity are central to the modern state, it is surprising that the methods most commonly used for evaluating its policies do not recognize these features. 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 are its important features. Heterogeneity across persons in preferences and policy outcomes are treated as second order problems and estimates of

[1] Indeed, as discussed by Porter (1995), the very definition of "objective" standards is often the topic of intense political debate. See also the discussion in Young (1994).

policy effects are based on macro time series per capita aggregates.

Standard cost-benefit analysis ignores both distributional and general-equilibrium aspects of a policy and enumerates aggregate costs and benefits at fixed prices. Harberger's paraphrase of Gertrude Stein that "a dollar is a dollar is a dollar" succinctly summarizes the essential features of his 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, these analyses usually reflect the personal preferences of the individuals conducting particular evaluations.

Access to microdata facilitates the estimation of the distributional consequences of alternative policies. Yet surprisingly, the empirical micro literature focuses almost exclusively on estimating mean impacts for specific demographic groups and estimates heterogeneity in program impacts only across demographic groups. It neglects heterogeneity in responses within narrowly defined demographic categories - variation shown to be empirically important both in the literature and in the empirical analysis I present below.

Microdata are no panacea, however, and they must be used in conjunction with aggregate time-series data to estimate the full general-equilibrium consequences of policies. Even abstracting from general-equilibrium considerations, the estimates produced from social experiments and the microeconometric "treatment effect" literature are not those required to conduct a proper cost-benefit analysis, unless agents with identical *observed* characteristics respond identically to the

policy being evaluated; or if they do not, their participation in the program being evaluated must not depend on differences across agents in gains from the program. The estimates produced from social experiments and the treatment effect literature improve on aggregate time series methods by incorporating heterogeneity in responses to the policies in terms of observed characteristics but ignore heterogeneity in unobserved characteristics, an essential feature of the microdata from program evaluations.

Unlike the macro-general-equilibrium literature, the literature on modern welfare economics (see. e.g., Sen, 1973) recognizes the diversity of outcomes produced under alternative policies but adopts a rigid posture about how the alternatives should be evaluated, invoking some form of "Veil of Ignorance" assumption as *the* "ethically correct" point of view. Initial positions are treated as arbitrary and redistribution is assumed to be costless. The political feasibility of a criterion is treated as a subsidiary empirical detail that should not intrude upon an "ethically correct" or "moral" analysis. In this strand of the literature, it is not uncommon to have the work of "contemporary philosophers" invoked as a source of final authority (see, e.g. Roemer, 1996), although the philosophers cited never 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 policies on their personal outcomes. As noted by Jeremy Bentham (1824), appeal to authority is the lowest form of argument. Thus the appeal to philosophical authority by many economists on matters of "correct distributional

criteria" is both surprising and disappointing.

In this essay, I question this criterion. Its anonymity postulates do not describe actual social decision making in which individuals evaluate policies by asking whether they (or groups they are concerned about) are better off compared to a benchmark position.[2] Agents know, or forecast, their positions in the distributions of outcomes under alternative policies and base their evaluations of the policies on them. From an initial base policy state, persons can at least partially predict their positions in the outcome distributions of alternative policy states. I improve on modern 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 parties in the democratic process.

In order to make my discussion specific and useful, I consider the evaluation of human capital policies for schooling and job training. Human capital is the largest form of investment in a modern economy. Human capital involves choices at the extensive margin (schooling) and at the intensive margin (hours of job training). Differences in ability are documented to affect the outcomes of human capital decisions in important ways. The representative-agent macro-general-equilibrium paradigm is poorly suited to accommodate these features; the cost-benefit approach ignores the distributional consequences of alternative human capital policies; and the approach

[2] Recall Ronald Reagan's devastating rhetorical question in the 1980 campaign: "Are you better off today than you were four years ago?".

taken in modern welfare economics denies that it is interesting to determine how policies affect movements of individuals across the outcome distributions of alternative policy states.

Using both micro-and macrodata, I establish the empirical importance of heterogeneity in the outcomes of human capital policies even conditioning on detailed individual and group characteristics. Using data from a social experiment evaluating a prototypical job training program, I compare evaluations under the different criteria. Theoretically important distinctions turn out to be empirically important as well and produce different descriptions of the same policy.

I present an approach to policy evaluation that unites the macro-general-equilibrium approach with the approach taken in modern welfare economics. Using an empirically based general-equilibrium model that combines micro-and macrodata, I examine the distributional consequences of various tax and tuition policies. I present evidence on the misleading nature of the micro evidence produced from social experiments and the 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. I first present alternative criteria that have been proposed to evaluate social programs and consider their limitations. I propose a position-dependent criterion to evaluate policies. I then consider the information requirements of the various criteria. Not surprisingly, the more interesting criteria are also more demanding in their requirements. I consider the consequences of heterogeneity in responses to policies by agents for the success of various

social experiment with what is required to perform a cost-benefit analysis. There is a surprising disconnect between the two approaches when agents respond differently to the same program.

I go on to consider the evidence on heterogeneity in program impacts across persons, using data from a protypical job training program. I use a variety of criteria to evaluate the same program, 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 among these alternative evaluation measures.

I then present an empirically based dynamic overlapping-generations general-equilibrium model fit on both micro-and macrodata that extends the pioneering analysis of Auerbach and Kotlikoff (1987) on intergenerational accounting to include human capital formation and heterogeneity in human ability. These extensions produce a framework that accounts for rising wage inequality and that can be used to evaluate alternative tax and tuition policies, including their distributional impacts. The estimates produced from the general-equilibrium framework are contrasted with those obtained from the widely used social experiment and treatment effect approaches. The contrasts are found to be substantial, casting doubt on the value of conventional methods that are

used to evaluate human capital policies.

I. Alternative Criteria for Evaluating Social Programs

In this section, I consider alternative criteria that have been set forth in the literature to examine the desireability 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:

$$W(j) = W(U_1(Y_{j1}), ..., U_N(Y_{jN})). \tag{I-1}$$

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

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

Criteria (I-1) and (I-2) implictly 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 $U_i(Y_{j1}, ..., Y_{jN})$ for all i.

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

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

Analysts who adopt criterion (I-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.

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, politicians and advocacy groups are interested in knowing the proportion of people who benefit from policy j as compared to policy k:

(I-4) $$PB(j\,|j,k) = \frac{1}{N}\sum_{i=1}^{N} 1(U_i(Y_{ji})) \geq U_i(Y_{ki})),$$

where "1" is the indicator function: $1(A) = 1$ if A is true; $1(A) = 0$ otherwise. In the median voter model, a necessary condition for j to be preferred to k is that $PB(j\,|j,k) \geq 1/2$. Other persons concerned about "social justice" are 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 centers on the gains to specific types of persons, e.g., the gains to persons with outcomes in the base state k less than $\underline{y}$: $\Delta_{jki} = Y_{ji} - Y_{ki} | Y_{ki} \leq \underline{y}$, or their distribution

$$F\left(\Delta_{jk} | Y_k = y_k, y_k \leq \underline{y}\right), \tag{I-5}$$

or the utility equivalents of these variables. Within a targeted subpopulation, there is sometimes interest in knowing the proportion of people who gain relative to specified values of the base state k:

$$\Pr\left(\Delta_{jk} > 0 | Y_k \leq \underline{y}\right). \tag{I-6}$$

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

The existence of merit goods like education or health implies that specific components of the vector Y_{ji} are of interest to certain groups. Many policies are paternalistic in nature and implicitly assume that 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:

$$Y_{ji} \geq \underline{y} \qquad \text{for } i \in S,$$

where S is a target set toward which the policy is directed, or in terms of a relative requirement compared to a base state k:

$$Y_{ji} \geq Y_{ki} \qquad \text{for } i \in S.$$

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, ex ante, these consequences are not known and, 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, ex post, agents know their outcome in a benchmark state, they may not know it ex ante, and they may always be uncertain about what they would have experienced in an alternative state.

In the literature on welfare economics and social choice, one form of decision-making under uncertainty plays a central role. The "Veil of Ignorance" of Vickrey (1945, 1961) and Harsanyi (1955, 1975) postulates that decision makers are completely uncertain about their positions in the distribution of outcomes under each policy, or *should* act as if they are completely uncertain, and they should use expected utility criteria (Vickrey-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 practical work in evaluating different income distributions (see Sen, 1973). It is an empirically tractable approach because it only requires

information about the marginal distributions of outcomes produced under different policies. The empirical literature on evaluating income inequality uses this criterion to compare the consequences of growing wage inequality in the past two decades: (See, e.g. 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 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 social decision making in a democratic welfare state recognizes that persons act in their own self-interest, or in the interest of certain other groups (e.g. the poor, the less able) and have at least partial knowledge about how they (or the groups they are interested in) will fare under different policies, and act on those perceptions, but only imperfectly anticipate their outcomes under different policy regimes. Even if outcomes in alternative policy regimes are completely unknown (and hence represent a random draw from the outcome distribution), the outcomes under the current policy are known. The outcomes in different regimes may be dependent so that persons who benefit under one policy may also benefit under another. For a variety of actual social choice mechanisms, both the initial and final positions of each agent are relevant for evaluation of social policy.[3] Politicians, policy makers and participants in the welfare state are more likely to be interested in how specific policies affect the fortunes

[3]This theme is developed in Heckman, Smith and Clements (1997), Heckman and Smith (1998), Coate (1998) and Besley and Coate (1998).

of specific groups measured from a benchmark state than in some abstract measure of "social justice".[4]

However, 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 $F(y_j, y_k \,|I_i)$ 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

$$E(U_i(Y_j)\,|I_i) > E(U_i(Y_k)\,|I_i).$$

Letting θ_i parameterize heterogeneity in preferences, so $U_i(Y_j) = U(Y_j;\theta_i)$, and using integrals to simplify the expressions, the proportion of people who prefer j is

$$PB\,(j|j,k) = \int 1(E\,(U\,(Y_j;\theta)\,|I) > E\,(U\,(Y_k;\theta)\,|I))dF\,(\theta, I), \tag{I-7}$$

where $F(\theta, I)$ is the joint distribution of θ and I in the population whose preferences over outcomes are being studied.[5] The voting criterion previously discussed is the special case where $I_i = (Y_{ji}, Y_{ki})$, so there is no uncertainty about Y_j and Y_k, and

$$PB\,(j|j,k) = \int 1\,(U\,(y_j;\theta) > U\,(y_k;\theta))\,dF\,(\theta, y_j, y_k)\,. \tag{I-8}$$

[4] 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) realizations.

[5] 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.

Expression (I-8) is an integral version of (I-4) when outcomes are perfectly predictable and when preference heterogeneity can be indexed by vector θ.

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 I_i by t, (I_{it}), to recognize that information about the outcomes of policies may accrue over time. Ex ante and ex post assessments of a voluntary program need not agree. Ex post assessments of a program through surveys administered to persons who have completed it (see Katz, Gutek, Kahn and Barton, 1975) may disagree with ex ante assessments of the program. Both may reflect honest valuations of the program but they are reported when agents have different information about it or have their preferences altered by participating in the program. Before participating in a program, persons may be uncertain of the consequences of participation in it. A person who has completed program j may know Y_j but can only guess at the alternative outcome Y_k which they have not experienced. In this case, ex post "satisfaction" for agent i is synonymous with the following inequality:

$$U_i(Y_{ji}) > E(U_i(Y_{ki}) \,|I_{it}), \tag{I-9}$$

where t is the post-program period in which the evaluation is made. In addition, survey ques-

tionnaries about "client" satisfaction with a program may capture subjective elements of program experience not captured by "objective" measures of outcomes that usually exclude psychic costs and benefits.

II. The Data Needed to Evaluate the Welfare State

To implement criteria (I-1) and (I-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 (I-3) only requires GNP (the sum of the program j net output). If interest centers solely on the distributions of outcomes of direct program participants, the measures can be defined solely for populations with $D_j = 1$. Criteria (I-4), (I-5), (I-6) and (I-8) require knowledge of outcomes and preferences *across* policy states. Criterion (I-7) requires knowledge of the joint distribution of information and preferences across persons. Tables 1A and 1B summarize the criteria and the data needed to implement them. The cost-benefit criterion is the least demanding; the voting criterion is the most demanding in that it requires information about the *joint* distributions of outcomes across alternative policy states.

Three distinct types of information are required to implement these criteria: (a) private preferences, including preferences toward the consumption and well being of others; (b) social preferences, as exemplified by social welfare function (I-1) and (c) distributions of outcomes in alternative states, and for some criteria, such as the voting criterion, *joint* distributions of outcomes *across* policy states. The reasons for the popularity of cost-benefit analysis are evident from these tables.

An important practical problem rarely raised in the literature on "social justice" is that many proposed criteria are not operational with current levels of knowledge.

There is a vast literature on the estimation of individual preferences defined 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 program reveal their preferences for it. Much of the standard literature on estimating consumer preferences abstracts from heterogeneity. However, a growing body of evidence summarized in Browning, Hansen and Heckman (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.[6] There are surprisingly few estimates of social welfare function (I-1) (Maital, 1973; Saez, 1998; and Gabaix, 1998 are exceptions), despite the widespread use of the social welfare function in public economics. The paucity of estimates of it suggests that the social welfare function is an empirically empty concept. It is a misleading, but traditional, intellectual crutch without operational content.[7]

Responses to income shocks, wages and the like vary widely across consumers. The evidence

[6] See, *e.g.*, Heckman, 1974a.

[7] Saez and Gabaix assume that tax schedules are set optimally using a social welfare function and derive the local curvature of the social welfare function that generates policy outcomes. They do not test that proposition. Ahmed and Stern (1984) test the proposition that taxes and subsidies in India are generated by optimizing a social welfare function.

speaks strongly against the representative agent model or the various simplifications used to justify RBC models. 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 I 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.

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

$$(Y_{ji}, Y_{ki}) \qquad i = 1, ..., I. \tag{II-1}$$

Macroeconomic approaches focus exclusively on mean outcomes or some other low dimensional representation of the aggregate (e.g. 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. The first case requires that we "adjust" the data on j and k to account for changes in the conditioning 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 could build up the joint distribution of policy outcomes by observing the same people under different regimes.

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.[8]

An entire literature on structural estimation in econometrics has emerged in an attempt 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.[9] If outcomes were indeed identical across persons, or if the representative consumer were a "reasonably good" representation, from knowledge of aggregate means, one could answer all of the policy evaluation questions in Tables 1A and 1B provided that preferences were known. This is a consequence of the implicit assumption of the representative consumer model that the joint distribution of (II-1) is degenerate.

The common form of the microeconomic evaluation problem is apparently more tractable. It considers evaluation of a program in which participation is voluntary although it may not have been intended to be so. Accordingly, it is not well suited to evaluating programs with universal

[8] A quotation from Knight is apt "The existence of a problem in knowledge depends on the future being different from the past, while the possibility of a solution of the problem depends on the future being like the past". (Knight, 1921, p. 313.)

[9] As summarized in Browning, Hansen and Heckman (1999), there is an emerging literature in macroeconomics that recognizes the evidence of microheterogeneity and its consequences for model construction and policy evaluation.

coverage such as a social security program.

Persons are offered a service through a program and may select into the program to receive it. A distinction is made between direct participation in the program and indirect participation. The latter occurs when people pay taxes or suffer the market consequences of changed supplies as a consequence of the program. Eligibility for the program may be restricted to subsets of persons in the larger society. Many "mandatory" programs allow that persons may attrite from them or fail to comply with program requirements. Participation in the program is thus equated with direct receipt of the service, and payments of taxes and general-equilibrium effects of the program are typically ignored.[10]

In this formulation of the evaluation problem, the no-treatment outcome distribution for a given program is used to approximate the distribution of outcomes in the no-program state. That is, the outcomes of the "untreated" within the framework of an existing program are used to approximate outcome distributions when there is no program. This approximation rests on two distinct 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

[10]The contrast between micro and macro analysis is overdrawn. Baumol and Quandt (1966), Lancaster (1971) and Domencich and McFadden (1975) are micro examples of attempts to solve what we have called a macro problem. Those authors consider the problem of forecasting the demand for a new good which has never previously been purchased.

distributions can be ignored or surmounted.[11] The treatment effect approach also converts the evaluation problem into a comparison between an existing program j and a benchmark no-program state rather than into a comparison between any two hypothetical states j and k.[12]

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

Let subscript "0" denote a policy regime without the program. Let $D_{ji} = 1$ if person i participates in program j. A crucial identifying assumption that is implicitly invoked in the microeconomic evaluation literature is

$$Y_{ji}^0 = Y_{0i} \tag{A-1}$$

i.e. that the no program outcome for i is the same as the no treatment outcome.

Letting $F(a \mid b)$ denote the conditional distribution of a given b, the assumption implies that $F(y_j^0 \mid D_j = 0, X) = F(y_0 \mid D_j = 0, X)$ for $y_j^0 = y_0$ 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

[11] As we note below, evidence from self-selection decisions can be used to evaluate private preferences for the program so that in principle we can use the "problem" of self selection as a source of information about private valuations. See, *e.g.* Heckman, (1974a,b), and Heckman and Honoré (1990) 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 program state.

policy j is operative. This assumption is consistent with a program 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 program participation decisions, it is possible to infer the implicit valuations of the program made by persons eligible for it. These evaluations constitute all of the data needed for a libertarian program evaluation, but more than these are required to evaluate programs in the interventionist welfare state. 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 (I-4) or (I-7), as I demonstrate below. I now consider how access to microdata and social experiments enables one to answer the evaluation questions posed in Section I.

III. What Can Be Learned From Micro Data and Social Experiments?

This section considers the information produced from social experiments and from ordinary observational data. Even abstracting from the problem that the analysis of these data typically ignores general-equilibrium effects, the information produced by them 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 programs.

Consider the analysis of program j and assume that assumption (A-1) is invoked. Within the framework of the "treatment effect" literature, we observe one of the following pair

$$(Y_i^0, Y_i^1)$$

for person i. To simplify the notation, I drop the j subscript in this section. At a point in time, we cannot observe a person simultaneously in the treated and untreated state. In general, we cannot form the gain of moving from "0" to "1" and $\Delta_i = Y_i^1 - Y_i^0$ for anyone. The evaluation problem is reformulated to the population level. The goal becomes to estimate some features of the distribution of Δ. 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. We observe Y_i

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

for each person.

The potential outcomes for person i can be written as

(III-1) $$Y_i^0 = \mu_0 + \varepsilon_{0i}$$

(III-2) $$Y_i^1 = \mu_1 + \varepsilon_{1i}$$

where $E(\varepsilon_0) = E(\varepsilon_1) = 0$. The means can be written in terms of observed characteristics X $(\mu_0(X); \mu_1(X))$ but for simplicity of notation we suppress this dependence. Thus we can may write

(III-3) $$Y_i = \mu_0 + (\mu_1 - \mu_0 + \varepsilon_{1i} - \varepsilon_{0i}) D_i + \varepsilon_{0i}.$$

Most of the evaluation literature formulates the parameters of interest as means. Two means receive the most attention. The first is

$$E(Y^1 - Y^0)$$

the average treatment effect ("ATE") that records the average gain of moving a randomly selected person from "0" to "1". A second mean is

$$E(Y^1 - Y^0 \mid D = 1)$$

the effect of treatment on the treated (TT). The two means are the same under one of the following conditions:

(C-1): $\varepsilon_{1i} = \varepsilon_{0i}$ so $\Delta_i = \Delta$

(No response heterogeneity given X)

or

(C-2): $E(\varepsilon_{1i} - \varepsilon_{0i} \mid D_i = 1) = 0$

(Agents do not enter the program 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 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. Write the traditional textbook model as:

(A) $\quad Y_i = \alpha_0 + \alpha_1 D_i + U_i, \qquad E(U_i) = 0.$

In this framework α_1 is a common coefficient for each i. It embodies assumption (C-1) where $\varepsilon_{1i} = \varepsilon_{0i}$ 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 $E(U_i \mid D_i = 1) \neq 0$.

In contrast, consider a second model:

(B) $Y_i = \alpha_0 + \alpha_{1i} D_i + U_i, \qquad E(U_i) = 0$ where $E(\alpha_{1i}) = \mu_1 - \mu_0$ but $V_i = \alpha_{1i} - E(\alpha_{1i}) = \varepsilon_{1i} - \varepsilon_{0i}$ satisfies $E(V_i \mid D_i = 1) = 0$ or equivalently $E(\varepsilon_{1i} - \varepsilon_{0i} \mid D_i = 1) = 0$.

In this framework, responses are different across persons (α_1 has an i subscript) but conditional on X, persons do not participate in the program based on these differential responses.[13] Again selection bias is said to arise if $E(U_i \mid D_i = 1) \neq 0$.

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

(C) $Y_i = \alpha_0 + \alpha_{1i} D_i + U_i, \qquad E(U_i) = 0$

[13] Another way to say this is that $\Pr(D_i = 1 \mid Z_i, V_i) = \Pr(D_i = 1 \mid Z_i)$. This is a "noncausality" condition.

$$E(U_i \mid D_i = 1) \neq 0 \neq E(\varepsilon_{1i} - \varepsilon_{0i} \mid D_i = 1).$$

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

Under models A and B, the parameters $E(Y_1 - Y_0)$ 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) and Heckman (1992), 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, we can go from a regression estimate of equation (A) to answer all of the policy questions posed in Section I comparing the policy being evaluated with a benchmark no policy state. The distribution of gains, Δ, *across and within* policy regimes is degenerate. Everyone either benefits or loses from the policy. 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 but assumes that persons do not act on it. Now the representative agent paradigm should be adjusted to account for variation in individual responses to the program; 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.[14] If outcomes are heterogeneous in the sense of model (B), conventional instrumental

[14] See *e.g.* Heckman, Smith and Clements, 1997.

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 study by Heckman, Clements and Smith (1997), there are few studies that estimate the distributions of program impacts.

Model (C) captures a fundamental form of heterogeneity. Agents know more than the observing economist and they act on this information in deciding whether or not to participate in a program. $E(Y_1 - Y_0) \neq E(Y_1 - Y_0 \mid D = 1)$. 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 (1999). In this case standard instrumental variable methods break down (see Heckman, 1997 or Heckman and Vytlacil, 1998). Heckman, Smith and Clements (1997) and Heckman and Smith (1998) present estimates of outcome distributions under Model (C). Heckman, Ichimura, Smith and Todd (1998) present evidence that Model (C) describes the data for the prototypical training program discussed in Section V below. While most of the thinking about program evaluation is in terms of Model (A) or more recently, in terms of Model (B), considerable evidence supports Model (C) for many programs.

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 α_1 is enough to answer all of the policy evaluation questions posed in Section I. The joint distribution of (II-1) is degenerate when k is the benchmark no-program state. Even if randomization alters

the composition of program participants (*i.e.* there is "randomization bias"), for any observed X in the experiment we can obtain α_1.

If Model (C) characterizes the data, all we can recover from social experiments administered to people who apply and are accepted into the program (the common point in the enrollment process where randomization is administered) are

$$F(y^1 \mid D = 1) \text{ and } F(y^0 \mid D = 1).$$

We cannot recover the joint distribution $F(y^1, y^0 \mid D = 1)$ either for persons who seek to participate in the program or for the general population. Below, we discuss what can be learned in this case. First, however, we consider what can be learned from participation decisions under Model (C).

Information From Revealed Preference

If agents act on the idiosyncratic gain from the program, so model (C) is the appropriate one, it is possible to use this information to infer the implicit valuations they place on the gains from the program 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 program or attrition from it.[15]

[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 labor supply.

The prototypical framework is the Roy (1951) model. In that setup,

$$D = 1(Y^1 \geq Y^0), \tag{III-4}$$

so participation depends only on the net gain from the program: $\Delta = Y^1 - Y^0$. Using equations (III-1) and (III-2), we obtain $Pr(D = 1|X) = \Pr(Y^1 - Y^0 \geq 0|X) = \Pr(\varepsilon_1 - \varepsilon_0 \geq -(\mu_1(X) - \mu_0(X)))$. Under conditions specified in Heckman and Honoré (1990), the joint distribution $F(y^0, y^1, D)$ can be identified nonparametrically.[16] Thus we can form all of the evaluation parameters presented in Tables 1A and 1B if the Roy model describes the data.

The crucial feature of the Roy model is that the decision to participate in the program is made solely in terms of potential outcomes. No new unobservable variables enter the model that do not appear in the outcome equations.[17] In this case, information about who participates also informs us about the distribution of the value of the program to participants $F(y^1 - y^0|Y^1 > Y^0, X)$. Thus, we acquire the distribution of implicit values of the program for participants, which is all that is required in a libertarian evaluation of the program. However, in the general case evaluation of the welfare state requires information about "objective" outcomes and their distributions that

[16]Heckman and Honoré (1990) demonstrate that if X is independent of $(\varepsilon_1, \varepsilon_0)$, $Var(\varepsilon_1) < \infty$ and $Var(\varepsilon_0) < \infty$. 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 $\mu_1(X)$ and $\mu_0(X)$ overlap or contain the supports of ε_1 and ε_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 normalizations. Precise conditions are given in Theorem A-1 in Appendix A of Heckman and Smith, 1998.

[17]We could augment decision rule (III-4) to be $D = 1(Y^1 - Y^0 - k(Z) \geq 0)$. Provided that we measure Z and condition on it, and provided that $(U_1 - U_0) \perp\!\!\!\perp (X, Z)$, the model remains nonparametrically 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 $(Y^1 - Y^0)$.

are needed to make the interpersonal comparisons that are an essential feature of the welfare state. Only in the Roy model do the "objective" and "subjective" evaluations coincide.[18]

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.

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 program. As noted by Heckman (1992) and Moffitt (1992), social experiments as typically conducted on persons who apply and are initially accepted into a program do not provide information about the determinants of program participation. Nonexperimental data can be used to infer the preferences of agents who select into the program.

Appendix A presents a discussion of the relationship between the parameters of cost-benefit analysis and the Roy model. Many of the parameters estimated in the micro evaluation literature are not the ones needed to conduct a rigorous cost-benefit analysis. For this reason, this literature

[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 $D = 1(IN > 0)$ 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 (1990a) demonstrates that in this more general case, provided that some structure is placed on η, we can nonparametrically identify $F(y^0, D \mid X)$ and $F(y^1, D \mid X)$ but not the full joint distribution $F(y^0, y^1, D \mid X)$. A generalization of his proof is given in Theorem A-2 of Appendix A of Heckman and Smith, 1998.

is not as informative about the economic aspects of program evaluation as one might hope.

The Problem of Recovering Joint Distributions

In the general case where textbook model (A) does not apply, and responses to programs are heterogeneous, we encounter a difficult evaluation problem. Unless the Roy model is invoked, we cannot identify the joint distribution of (Y^0, Y^1). At best we can extract the marginal distributions of Y^0 and Y^1, even from ideal social experiments. This leaves considerable uncertainty about our ability to implement the voting criterion and many other position-dependent majority voting criteria discussed in Section I.

To see this problem, suppose that we have data from an ideal social experiment so that standard self-selection problems can be ignored. Suppose that there are N treated and N untreated 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, we obtain two data distributions the gain:

$$\underline{\text{Treatment Outcome}}: F(y^1|D=1) \qquad \underline{\text{Non-Treatment Outcome}}: F(y^0|D=1)$$

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

We know the marginal data distributions $F(y^1|D=1)$ and $F(y^0|D=1)$, but we do not know where person i in the treatment distribution would appear in the non-treatment distribution.[19]

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:

$$\underset{\sim}{\Delta}=\underset{\sim}{Y}^1 - \Pi_\ell \underset{\sim}{Y}^0 \qquad \ell = 1, \ldots, N!$$

where Π_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 $\underset{\sim}{\Delta}$, $\underset{\sim}{Y}^1$ and $\underset{\sim}{Y}^0$ are $N \times 1$ vectors of impacts, treated and untreated outcomes. By considering all possible permutations, one can obtain all possible sortings of treatment, Y^1, and non-treatment, Y^0, outcomes using realized 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 the common effect case, Y^1 and Y^0 differ by a constant for each person. A generalization of that model preserves perfect dependence in the ranks between the

[19] These distributions can also be defined conditional on X.

two distributions but does not require the impact to be the same at all quantiles of the base state distribution.

In place of ranks, it is easier to work with the percentiles of the Y^1 and Y^0 distributions. which have much better statistical properties.[20] 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 $\Delta(y^0) = F_1^{-1}(F_0(y^0|D=1)) - y^0$. One can test non-parametrically for the classical common effect model by determining if percentiles 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 $T : q_1 \rightarrow q_0$. 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 Π_ℓ is a Markov transition matrix and we consider all possible Markov matrices.

I now consider empirical evidence on the question of the constancy of the gross gain Δ across base state quantiles using earnings data from an experimental evaluation of a major U.S. job training program described in Appendix B. Figure 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"

[20] See Heckman and Smith, 1993, Heckman, Smith and Clements, 1997.

distribution. More formally, Figure 1 assumes that the permutation matrix $\Pi = I$. No conditioning is made, so the full sample is utilized. Between the 25th and 85th percentiles the assumption of a constant impact is roughly correct. This evidence supports Model (A). However, the data are grossly at odds with this model at the highest and lowest percentiles.[21] Heckman, Smith and Clements (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.

IV. Evidence on Impact Heterogeneity and the Value of Self-Assessments and Revealed Preference Information

This section of the paper address three questions. Question (1) is: "What is the empirical evidence on heterogeneity in program 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 program described in detail in Appendix B, I implement the criteria discussed in Section I to bound or identify the joint distribution of outcomes conditional on $D = 1$. There is considerable evidence of heterogeneity of program impacts, so that conventional econometric methods do not take one very far in constructing the evaluation criteria discussed

[21] Standard errors for the quantiles are obtained using methods described in Csörgo (1993).

in Section I. Use of experimental data enables us to avoid the self-selection problems that plague ordinary observational data, and simplifies our analysis.

Given the evidence on impact heterogeneity, I ask question (2): "How sensitive are the estimates of the proportion of people who gain from the program - what I have called 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 program exceeds 50 percent in every case I consider but one, and is close to 100 percent 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 program. These assessments are all that are required for a libertarian evaluation of the welfare state. I examine the implicit value placed on the program by addressing the following questions: (3a) "Are persons who applied to the program and were accepted into it but then randomized out of it placed in an inferior position relative to those accepted applicants who were not randomized 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 program. I also consider ex post evaluations of participants by asking: (3b) "How 'satisfied' are participants with their experience in the program?" Self-assessments of programs are widely used in evaluation research (see e.g., 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 program (its process) or an evaluation of the benefits of the program? The evidence presented here suggests that respondents report a net benefit inclusive of their costs of participating in the program. Groups for whom the program has a negative average impact as estimated by the "objective" experimental data express as much (or more) enthusiasm for the program as groups with positive average impacts. A third source of revealed preference evaluations uses the revealed choices of attriters from the program. Econometric models of self-selection since Heckman (1974a,b) have used revealed choice behavior to infer the evaluations people place on programs 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 program do attriters place on it?" I do not examine this question in this paper. Heckman and Smith (1998) present evidence on it.

Evidence on Heterogeneity

Heckman and Smith (1993, 1998) apply the nonparametric Frechet Bounds of classical probability theory to the JTPA data to establish that the variance of the gain Δ is positive for a variety of conditioning sets. Their estimate for the JTPA data is reported in the first row of Table 6. The $675 lower bound on the standard deviation is to be compared with a $400 gain and mean $7200

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.[22]

Using the sample data from the JTPA experiment (see Orr, et al., 1995) and discussed in Appendix A, and in Heckman and Smith (1998), we may pair percentiles of the Y^1 and Y^0 distributions for any choice of rank correlation τ between -1.0 and 1.0. The case of $\tau = 1.0$ corresponds to the case of perfect positive dependence, where $\Pi = \mathrm{I}$ and $q_1 = q_0$. The case where τ = -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, Smith and Clements (1997) show how to obtain random samples of permutations conditional on values of τ between 1.0 and -1.0. Table 2 displays two sets of estimates from their work. The first set assumes positive but not perfect dependence between the percentiles of Y^1 and Y^0, with $\tau = 0.95$. Estimates based on a random sample of 50 percentile permutations with this value of τ 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

[22]The classical solution to bounding a joint distribution from its marginals uses the Frechet-Hoeffding bounds:

$$Max[F(y^1(D=1) + F(y^0 \mid D=1), -1.0] \leq F(y^1, y^0 \mid D=1) \leq Min(F(y^1 \mid D=1), F(y^0 \mid D=1)).$$

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.

striking still are the results in the third column of Table 2, which correspond to the case where $\tau = 0.0$. This value of τ is implied by independence between the percentiles of Y^1 and Y^0. Here (as in the case with τ = -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 program is robust to the choice of τ.[23]

Even though many joint distributions of outcomes are consistent with the marginals produced from a social experiment, one model is not: common effect model (A). Heckman, Smith and Clements 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 demographic categories.

Assuming the Gain Is Independent of the Base

Suppose that random coefficient model (B) of Section III is true. In that framework, suppose that Δ is not known at the time decisions to go into the program are made. Then if Y^0 is known, Y^0 is independent of Δ. Otherwise the coefficient α_{1i} is correlated with D_i. In applying this to experimental data, let $R = 1$ if a person who applies and is provisionally accepted into the program

[23]Heckman, Smith and Clements (1997) present methods for allowing for mass points of zero earnings in the population, and some evidence derived from such methods. Their qualitative conclusions on variability are similar to ours.

is randomized into the program, and $R = 0$ if a provisionally accepted applicant is randomized out. $Y = Y^0 + R\Delta$, and $R\Delta$ is statistically independent of Y^0.

In the notation of Section III, we obtain a conventional random coefficient model for a regression: $Y = RY^1 + (1 - R)Y^0 = \alpha_0 + \alpha_{1i}R_i + \varepsilon_0$. Using a components of variance model, one may write $E(\Delta) = \bar{\alpha}$ and $V_i = \Delta_i - \bar{\Delta} = \alpha_{1i} - \bar{\alpha}$ so that

$$Y_i = \alpha_0 + \bar{\alpha}_1 R_i + V_i R_i + \varepsilon_0$$

where

$$E(V_i R_i + \varepsilon_0) = 0.$$

Using a standard random coefficient model, we can estimate the variance of V_i. The first row of Table 3 presents estimates of the random coefficient using these assumptions. The evidence supports the hypothesis that $\text{VAR}(\Delta) > 0$, suggesting that a more elaborate approach to estimating the distribution of Δ based on deconvolution is likely to be fruitful. If we maintain normality of Y^1 and Y^0 (given $D = 1$ and X), the distribution of Δ is normal with mean $\overline{\Delta}$ and variance $\text{VAR}(\Delta)$ and deconvolution is easy to perform. Under this assumption, we can estimate the voting criterion and determine the estimated proportion of people who benefit from the program.

More generally, it is not necessary to assume that the distribution of Δ is normal. I use the deconvolution procedure presented in Heckman, Smith and Clements (1997), to estimate the distribution of impacts nonparametrically. Table 3 presents parameters calculated from this

distribution. The evidence suggests that under this assumption, about 43% of adult women were harmed by participating in the program. The estimated density is presented in Figure 2 and is clearly non-normal. Nonetheless, the estimated variance of the nonparametric gain distribution matches the variance for the gain distribution obtained from the random coefficient model within the range of the sampling error of the two estimates. The estimates of the proportion who benefit are in close agreement across the two models when normality is imposed on the random coefficient model. The fact that a positive density is estimated indicates that the assumption underlying model (B) of Section III is consistent with the data for women and provides some support for the hypothesis that agents do not select into the program based on Δ.[24]

However, the evidence against matching as a method of evaluating social programs that is presented in Heckman, Ichimura, Smith and Todd (1998) and in Heckman, Ichimura and Todd (1997), suggests that in most demographic groups, persons act on unobservable gains in making program enrollment decisions. Matching assumes a (nonparametric) version of model B. Since the cited papers test, and reject, the matching assumption using the same JTPA data as used in this paper, a model of purposive selection on unobserved gains (Model C) is a more appropriate description of the JTPA data.

Testing For Ex Ante Stochastic Rationality of Participants

If individuals choose whether or not to participate in the program based on the gross gains

[24]These calculations were first presented in Heckman and Smith (1993).

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, then second-order stochastic dominance should imply that they order the distributions of outcomes for persons who sought to go into the program. For non-negative y^1, y^0 this form of rationality implies that

$$\text{(IV-1)} \qquad \int_0^{\alpha} F_1(y^1|D=1)dy^1 < \int_0^{\alpha} F_0(y^0|D=1)dy^0 \text{ for all } \alpha \in R_+.$$

Draws from the Y^1 distribution produce higher expected utility than draws from the Y^0 distribution among participants. The difference between the two integrals is a measure of regret among persons randomized out from the program 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 program on a principle other than expected utility formulated in terms of gross outcomes. Condition (IV-1) is a sufficient condition. Agents might still prefer distribution 1 even if it is not satisfied. Thus failure to reject is informative; rejection is not.

I test condition (IV-1) by comparing the integrals of the empirical CDFs of the control and treatment group earnings distributions for various values of α. Table 4 displays the results of tests of the null hypothesis of equality of the integrated distributions in (IV-1) for adult males and females and male and female youth using self-reported earnings in the eighteen months after random assignment. The table displays test results for $\alpha \epsilon \{\$2,500, \$5,000, \$10,000,$

$\$15,000, \$20,000, \$25,000\}$. Standard errors are obtained by bootstrapping. For adult males, the integrated CDF of earnings for the control group exceeds that for the treatment group at every point, with a p-value below 0.05 for $\alpha < \$16,500$, and below 0.10 for $\alpha < \$22,500$, which includes most of the supports of the two earnings distributions. The data for adult females provide strong evidence of rational behavior in the sense of (IV-1), passing the test at the five percent level or better for every value of α. This evidence suggests that personal objectives and program objectives are aligned for adult women. Results for youth are mixed. For male youth, for whom the mean experimental impact is significantly negative, the difference in integrated CDF's is negative for most values of α, though not statistically significant. For female youth, the difference switches sign around $\alpha = \$11,000$, but is never close to statistical significance.[25]

Evidence from Self-Assessments of Program Participants

Self-assessments of program participants are an alternative to comparisons of observed outcomes as a measure of program impact. Unlike the ex ante measures based on second-order stochastic dominance, these measures are statements about ex post expectations. There is no reason why the two measures should agree if people revise their assessments based on what they learn about a program by participating in it. In this section, I consider the strengths and limitations of self-reported assessments of satisfaction with the program as an evaluation criterion,

[25] These tests of second order stochastic dominance of treatment distributions were first presented in Heckman and Smith (1993, 1998) and Heckman, Smith and Clements (1997).

and report on self-evaluations by participants in the JTPA experimental treatment group. I also consider what can be learned from self-assessment data regarding the heterogeneity of individual treatment effects and the rationality of program participants.

Using participant assessments to evaluate a program has two main advantages relative to the approaches already discussed. First, participants have information not available to external program evaluators. They typically know more about certain components of the cost of program 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 program effectiveness, unless they are paid by the program through subsidies. Participants are likely to include such information in arriving at their self-assessments of the program. 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 program'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 program. They make an input-based measure of program evaluation and not an output-based 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 chose to go into the program may rationalize 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.

The top panel of Table 5 reports JTPA participant responses to a question about whether or not the program made them better off.[26] Assuming people answer honestly, and are reporting a gross impact, the self-assessment data clearly contradict the hypothesis of impact homogeneity. For all four demographic groups, 65 to 70 percent of self-reported participants give a positive self-assessment, not the 100% or 0% predicted if impacts were homogeneous. However, if respondents are reporting a perceived net impact, the evidence reported in Table 5 does *not* necessarily contradict an assumption of gross impact homogeneity if there is heterogeneity in costs across participants. The entries in the third row of Table 5 reveal that the fractions reporting a positive impact are far lower than those obtained from all of the analyses using outcome data. This evidence is consistent with one of two hypotheses: (a) that respondents are reporting net outcomes and that costs borne by participants are a substantial fraction of gross outcomes or (b) that self-assessments are inaccurate.

[26]The exact wording of the survey question is "Do you think that the training or other assistance you got from the program 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.

The evidence suggests that the self-assessments are at least partly based on inputs received rather than on outputs produced by the program. The lower panel of Table 5 shows the fraction of persons receiving each type of training whose self-assessment of the program was positive. The fraction increases with the level of treatment intensity for all four demographic groups. Expensive and more intensive services such as classroom training in occupational skills (CT-OS) and on-the-job training at a private firm (OJT) elicit a higher proportion of positive self-assessments than do less expensive services such as job search assistance (JSA) or basic education. However, the experimental impact estimates presented in Bloom, et al. (1993) reveal that treatment effectiveness and treatment intensity are not positively related. For example, for female youth, classroom training in occupational skills has a more negative mean impact than the less expensive services in the "other" treatment stream. This evidence suggests that participants may have difficulty correctly constructing what would have happened to them in the absence of treatment, and so rely in part on treatment intensity or program inputs as a proxy for treatment impact.

Finally, for adult women we 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 Figure 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 program for about 85 percent of participants - all except those with zero earnings in both states. This value far ex-

ceeds the overall self-reported effectiveness rate of 44 percent reported in row 3 of Table 5. The 44 percent 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 program 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 program, and the attriters have lower earnings than the non-attriters.

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

Summary of the Evidence on Impact Heterogeneity and Its Consequences

Table 6 presents a summary of the main findings of this section. (1) Under a variety of different assumptions, there is evidence of substantial heterogeneity in net impacts, Δ. (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. (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). However, positive dependence is implied by economic models of self selection. These narrow the range of dependence across outcome states. (4) The range of the estimated proportion of people benefitting from the program in the sense of gross outcomes (the "voting criterion") varies widely under different assumptions about the dependence in outcomes. The data from the self-report and attrition studies show a lower proportion benefitting - a phenomenon consistent with the hypothesis that net returns and not gross returns are being reported by participants.

V. 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 programs being evaluated. Many human capital policies are large scale in 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 micro economic evaluations.

In addition, many policies do not fall into the "treatment effect" category at all. A tax on labor 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 analyze distributional issues. An important exception is the class of overlapping gener-

ations models which explicitly consider inequality among generations. The goal of this section of the paper is to extend the important empirical overlapping generations model of Auerbach and Kotlikoff 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 I for a dynamic economy. It is a vehicle for examining the performance of micro evaluation methods within a general equilibrium setting.

Heckman, Lochner and Taber (HLT, 1998a) formulate and estimate dynamic general equilibrium models with endogenous heterogeneous human capital accumulation. Their model explains rising wage inequality in the U.S. economy. In this section of the paper, I use their model to study the impacts on skill formation of proposals to switch from progressive taxes to flat income and consumption taxes.

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

HLT build on the model of Alan Auerbach and Laurence 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. HLT discard the Auerbach-Kotlikoff efficiency units assumption for labor services. Models with efficiency units for labor services do not explain rising wage inequality among skill groups. (3) HLT use micro data joined with macro time series evidence 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 HLT 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 θ 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 HLT model extends the analysis of James Davies and John Whalley (1991) who introduce human capital into the Auerbach-Kotlikoff model but assume only one skill. HLT allow for multiple skills, incorporate both schooling and on-the-job training, and allow for rational expectations in calculating transition paths.

In the HLT 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 the highest level of lifetime utility. Define K_{at}^S as the stock of physical capital held at time t by a person age a: H_{at}^S 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_{at}) and post-school investment time (I^S_{at}) for each schooling level. Second, select among schooling levels to maximize lifetime welfare.

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

$$\text{(V-1)} \qquad V_{at}(H^S_{at}, K^S_{at}, S) = \max_{C_{at}, I^S_{at}} \frac{(C^S_{at})^\gamma - 1}{\gamma} + \delta V^S_{a+1,t+1}(H^S_{a+1,t+1}, K^S_{a+1,t+1}, S),$$

where δ is a time preference discount factor. HLT follow Laurence Kotlikoff, Kent Smetters, and Jan Walliser (1997, henceforth KSW), by assuming that the tax schedule can be approximated by a progressive tax on labor income and a flat tax on capital income. This gives a dynamic budget constraint,

$$\text{(V-2)} \qquad K^S_{a+1,t+1} \leq K^S_{a,t}(1 + (1 - \tau_k)r_t) + R^S_t H^S_{at}(1 - I^S_{at}) - \tau_l \left(R^S_t H^S_{at}(1 - I^S_{at})\right) - C^S_{at},$$

where τ_k is the proportional tax rate on capital, τ_l is the progressive tax schedule on labor 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. HLT abstract from labor supply. Estimates of intertemporal substitution in labor supply estimated on annual data are small, so ignoring labor supply does not affect our analysis. (Browning, Hansen and Heckman, 1999). This simplification makes the HLT 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

(V-3) $$H^S_{a+1,t+1} = A^S(\theta) I^{\alpha_S}_{at} H^{\beta_S}_{at} + (1 - \sigma^S) H^S_{at},$$

where the conditions $0 < \alpha_S < 1$ and $0 \leq \beta_S \leq 1$ guarantee that the problem is concave, and σ^S is the rate of depreciation of skill-S specific human capital. "θ" 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 α and β are also permitted to be S-specific, which emphasizes that schooling affects the process of learning on the job in a variety of different ways.

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 (1998, 1999) that long-run family factors correlated with income (the θ operating through $A^S(\theta)$ and the initial condition for (3)) 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 maximizes lifetime utility:

(V-4) $$\hat{S} = \underset{S}{\text{Argmax}} \left[PV^S(\theta) - D^S + \varepsilon^S \right]$$

where $PV^S(\theta)$ is the tax-adjusted present value of lifetime earnings given schooling level S, tuition costs are D^S, and ε^S represents monetized nonpecuniary benefits of schooling level S, or else unobserved components of tuition subsidies (negative costs).[27] All values and costs are discounted

[27] 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.

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. In order to compute service flow prices for capital and the different types of human capital, it is necessary to construct aggregates for 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 prices. In the two-skill economy HLT analyze, the production function at time t is defined over the inputs $\bar{H}_t^1$, $\bar{H}_t^2$ and $\bar{K}_t$, where $\bar{H}_t^1$ and $\bar{H}_t^2$ are aggregates of *utilized* skills (high school and college, respectively) supplied to production, and $\bar{K}_t$ is the aggregate stock of capital. The technology is

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

HLT estimate that $\rho_2 = 0$ but $\rho_1 = .693$, which yields an elasticity of substitution between high school and college human capital of 1.441. HLT 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 U.S. economy and I use that version.

Human capital accumulation functions (V-3) 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 U.S. economy, misspecification of the tax system does not affect parameter estimates if the model is recalibrated on aggregate data. (See Heckman, Lochner, Taber, 1998a, Appendix B). HLT (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. I now use the HLT model to evaluate the effects of alternative tax policies and tuition policies on efficiency and distribution.

Tax Effects on Human Capital Accumulation

In the absence of labor 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 (*i.e.* 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 (see 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 HLT to study alternative proposals for tax reform, their consequences for inequality, and their ability to construct the policy counterfactuals discussed in Section I.

Analyzing Two Tax Reforms[28]

Following KSW, I assume that the U.S. income tax can be captured by a progressive tax on labor income and a flat tax on capital income. Each earner has 1.22 children and is single. For

[28] This section is based in part on Heckman, Lochner, Taber (1998b).

each additional dollar beyond $9660, there is an increase in itemized deductions of 7.55 cents. An individual with labor income Y has taxable income $(Y - 9660)(1 - .0755)$. 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 labor 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, I consider two models: (1) a partial-equilibrium model in which skill prices and interest rates are fixed, and (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 methods from the "treatment effect" literature because that approach is ineffective in this context.

Table 7 presents both partial-equilibrium and general-equilibrium results measured relative to a benchmark economy with the KSW tax schedule. 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 (OJT) 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 enrollment. 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 favors 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 OJT 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 enrollment 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 enrollment. 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, is ordered in the same way. 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.

Figure 3 shows 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 capital. On a strict voting criterion for those in the current generation, the reform would not pass (43% in favor; 57% against). Evaluated at the final steady state, the reform would not be favored. (See the numbers in Table 8).

Next, consider a move to a "Flat Consumption Tax." This reform is more pro-capital and is less favorable to human capital. 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 favorable 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 increases slightly by both conventional measures. 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.

Figure 4 reveals 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 barely favored (52% in favor; 48% against). Comparing steady states, the reform would be passed by a substantial majority.

When deductibility of tuition is introduced in both reforms, and revenue neutrality is preserved, there is virtually no effect on skill formation (or anything else) in general equilibrium. This is consistent with Heckman, Lochner and Taber (1998b) and the analysis presented below which shows that general-equilibrium effects of tuition subsidies are small. The lessons from partial-equilibrium analyses are substantially misleading guides in analyzing the effects of tax and tuition 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 conclusions also hold for open economy simulations in which the interest rate is set in world markets. (Heckman, Lochner and Taber, 1999). They are robust to a variety of tax schedules and empirically grounded parameter estimates. However, the welfare criteria disagree. On the basis of a voting criterion, the switch to a consumption tax would be preferred. On the basis of the "Veil of Ignorance" evaluation criterion applied to steady states population it would not be.[29] I now turn to an analysis of tuition policy.

B. General-equililbrium Treatment Effects: A Study of Tuition Policy

This section of the paper uses the general equilibrium model developed by HLT to consider the effects of changes in tuition on schooling and earnings, accounting for general-equilibrium effects on skill prices and considering how the reform would be evaluated under different criteria. The typical microeconometric evaluation estimates the response of college enrollment to tuition variation using geographically dispersed cross-sections of individuals facing different tuition rates. These estimates are then used to determine how subsidies to tuition will raise enrollment. The impact of tuition policies on earnings is evaluated using a schooling-earnings relationship fit on pre-intervention data which does not account for the effects of anticipated skill price changes on

[29] Recall that second order stochastic dominance and Gini ranking are equivalent. See Rothschild and Stiglitz (1970).

enrollment and on the job training decisions or the response to the taxes raised to finance the tuition subsidy. Kane (1994) exemplifies this approach.

The danger in this widely used practice is that what is true for policies affecting a small number of individuals need not be true for policies that affect the economy at large. A national tuition-reduction policy that stimulates substantial college enrollment will likely compress skill prices, as advocates of the policy claim. However, agents who account for these changes will not enroll in school at the levels calculated from conventional procedures which ignore the impact of the induced enrollment on earnings. As a result, standard policy evaluation practices are likely to be misleading about the effects of tuition policy on schooling attainment and wage inequality. The empirical question is how misleading? Heckman, Lochner and Taber (1998c) show that these practices lead to estimates of enrollment responses that are ten times larger than the long-run general-equilibrium effects. HLT also improve on current practice in the treatment effects literature by considering both the gross benefits of the program and the tax costs of financing the treatment as borne by different groups.

Evaluating the general-equilibrium effects of a national tuition policy requires more information than the tuition-enrollment parameter that is the centerpiece of partial-equilibrium policy analysis. Most policy proposals extrapolate well outside the range of known experience and ignore the effects of induced changes in skill quantities on skill prices. I apply the closed economy general-equilibrium framework to evaluate tuition policies that attempt to increase college enrollment.

As noted in Section II, the standard framework for microeconometric program evaluation is partial equilibrium in character. For a given individual i, Y_i^1 is defined to be the outcome the individual receives if he participates in the program, and Y_i^0 is the outcome he receives if he does not participate. The treatment effect for person i is $\Delta_i = Y_i^1 - Y_i^0$. When interventions have general-equilibrium consequences, these effects depend on who else is treated and the market interaction between the treated and the untreated.

To see the problems that arise in the standard framework, consider instituting a national tuition policy. In this case, Y_i^0 is person i's wage if he does not attend college, and Y_i^1 is his wage if he does attend. The "parameter" Δ_i then represents the impact of college, and it can be used to estimate the impact of tuition policies on wages. It is a constant, or policy-invariant, parameter only if wages (Y_i^0, Y_i^1) are invariant to the number of college and high school graduates in the economy.

In a general-equilibrium setting, an increase in tuition increases the number of individuals who attend college, which in turn decreases the relative wages of college attendees. In this case, the program not only impacts the wages of individuals who are induced to move by the program, but it also has an impact on the wages of those who do not. For two reasons, then, the "treatment effect" framework is inadequate. First, the parameters of interest depend on who in the economy is "treated" and who is not. Second, these parameters do not measure the full impact of the program.

For example, increasing tuition subsidies may increase the earnings of uneducated individuals who do not take advantage of the subsidy. To pay for the subsidy, the highly educated would be taxed and this may affect their investment behavior. Additional educated workers enter the market as a result of the policy, depressing the earnings of other college graduates. Conventional methods ignore the effect of the policy on nonparticipants. In order to account for this effect, it is necessary to conduct a general-equilibrium analysis.

Exploring Increases in Tuition Subsidies in a General-Equilibrium Model

Heckman, Lochner and Taber (1998c) simulate the effects of a revenue-neutral $500 increase in tuition subsidy (financed by a proportional tax) on enrollment in college and wage inequality starting from our baseline economy described in the previous section. The partial-equilibrium model predicts an increase in college attendance of 5.3 percent 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 percent. 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 behavior accordingly.

Policy analysis of the type offered in the "treatment effect" literature ignores the responses of rational agents to the policies being evaluated. There is substantial attenuation of the effects of tuition policy on capital and the stocks of the different skills in our model. Simulating the effects of this policy under a number of additional alternative assumptions about the parameters of the economic model, including analysis of a case where tuition costs rise with enrollment, reproduces the basic result of substantial partial-equilibrium effects and much weaker general-equilibrium effects.

The steady state results are long-run effects. When HLT simulate the model with rational expectations, the short-run enrollment 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 at higher rates. When they simulate using myopic expectations, the short-run enrollment 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 enrollment responses in general equilibrium. Using the Gini coefficient as a measure of welfare, both partial-equilibrium and general-equilibrium simulations suggest that the reform is welfare improving. In general equilibrium, the overall variance of log wages is reduced.

Next consider the impact of a policy change on discounted earnings and utility. Decompose the

total effects into benefits and costs, including tax costs for each group. Table 9 compares outcomes in two steady states: (a) the benchmark steady state and (b) the steady state associated with the new tuition policy. Given that the estimated schooling response to a $500 subsidy is small, we instead use an extremely high $5,000 subsidy for the purpose of exploring general-equilibrium effects. The rows High School - High School report the changes in a variety of outcome measures for those persons who would be in high school under the benchmark or new policy regime; the High School - College rows report the changes in the same measures for high school students in the benchmark who are induced to attend college only by the new policy; College - High School outcomes refer to those persons in college in the benchmark economy who only attend high school after the new policy is put in place; and so forth.

By the measure of the present value of earnings (column 3), some of those induced to change are worse off. Contrary to the monotonicity assumption built into the LATE parameter of Imbens and Angrist (1994), defined in this context as the effect of tuition change on the earnings of those induced to go to college, HLT find that the tuition policy produces a two-way flow. Some people who would have attended college in the benchmark regime no longer do so. The rest of society also is affected by the policy–again, contrary to the implicit assumption built into LATE that only those who change status are affected by the policy. People who would have gone to college without the policy and continue to do so after the policy are financially worse off for two reasons: (a) the price of their skill is depressed and (b) they must pay higher taxes to finance the policy.

However, they now receive a tuition subsidy and for this reason, on net, they are better off both financially and in terms of utility. Those who would abstain from attending college in both steady states are also better off in the second steady state. They pay higher taxes, but their skill becomes more scarce and their wages rise. Those induced to attend college by the policy are better off in terms of utility but are not always better off in terms of income. For example, individuals from ability quartiles 2 and 3 have lower net incomes as a result of the tuition policy; however, their utility rises due to a strong taste for college education. While most groups gain about the same in terms of utility, there is substantial variation in the effects on lifetime earnings. Note that neither category of non-changers is a natural benchmark for a "difference in differences" estimator. The movement in their wages before and after the policy is due to the policy and cannot be attributed to a benchmark "trend" that is independent of the policy. The implicit assumptions that justify the widely used difference in differences estimator do not apply here. The tax system and the market make the "nontreated" affected by the policy. (See the discussion in Heckman, LaLonde and Smith, 1999). These conclusions are robust as to whether a closed-economy or open-economy model is used. (Heckman, Lochner and Taber, 1999).

C. General-Equilibrium vs. Partial-Equilibrium Approaches

The sharp contrast between the general-equilibrium estimates of program 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 evaluation of programs with economy-wide effects, it also offers an economically interpretable evaluation of a policy. As discussed in Appendix A, many of the parameters estimated in the micro-economic "treatment effect" literature, are not those required for the cost-benefit evaluations.

Critics of the general-equilibrium approach dismiss it because it is based on "questionable" empirical foundations. They ignore, or trivialize, 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 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 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, Hansen and Heckman (1999) summarize current developments in this field and present an agenda for research in uniting micro evidence with macro general-equilibrium models. The empirical analysis presented in Heckman, Lochner and Taber (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.

VI. Summary and Conclusions

Diversity, heterogeneity and involuntary redistribution are the defining features of the modern welfare state. Disagreements over outcomes and their valuation gives rise to the demand for publicly justified evaluations of social programs.

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 program impacts as either uninteresting or empirically irrelevant. Modern welfare economics explicitly recognizes heterogeneity in outcomes but focuses on one measure of alternative policies because of its "ethical correctness". All of the measures proposed in the modern literature on welfare economics ignore or deplore self-interested agents who evaluate policies

by comparing *their* initial positions under current policies with their positions under proposed alternative policies. A measure that enumerates gainers or losers and quantifying the magnitudes of their losses comes closer to capturing the information useful in a modern democracy than a criterion based on axioms of correct behavior 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 their 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 "treatment effect" approaches ignore general-equilibrium considerations, which are also left implicit in the approach of modern welfare economics. General equilibrium is center stage in the macro policy evaluation model. Virtually all methods ignore the heterogeneity in program 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 program outcomes is an empirically important feature of the data even after conditioning on the observed characteristics. 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 analyze the consequences of heterogeneity and diversity for the evaluation of social programs. The benefits of this approach are examined in evaluating several proposed tax reforms and in evaluating a proposed tuition policy.

Accounting for heterogeneity, diversity and general equilibrium has important implications for the way we evaluate social policies. An evaluation criterion that counts gainers and losers produces a very different asssesment of the suitability of policies than the "ethically correct" criteria advocated in modern welfare economics, and one that is more closely attuned to the requirements of positive political economy.

References

[1] Aakvik, A., J. Heckman and E. Vytlacil (1998), "Local Instrumental Variables and Latent Variable Models For Estimating Treatment Effects," unpublished manuscript, University of Chicago.

[2] Ahmad, and N. Stern (1984), *The Theory of Tax Reform and Indian Indirect Taxes",* 25(3), 259-298.

[3] Auerbach, A. and L. Kotlikoff (1987), *Dynamic Fiscal Policy*, Cambridge: Cambridge University Press.

[4] Baumol, W. and Quandt, R. (1966), "The Demand For Abstract Transport Modes: Theory and Measurement," **Journal of Regional Science**, 6, 13-26.

[5] Bell, S. and C. Reesman (1987), "AFDC Homemaker - Health Aide Demonstration: Trainee Potential And Performance" (Abt Associates: Cambridge, Massachusetts).

[6] Bentham, J., (1824), *Handbook of Political Fallacies*, republished by Johns Hopkins Press, 1952.

[7] Besley, T. and S. Coate (1998), "The Public Choice Critique of Welfare Economics: An Exploration," unpublished manuscript, London School of Economics, December.

[8] Bloom, H., L. Orr, G. Cave, S. Bell, and F. Doolittle (1993), *The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months.* Bethesda, MD: Abt Associates.

[9] Boadway, R. and Bruce, N. (1984), *Welfare Economics.* Oxford, England: Basil Blackwell.

[10] Browning, M., Hansen, and L, Heckman, J. (1999), "Estimating Dynamic General Equilibrium Models," forthcoming in *Handbook of Macroeconomics*, ed. by J. Taylor and M. Woodford. Amsterdam: North Holland.

[11] Cameron, S. and J. Heckman (1999), "Can Tuition Policy Combat Rising Wage Inequality," ed. by M. Kosters, in *Financing College Tuition: Government Policies and Social Priorities,* AEI Press: Washington, DC., 75-125.

[12] Chipman, J. and J. Moore (1976), "Why An Increase in GNP Need Not Imply An Improvement in Potential Welfare," **Kyklos**, 29, 391-418.

[13] Coate, S. (1998), "Welfare Economics and The Evaluation of Policy Changes," unpublished paper, Department of Economics, Cornell University.

[14] Couch, K. (1992), "New Evidence on the Long-Term Effects of Employment Training," **Journal of Labor Economics**, 10:4, 380-388.

[15] Csörgo (1993), *Quantile Processes with Statistical Applications.* Philadelphia: Society for Industrial and Applied Mathematics.

[16] Davies, J. and J. Whalley (1991), "Taxes and Capital Formation: How Important Is Human Capital?," in B. Bernheim and J. Shoven, eds., *National Saving and Economic Performance*, Chicago: University of Chicago Press.

[17] Domencich, T. and D. McFadden (1975), *Urban Travel Demand: A Behavioral Analysis.* North Holland: Amsterdam.

[18] Frechet, M. (1951), "Sur Les Tableux de Correlation Dont Les Marges Sont Donnes," **Ann. University Lyon: Sect. A**, 14, 53-77.

[19] Gabaix, X., (1998), "Cost of Inequality" unpublished manuscript, Department of Economics, MIT.

[20] Gueron, J. and E. Pauly (1991), *From Welfare to Work*, New York: Russell Sage Foundation.

[21] Harberger, A. (1971), "Three Basic Postulates for Applied Welfare Economics: An Interpretive Essay," **Journal of Economic Literature**, 9, 785-797.

[22] _____ (1978), "On the Use of Distributional Weights in Social-Cost Benefit Analysis," **Journal of Political Economy**, 86, S87-S120.

[23] Harsanyi, J. (1955), "Cardinal Welfare, Individualistic Ethics and Interpersonal Comparisons of Utility," **Journal of Political Economy**, 63, 309-321.

[24] _____ (1975), "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls' Theory," **American Political Science Review**, 69(2), 594-606.

[25] Heckman, J. (1974a), "The Effect of Child Care Programs on Women's Work Effort," **Journal of Political Economy,** 82(2), 5136-5163. Reprinted in *Economics of the Family: Marriage, Children, and Human Capital,* ed. by T.W. Schultz. Chicago: University of Chicago Press.

[26] _____ (1974b), "Shadow Prices, Market Wages and Labor Supply," **Econometrica,** 42(4) 679-94.

[27] _____ (1976), "A Life Cycle Model of Earnings, Learning and Consumption," **Journal of Political Economy**, 84(4, pt.2), S11-S44.

[28] _____ (1978), "Dummy Endogenous Variables In A Simultaneous Equation System," **Econometrica**, 46(4), 931-959.

[29] _____ (1990a), "Varieties of Selection Bias," **American Economic Review**, 80(2), 313-318.

[30] _____ (1992), "Randomization and Social Program Evaluation," in *Evaluating Welfare and Training Programs*, ed. by C. Manski and I. Garfinkel. Boston: Harvard University Press, 201-230.

[31] _____ (1997), "Instrumental Variables: A Study of Implicit Behavioral Assumptions in One Widely-Used Estimator Used in Making Program Evaluations," **Journal of Human Resources,** Summer.

[32] _____ (1999), "Econometric Evaluation of Social Programs," forthcoming in *Handbook of Econometrics*, Vol. V, Amsterdam: North Holland.

[33] Heckman, J., L. Lochner and C. Taber (1998a), "Explaining Rising Wage Inequality: Explorations With A Dynamic General Equilibrium Model of Labor Earnings With Heterogeneous Agents," **Review of Economic Dynamics,** 1(1), January.

[34] Heckman, J., N. Hohmann, M. Khoo and J. Smith (1997), "Did We Learn the Right Lesson from the National JTPA Study? Substitution Bias in Social Experiments," Mimeo, University of Chicago, under revision, **Quarterly Journal of Economics**.

[35] Heckman, J. and B. Honoré (1990), "The Empirical Content of the Roy Model," **Econometrica**, 58(5), 1121-1149.

[36] Heckman, J., H. Ichimura, J. Smith and P. Todd. (1998), "Characterizing Selection Bias Using Experimental Data," **Econometrica**, 66(5), 1017-1098.

[37] Heckman, J., H. Ichimura and P. Todd (1997a), "Matching As An Econometric Evaluation Estimator: Evidence on Its Performance Applied To The JTPA Program, Part I. Theory and Methods," **Review of Economic Studies,** October.

[38] _____ (1998), "Matching As An Econometric Estimator: Evidence on Its Performance Applied to the JTPA Program, Part II. Empirical Evidence," **Review of Economic Studies**, April.

[39] Heckman, J. and R. Robb (1985), "Alternative Methods For Evaluating The Impact of Interventions," in *Longitudinal Analysis of Labor Market Data,* ed. by J. Heckman and B. Singer. New York: Cambridge University Press, 156-245.

[40] Heckman, J. and J. Smith (1998), "Evaluating The Welfare State," in S. Strom, ed, *Econometrics and Economic Theory in the Twentieth Century*, Econometric Monograph Series, 31, Cambridge: Cambridge University Press.

[41] _____ (1993), "Assessing The Case For Randomized Evaluation of Social Programs," in *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policies*, ed. by K. Jensen and P. K. Madsen. Copenhagen: Danish Ministry of Labor, 35-96.

[42] _____ (1995), "Assessing The Case For Social Experiments," **Journal of Economic Perspectives**, 9, 85-110.

[43] _____ (1996), "Experimental and Nonexperimental Evaluation," in *International Handbook of Labor Market Policy and Evaluation*, ed. by G. Schmidt, J.O' Reilly and K. Schömann. Cheltenham, U.K: Elgar Publishers.

[44] Heckman, J., R. LaLonde, and J. Smith (1999), "The Economics and Econometrics of Active Labor Market Programs," in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, Vol. III, North Holland.

[45] Heckman, J., J. Smith and N. Clements (1997), "Making The Most Out of Program Evaluations and Social Experiments: Accounting for Heterogeneity in Program Impacts," **Review of Economic Studies**, October.

[46] Heckman, J. and E. Vytlacil (1998), "Instrumental Variable Estimation of A Correlated Random Coefficient Model," **Journal of Human Resources**, Fall.

[47] _____ (1999), "Local Instrumental Variables And Latent Variable Models For Identifying and Bounding Treatment Effects", forthcoming in **The Proceedings of The National Academy of Sciences of the USA.**

[48] Imbens, G. and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," **Econometrica**, 62, 467-471.

[49] Kane, T., (1994), "College Entry by Blacks Since 1970: The Role of College Costs, Family Background and the Returns to Education," **Journal of Political Economy**, 102, 878-911.

[50] Karoly, L. (1992), "Changes In the Distribution of Individual Earnings in the United States: 1967-1988," **Review of Economics and Statistics**, 74(1), 107-115.

[51] Katz, D., B. Gutek, R. Kahn and E. Barton (1975), *Bureaucratic Encounters: A Pilot Study in the Evaluation of Government Services*. Ann Arbor, MI: Institute for Social Research.

[52] Knight, F. (1921), *Risk, Uncertainty and Profit.* New York: Houghton Mifflin Company.

[53] Kotlikoff, L., K. Smetters, and J. Walliser (1997), "The Economic Impact of Privatizing Social Security," unpublished manuscript, Boston University.

[54] Laffont, J. J. (1989), *Fundamentals of Public Economics.* Cambridge, MA: MIT Press.

[55] Lancaster, K. (1971), *Consumer Demand: A New Approach.* New York: Columbia University.

[56] Lewis, H. Gregg (1963), *Unionism and Relative Wages*, Chicago: University of Chicago Press.

[57] Lucas, R. (1987), *Models of Business Cycles*, Oxford: Basil Blackwell.

[58] Lucas, R. and T. Sargent (1981), "Introduction," in *Essays on Rational Expectations and Econometric Practice.* Minneapolis: University of Minnesota Press, xi-xl.

[59] Maital, Schlomo (1973), "Public Goods and Income Distribution: Some Further Results," **Econometrica**, 41(3), 561-568.

[60] Marschak, J. (1953), "Economic Measurements For Policy and Prediction," in *Studies in Econometric Method,* ed. by W. Hood and T. Koopmans. New York: John Wiley, 1-26.

[61] Moffitt, R. (1992), "Evaluation of Program Entry Effects," in *Evaluating Welfare and Training Programs,* ed. by C. Manski and I. Garfinkel. Boston: Harvard University Press, 231-252.

[62] Moulin, H. (1983), *The Strategy of Social Choice.* Amsterdam: North Holland.

[63] Orr, L., H. Bloom, S. Bell, W. Lin, G. Cave, F. Doolittle (1995), *The National JTPA Study: Impacts, Benefits and Costs of Title II-A.* Bethesda, MD: Abt Associates.

[64] Rawls, J., (1971), *A Theory of Justice.* Cambridge, MA: Harvard University Press.

[65] Roemer, J., (1996), *Theories of Distributive Justice*, Cambridge: Harvard University Press.

[66] Rothschild, M. and J. Stiglitz (1970), "Increasing Risk I: A Definition," **Journal of Economic Theory**, 2, 225-243.

[67] Roy, A. (1951): "Some Thoughts on the Distribution of Earnings," **Oxford Economic Papers**, 3, 135-146.

[68] Saez, E. (1998), "Using Elasticities to Derive Optimal Income Tax Rates," unpublished paper, MIT, Department of Economics.

[69] Sen, A. (1973), *On Economic Inequality.* Oxford: Clarendon Press.

[70] _____ (1979), "Strategies and Revelation: Informational Constraints in Public Decisions," in *Aggregation and Revelation of Preferences*, ed. by J. J. Laffont. Amsterdam: North Holland.

[71] Smith, J. (1997), "The JTPA Selection Process: A Descriptive Analysis," in *Performance Standards in a Federal Bureaucracy: Analytical Essays on the JTPA Performance Standards System*, ed. by J. Heckman. Kalamazoo, MI: W.E. Upjohn Institute.

[72] Theil (1961), *Economic Forecasts and Policy*, Amsterdam: North Holland.

[73] Tinbergen, J. (1956), *Economic Policy: Principles and Design.* Amsterdam: North Holland.

[74] Trostel, P. (1993), "The Effect of Taxation on Human Capital," **Journal of Political Economy**, 101(2), 327-50.

[75] U.S. General Accounting Office (1996): "Job Training Partnership Act: Long-Term Earnings and Employment Outcomes." GAO/HEHS-96-40.

[76] Vickrey, W. (1945), "Measuring Marginal Utility By Reactions To Risk," **Econometrica,** 13, 319-333.

[77] _____ (1961), "Risk Utility and Social Policy," Social Research.

[78] Young, P. (1994), *Equity in Theory and Practice*, Princeton: Princeton University Press.

</ref_section>

Appendix A
The Relationship Between The Requirements of Cost-Benefit Analysis and The Information Produced From Social Experiments and The Microeconometric "Treatment Effect" Literature

In this appendix I relate the parameters estimated in the micro-econometric evaluation literature or the literature on social experiments to the parameters needed to perform cost-benefit analysis. I follow the literature in cost-benefit analysis and assume that the policy being evaluated has a voluntary component and that valid evaluations of a policy can be derived from looking at the impact of the policy on self-selected participants and nonparticipants.

I postulate the following framework. For a given program associated with policy j, there are two discrete outcomes corresponding to direct receipt of treatment ($D_j = 1$, for program participation) or not ($D_j = 0$), and a set of program intensity variables φ_j defined under policy j that affect outcomes in the two states and the allocation of persons to "treatment" or nontreatment. The program intensity variables φ_j may be discrete or continuous. Policy "0" is a no-intervention benchmark with program intensity φ_0.

Assuming that costless lump-sum transfers are possible, that a single social welfare function governs the distribution of resources and that prices reflect true opportunity costs, traditional cost-benefit analysis (see, *e.g.*, Harberger (1971) or Boadway and Bruce (1984)) seeks to determine the impact of programs on the total output of society. Efficiency becomes the paramount criterion in this framework, with the distributional concerns assumed to be taken care of through lump sum transfers and taxes. In this framework, impacts on total output, as in the evaluation criterion (I-3), are the only objects of interest in evaluating policies.

For policy j let Y_{ji}^1 and Y_{ji}^0 be individual output for person i in the direct participation ($D_j = 1$) and direct non-participation ($D_j = 0$) state, respectively. The vector of program intensity variables φ_j operates on all persons within the context of program j, although its effect need not be uniform. It determines, in part, participation in the program. One may write $D_j(\varphi_j)$ as the indicator for participating in program j when program intensity is φ_j. To simplify notation I keep implicit any conditioning on personal characteristics that may affect both participation and outcomes. I define $c_j(\varphi_j)$ as the social cost of φ_j denominated in units of output. In general, policies could be designed for specific persons but we do not consider that possibility here. I assume that $c_j(0) = 0$ and that c is convex and increasing in φ_j. The value φ_0 defines another benchmark policy, "0", in which there is no program and therefore no participants. This policy has associated cost function $c_0(\varphi_0)$.

When $\varphi_j = 0$, there might be effects of policy j on output that distinguish that policy from the no policy regime "0". A law that is universally assented to and accepted may raise output at no cost (*e.g.*, adopting a convention about driving on the right hand side of the road). Output could be different in a policy without the law (policy "0") but the direct costs of enforcement would be

the same under both policies.

Letting $N_1(\varphi_j)$ be the number of direct program participants and $N_0(\varphi_j)$ be the rest of the population, the total output of society under policy j at program intensity level φ_j is

$$N_1(\varphi_j)E(Y_j^1 \,|D(\varphi_j) = 1,\varphi_j) + N_0(\varphi_j)E(Y_j^0 \,|D(\varphi_j) = 0,\varphi_j) - c(\varphi_j),$$

where $N_1(\varphi_j) + N_0(\varphi_j) = \bar{N}$ is the total number of persons in society. "φ_j" appears twice in the conditioning arguments: as a determinant of D_j and as a determinant of the output levels in the different states. Vector φ_j is general enough to include financial incentive variables as well as mandates that assign persons to a particular treatment state. Recall that I keep conditioning on personal characteristics implicit.

Assume for simplicity the differentiability of the treatment choice and mean outcome functions and further assume that φ_j is a scalar, a simplifying assumption that is easily relaxed. The change in output in response to a marginal increase in the policy intensity parameter φ_j from any given position is:

$$\begin{aligned} M\left(\varphi_j\right) \;=\; & \frac{\partial N_1\left(\varphi_j\right)}{\partial \varphi_j}\left[E\left(Y_j^1|D_j(\varphi_j)=1,\varphi_j\right) - E\left(Y_j^0|D_j(\varphi_j)=0,\varphi_j\right)\right] \\ & + N_1\left(\varphi_j\right)\left[\frac{\partial E\left(Y_j^1|D(\varphi_j)=1,\varphi_j\right)}{\partial \varphi_j}\right] \\ & + N_0(\varphi_j)\left[\frac{\partial E\left(Y_j^0|D(\varphi_j)=0,\varphi_j\right)}{\partial \varphi_j}\right] - c_j'\left(\varphi_j\right). \end{aligned}$$

The first term arises from the change in the number of participants induced by the policy change. The second and third terms arise from changes in output among participants and nonparticipants induced by the policy change. The fourth term is the marginal direct output cost of the change in the intensity of policy φ_j.

In principle, this measure could be estimated from time-series data on the change in aggregate GNP occurring after the policy intensity parameter is varied. Under the assumption of a well-defined social welfare function with interior solutions and the additional assumption that prices are constant at initial values, an increase in GNP at base period prices raises social welfare.[1]

If marginal program intensity changes under policy regime j have no effect on intra-sector mean output, the bracketed expressions in the second and third terms are zero. In this case, the parameters of interest are:

[1]See, *e.g.*, Laffont (1989, p. 155), or the comprehensive discussion in Chipman and Moore (1976).

(i) $\dfrac{\partial N_1\left(\varphi_j\right)}{\partial \varphi_j}$ — the number of people induced into program j by the change in φ_j,

(ii) $E\left(Y_j^1 | D_j(\varphi_j)=1, \varphi_j\right) - E\left(Y_j^0 | D_j(\varphi_j)=0, \varphi_j\right)$ — the mean output difference between participants and nonparticipants.

(iii) $c_j'(\varphi_j)$ — the direct social marginal cost of policy j at program intensity level φ_j.

It is revealing that nowhere on this list are the parameters that receive the most attention in the micro econometric policy evaluation literature or the literature on social experiments. (See, *e.g.*, Heckman and Robb, 1985, 1986 or Heckman, LaLonde and Smith, 1999). These are:

(a) $E\left(Y_j^1 - Y_j^0 | D_j(\varphi_j)=1, \varphi_j\right)$ The effect of "treatment on the treated" for persons in regime j at policy intensity φ_j.

(b) $E\left(Y_j^1 - Y_j^0 | \varphi_j = \bar{\varphi}\right)$ where $\varphi_j = \bar{\varphi}$ sets $N_1(\bar{\varphi}) = \bar{N}$. This is the effect of universal direct participation in program j compared to universal nonparticipation in j at level of program intensity $\bar{\varphi}$.

(c) $E\left(Y_j^1 - Y_j^0 | \varphi_j\right)$ The effect of randomly selecting a single person for direct treatment and forcing their compliance with this treatment compared to their position in the no participation state under policy j at program intensity level φ_j.

Parameter (ii) can be obtained from simple mean differences between the treated and the nontreated. No adjustment for selection bias is required. Parameter (i) can be obtained from knowledge of the net movement of persons into or out of direct participation in the program in response to the policy change, something usually not measured in micro policy evaluations or social experiments (for discussions of this problem, see Moffitt, 1992 or Heckman, 1992). Parameter (iii) can be obtained from cost data. It should include full social costs of the program, including the welfare cost of raising public funds, although these are often ignored.

It is informative to place additional structure on this model. This leads to a representation of a criterion that is widely used in the literature on microeconomic program evaluation and also establishes a link with the discrete choice literature in econometrics. Assume a binary choice random utility framework like that used in the Roy model. Suppose that under policy regime j with program intensity level φ_j agents make choices to directly participate or not based on net utility

and that policies affect participant utility through an additively-separable term, $k(\varphi_j)$, that is assumed scalar and differentiable. Net utility from participating in the program is $U_j = X + k(\varphi_j)$, where k is monotonic in φ_j and where the joint distributions of (Y_j^1, X) and (Y_j^0, X) are $F(y_j^1, X)$ and $F(y_j^0, X)$, respectively.[2] In the special case of the Roy model, $X = Y_j^1 - Y_j^0$ and $k = 0$. In general, $D_j(\varphi_j) = 1(U_j \geq 0) = 1(X \geq -k(\varphi_j))$, so

$$N_1\left(\varphi_j\right) = \bar{N}\Pr\left(U_j \geq 0\right) = \bar{N}\int_{-k(\varphi_j)}^{\infty} f\left(x\right)dx$$
$$N_0\left(\varphi_j\right) = \bar{N}\Pr\left(U_j < 0\right) = \bar{N}\int_{-\infty}^{-k(\varphi_j)} f\left(x\right)dx.$$

Total output is

$$\bar{N}\int_{-\infty}^{\infty} y^1 \int_{-k(\varphi_j)}^{\infty} f\left(y^1, x|\varphi_j\right)dxdy^1 + \bar{N}\int_{-\infty}^{\infty} y^0 \int_{-\infty}^{-k(\varphi_j)} f\left(y^0, x|\varphi_j\right)dxdy^0 - c_j\left(\varphi_j\right).$$

Under standard conditions, one may differentiate under the integral sign to obtain the following expression for the marginal change in output with respect to a change in intensity parameters φ_j within policy regime j :

$$M\left(\varphi_j\right) =$$
$$\bar{N}k'\left(\varphi_j\right) f_x\left(-k\left(\varphi_j\right)\right)\left[E\left(Y_j^1|D(\varphi_j) = 1, X = -k\left(\varphi_j\right), \varphi_j\right) - E(Y_j^0|D(\varphi_j) = 0, X = -k\left(\varphi_j\right), \varphi_j)\right]$$
$$+\bar{N}\left[\int_{-\infty}^{\infty} y^1 \int_{-k(\varphi_j)}^{\infty} \frac{\partial f\left(y^1, x|\varphi_j\right)}{\partial \varphi_j} dxdy^1 + \int_{-\infty}^{\infty} y^0 \int_{-\infty}^{-k(\varphi_j)} \frac{\partial f\left(y^0, x|\varphi_j\right)}{\partial \varphi_j} dxdy^0\right] - c_j'\left(\varphi_j\right),$$

where f_x, the marginal density of X, is evaluated at $X = -k(\varphi_j)$.

This model has a well-defined marginal entry condition: $X \geq -k(\varphi_j)$. The first set of terms corresponds to the gain arising from the movement of persons at the margin (the term in brackets) weighted by the proportion of the population at the margin, $f_x(-k(\varphi_j))$, times the number of people in the population. This term is the net gain from switching from nonparticipant to participant status. The expression in brackets in the first term is a limit form of the "local average treatment effect" of Imbens and Angrist (1994). This term is estimated in Aakvik, Heckman and Vytlacil (1998) and further analyzed in Heckman and Vytlacil (1999). The second set of terms is the within-treatment-status change in output resulting from the change in the program intensity parameter. This term is ignored in many microeconomic evaluation studies. It describes how people who do not switch their participation status are affected by the policy change. The third term is the direct marginal social cost of the policy change, which is rarely estimated. At a social planner's optimum, $M(\varphi_j) = 0$, provided standard second order conditions are satisfied. Marginal benefit should equal the marginal cost. Either a cost-based measure of marginal benefit or a

[2] These are assumed to be absolutely continuous with respect to Lebesgue measure.

benefit-based measure of cost can be used to evaluate the marginal gains or costs of the change in policy intensity.

Observe that the local average treatment effect is simply the effect of treatment on the treated for persons at the margin ($X = -k(\varphi_j)$) :

$$\begin{aligned} & E\left(Y_j^1 | D_j(\varphi_j) = 1, X = -k\left(\varphi_j\right), \varphi_j\right) - E\left(Y_j^0 | D_j(\varphi_j) = 0, X = -k\left(\varphi_j\right), \varphi_j\right) \\ = \; & E\left(Y^1 - Y^0 | D(\varphi_j) = 1, X = -k\left(\varphi_j\right), \varphi_j\right). \end{aligned}$$

The proof of this result is immediate once it is recognized that the set $X = -k(\varphi_j)$ is the indifference set for this problem. Thus, the LATE parameter is a marginal version of the conventional "treatment on the treated" evaluation parameter for gross outcomes. This parameter is but one of the three ingredients required to produce an evaluation of social welfare under the cost-benefit criterion.

The conventional evaluation parameter "treatment on the treated"

$$E\left(Y_j^1 - Y_j^0 | D_j\left(\varphi_j\right) = 1, X, \varphi_j\right),$$

produced in the microeconometric program evaluation or from social experiments does not incorporate costs, does not correspond to a marginal change and includes the effect of intramarginal changes. This parameter is in general inappropriate for evaluating the effect of a policy change on GNP. If model (A) of Section III is true, however, then treatment on the treated is the same as the average treatment effect and it is correct economic parameter for conducting a cost-benefit analysis. Under certain conditions which I now make precise, the treatment on the treated parameter is sometimes informative about the gross gain accruing to the economy from the existence of program j at level φ_j compared to the alternative of shutting it down and switching to policy "0". The social cost associated with policy "0" is $c_0(\varphi_0)$, which we assume is zero: $c_0(\varphi_0) = 0$. The essential condition is assumption (A-1) in Section II.

The appropriate criterion for an all or nothing evaluation of a policy at level φ_j is

$$\begin{aligned} A\left(\varphi_j\right) \; = \; & \left\{N_1\left(\varphi_j\right) E\left(Y_j^1 | D_j(\varphi_j) = 1, \varphi_j\right) + N_0\left(\varphi_j\right) E\left(Y_j^0 | D_j(\varphi_j) = 0, \varphi_j\right) - c_j\left(\varphi_j\right)\right\} \\ & - \bar{N} E\left(Y_0 | \varphi_0\right). \end{aligned}$$

In the no policy regime, there is only one output Y_0 and everyone is in the "no program" state. If $A(\varphi_j) > 0$, total output is increased by establishing program j at level φ_j. In the special case where the outcome in the nonparticipation state under regime j, Y_j^0, is the same as the outcome in the no-program state (Y_0) both for participants and nonparticipants under regime j, we have

(AA-1) $$E(Y_j^0 \,|\, D_j(\varphi_j) = 0,\ \varphi_j) \; = E(Y_0 \,|\, D_j(\varphi_j) = 0,\ \varphi_0)$$

and

(AA-2) $$E(Y_j^0 \,|\, D_j(\varphi_j) = 1,\ \varphi_j) = E(Y_0 \,|\, D_j(\varphi_j) = 1,\ \varphi_0).$$

The right hand sides of both expressions describe hypothetical conditional expectations. The right hand side of (AA-1) is what the outcome in the no-program state would be for persons who do not directly participate in the program under policy j with parameters φ_j, *i.e.*, those for whom $D_j(\varphi_j) = 0$. The right hand side of (AA-2) is the corresponding expression for persons who would participate in the program under policy j with intensity parameters φ_j, *i.e.*, those for whom $D_j(\varphi_j) = 1$. These conditioning statements select out, respectively, non-participants and participants in policy regime j and compute the expected values of output in the policy "0" regime.

Assuming that the probability of participation in regime j under program intensity level φ_j does not depend on the value of φ_0 in the no-program state:

(A-2) $$\Pr\left(D_j = 1 | \varphi_j, \varphi_0\right) = \Pr\left(D_j = 1 | \varphi_j\right),$$

under assumption (A-1) we may use the law of iterated expectations to write

$E(Y^0 \,|\, \varphi_0) =$

$$E(Y_0 \,|\, D_j(\varphi_j) = 1, \varphi_0) \Pr(D_j(\varphi_j) = 1 \,|\, \varphi_j) + E(Y_0 \,|\, D_j(\varphi_j) = 0, \varphi_0) \Pr(D_j(\varphi_j) = 0 \,|\, \varphi_j).$$

From (AA-1) and (AA-2) and (A-2) one obtains

$E(Y^0 \,|\, \varphi_0) =$

$$E(Y_j^0 \,|\, D_j(\varphi_j) = 1, \varphi_j) \Pr(D_j(\varphi_j) = 1 \,|\, \varphi_j) + E(Y_j^0 \,|\, D_j(\varphi_j) = 0, \varphi_j) \Pr(D_j = 0 \,|\, \varphi_j).$$

Substituting for $E(Y_j^0 \,|\, \varphi_0)$ in the expression for $A(\varphi_j)$, we obtain

(AA-3) $$A(\varphi_j) = N(\varphi_j) E(Y_j^1 - Y_j^0 \,|\, D_j(\varphi_j) = 1, \varphi_j) - c_j(\varphi_j),$$

which vindicates the use of the parameter "treatment on the treated" as an evaluation parameter in the case in which there are no general equilibrium effects in the sense of assumption (A-1). This important case is applicable to small-scale social programs with partial participation. For evaluating the effect of "fine-tuning" the intensity levels of existing policies, measure $M(\varphi_j)$ is more appropriate. Neither parameter captures the distributional consequences of the policy change.

As a matter of practice the treatment effect literature focuses its exclusion attention on gross gains and rarely measures the full costs of the program it evaluates. Heckman and Smith (1998) demonstrate the empirical importance of adjusting for costs in evaluating job training programs. Accounting for full costs substantially revises the estimates of the returns.

Appendix B

The data analyzed in Sections III and IV of this paper were gathered as part of an experimental evaluation of the training programs financed under Title II-A of the Job Training Partnership Act (JTPA). The experiment was conducted at a sample of sixteen JTPA training centers around the country. Data were gathered on JTPA applicants randomly assigned to either a treatment group allowed access to JTPA training services or a control group denied access to JTPA services for 18 months. Random assignment covered some or all of the period from November 1987 to September 1989 at each center. A total of 20,601 persons were randomly assigned.

Follow-up interviews were conducted with each person in the experimental sample during the period from 12-24 months after random assignment. This interview gathered information on employment, earnings, participation in government transfer programs, schooling, and training during the period after random assignment. The response rate for this survey was around 84 percent. The sample used here includes only those adult women who (1) had a follow-up interview scheduled at least 18 months after random assignment, (2) responded to the survey, and (3) had useable earnings information for the 18 months after random assignment.

The sample was chosen to match that used in the 18-month experimental impact study by Bloom, et.al. (1993). As in that report, the earnings measure is the sum of self-reported earnings during the 18-months after random assignment. This earnings sum is constructed from survey questions about the length, hours per week, and rate of pay on each job held during this period. Outlying values for the earnings sum are replaced by imputed values as in the impact report. However, imputed earnings values used in the report for adult female non-respondents are not used. For a more complete description of this data, see Heckman and Smith (1998) or Heckman, Ichimura, Smith and Todd (1998).

Table 1A

Population Data Requirements To Implement Criterion
General Population (Compulsory Programs)
Program j compared to program k

	Cost Benefit	Benthamite Criterion	General Social Welfare Function With Interdependent Peferences	Selfish Voting
		$E(U(Y_j,0))-$ $E(U(Y_k,\theta)) \geq 0$	$W(j) > W(k)$	
Criterion	$E(Y_j) - E(Y_k) \geq 0$	$E(U(Y_\ell,\theta)) =$ $\int U(y_\ell,\theta)dF(y_\ell,\theta)$ $\ell = j,k$	$W(\ell) = W(U_1(Y_{1\ell},...,Y_{N\ell}),...,U_N(Y_{1\ell},...,Y_{N\ell}))$ $\ell = j,k$**	$\int 1(U(y_j,\theta) \geq U(y_k,\theta))$ $dF(y_j,y_k,\theta) \geq)$
Require	Population Means $E(Y_j), E(Y_k)$	U(Y_ℓ,θ) and distribution of $(Y_\ell,\theta)F(y_\ell,\theta)$ $\ell = j,k$	Need each $U_i(Y_{1\ell},...,Y_{N\ell})$ for all i. Need outcomes for each person*	Need $U(Y,\theta), F(y_j,y_k,\theta)$***
Estimable on Aggregate Time Series	Yes, if data exist on aggregate economy in both regimes and can eliminate trend	No, unless θ the same for everyone (homogeneity); $U(Y_\ell,\theta)$ known and the moment $\int$ U($y_\ell,\theta)dF(y_\ell)$ known or estimable $\ell = j,k$	No, except in the special cases previously considered.	No

Notes:

* In special cases, summary statistics of the distribution of Y may suffice.

** This includes the special case where individual utility depends only on individual consumption.

*** For altruistic voting, U depends on $Y_{1j}, ..., Y_{Nj}$ or various sub-aggregators.

Table 1B

Population Data Requirements To Implement Criterion
General Population (Compulsory Programs)
Program j compared to program k

	Cost Benefit	Benthamite Criterion	General Social Welfare Function With Interdependent Preferences*	Selfish Voting
Criterion	$E(Y_j \mid D_j = 1) - E(Y_k \mid D_j = 1)$ What j participants gain over state k	$E(U(Y_\ell, \theta \mid D_j = 1)) - E(U(Y_k, \theta) \mid D_j = 1)) > 0$	$W(j) > W(k)$	
		where $E(U(Y_\ell, \theta) \mid D_j = 1) = \int U(y_\ell, \theta) dF(y_\ell, \theta \mid D_j = 1)$ $\ell = j, k$	$W(\ell) = W(U_1(Y_{1\ell}, ..., Y_{N\ell}), ..., U_N(Y_{1\ell}, ..., Y_{N\ell}))$ $\ell = j, k$	$\int 1(U(y_j, \theta) > U(y_k, \theta)) dF(y_j, y_k, \theta \mid D_j = 1) \geq 0$
Require	Population Conditional Means $E(Y_j \mid D_j = 1)$, $E(Y_k \mid D_j = 1)$	$U(Y_\ell, \theta)$ for $D_j = 1$, $\ell = j, k$	Need each $U_i(Y_{1\ell}, ..., Y_{N\ell})$ for whom $D = 1$ Need identity of outcomes for each persons*	Need $U(Y, \theta)$, $F(y_j, y_k, \theta) \mid D_j = 1)$.
Estimable on Aggregate Time Series Data?	Yes, if aggregate data for participants exist in both regimes, can eliminate trend.	No, unless θ the same for everyone (homogeneity); $U(Y_\ell, \theta)$ known and the moment $\int U(y_\ell, \theta) dF(y_\ell \mid D_j = 1)$ known or estimable $\ell = j, k$	No, except in the special cases previously considered.	No

Notes:

* This criterion is not well defined with restricted to subsets of the population. If only the utility of voluntary participants is considered, some position about the utility of nonparticipants must be taken, and the feedback between participants and nonparticipants must be explicitly modeled. When individual utility only depends on individual consumption, the criterion is well defined.

TABLE 2

ESTIMATED PARAMETERS OF THE IMPACT DISTRIBUTION PERFECT POSITIVE DEPENDENCE, POSITIVE DEPENDENCE WITH $\tau = 0.95$, INDEPENDENCE AND PERFECT NEGATIVE DEPENDENCE CASES

National JTPA Study 18 Month Impact Sample
Adult Females

Statistic	Perfect Positive Dependence ($\tau = 1.0$)	Positive Dependence with $\tau = 0.95$	Independence of Y^1 and Y^0 ($\tau = 0.0$)	Perfect Negative Dependence ($\tau = -1.0$)
5th Percentile	0.00 (47.50)	0.00 (360.18)	-18098.50 (630.73)	-22350.00 (547.17)
25th Percentile	572.00 (232.90)	125.50 (124.60)	-6043.00 (300.47)	-11755.00 (411.83)
50th Percentile	864.00 (269.26)	616.00 (280.19)	0.00 (163.17)	580.00 (389.51)
75th Percentile	966.00 (305.74)	867.00 (272.60)	7388.50 (263.25)	12791.00 (253.18)
95th Percentile	2003.00 (543.03)	1415.50 (391.51)	19413.25 (423.63)	23351.00 (341.41)
Percent Positive	100.00 (1.60)	96.00 (3.88)	54.00 (1.11)	52.00 (0.81)
Impact Std Dev	1857.75 (480.17)	6005.96 (776.14)	12879.21 (259.24)	16432.43 (265.88)
Outcome Correlation	0.9903 (0.0048)	0.7885 (0.0402)	-0.0147 (0.0106)	-0.6592 (0.0184)

TABLE 3

RANDOM COEFFICIENT AND DECONVOLUTION ESTIMATES IMPACT ON EARNINGS IN THE 18 MONTHS AFTER RANDOM ASSIGNMENT

National JTPA Study 18 Month Impact Sample
Adult Females

Analysis	Estimated Mean Impact	Estimated Impact Std Dev	Estimated Percent Positive
Random coefficient model	601.74 (201.63)	2271.00 (1812.90)	60.45
Deconvolution	614.00	1675.00	56.35

TABLE 4

TESTS OF SECOND ORDER STOCHASTIC DOMINANCE OF EXPERIMENTAL TREATMENT GROUP OVER EXPERIMENTAL CONTROL GROUP EARNINGS IN THE 18 MONTHS AFTER RANDOM ASSIGNMENT

National JTPA Study 18 Month Impact Sample

Earnings Value (α)	Adult Males	Adult Females	Male Youth	Female Youth
2,500	0.8836	1.0296	-0.3357	0.6674
	(0.3162)	(0.2978)	(0.4250)	(0.5094)
	[0.0052]	[0.0005]	[0.4296]	[0.1901]
5,000	1.8067	1.9343	-1.0482	0.7137
	(0.6582)	(0.5955)	(0.9344)	(1.0022)
	[0.0061]	[0.0012]	[0.2620]	[0.4764]
7,500	2.3903	2.7811	-1.8742	0.4428
	(0.9983)	(0.8933)	(1.4610)	(1.4507)
	[0.0166]	[0.0019]	[0.1995]	[0.7602]
10,000	2.9839	3.7315	-2.8489	0.1308
	(1.3334)	(1.1504)	(1.9790)	(1.8486)
	[0.0252]	[0.0012]	[0.1500]	[0.9436]
15,000	4.0191	5.2659	-4.0631	-0.2717
	(1.9826)	(1.5768)	(2.8333)	(2.4032)
	[0.0435]	[0.0008]	[0.1516]	[0.9100]
20,000	4.4428	6.2660	-5.8554	-0.4484
	(2.5434)	(1.8551)	(3.5386)	(2.1750)
	[0.0807]	[0.0007]	[0.0980]	[0.8688]
25,000	4.6171	7.0279	-6.3804	-0.4503
	(2.9192)	(1.9980)	(4.0905)	(2.8641)
	[0.1137]	[0.0004]	[0.1188]	[0.8751]

TABLE 5

SELF-ASSESSMENTS OF JTPA IMPACT EXPERIMENTAL TREATMENT GROUP

National JTPA Study 18 Month Impact Sample

	Adult Males	Adult Females	Male Youth	Female Youth
Full Sample Percentages				
Percent who self-report participating:	61.63 (0.81)	68.10 (0.68)	62.62 (1.29)	66.29 (1.09)
Percent of self-reported participants with a positive self-assessment:	62.46 (1.04)	65.21 (0.85)	67.16 (1.59)	71.73 (1.29)
Overall percent with positive self-assessments:	38.49 (0.81)	44.41 (0.73)	42.06 (1.32)	47.55 (1.16)
Percent of Self-Reported Participants with a Positive Self-Assessment by Primary Treatment Received				
None (dropouts)	48.89 (2.07)	51.44 (1.85)	58.90 (3.33)	61.56 (2.79)
Classroom training in occupational skills	74.10 (2.15)	73.47 (1.36)	72.73 (3.60)	75.28 (2.30)
On-the-job training at private firm	75.13 (2.18)	78.90 (2.14)	71.00 (4.56)	75.00 (4.04)
Job search assistance	59.57 (2.27)	59.80 (2.18)	68.09 (3.94)	68.94 (4.04)
Basic education	62.96 (4.67)	56.55 (3.84)	70.97 (4.09)	78.44 (3.19)
Work experience	66.67 (9.83)	68.75 (5.84)	82.76 (7.14)	73.17 (7.01)
Other	58.47 (3.65)	66.40 (2.98)	62.50 (4.77)	77.98 (3.99)

Table 6

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

	Description of Analysis	Evidence of Heterogeneity?	Standard Deviation of Impacts	Evidence on Voting Criterion	Dependence Between Y^1 and Y^0
Frechet Bounds	Statistical bounds on the joint distribution of outcomes, $F(y^0, y^1 \mid D = 1)$, and on super- and sub-additive functions of the joint distribution.[1] See equation (17) in the text.	Yes, the impact standard deviation is bounded away from zero.[2]	Bounded between \$675 and \$14969.[2]	The bounds do not apply to the indicator function $1(Y^1 \geq Y^0)$ as this function is not super- or sub-additive. Thus, the voting criterion cannot be bounded.[2]	Product-moment correlation ρ between Y^1 and Y^0 bounded between -0.760 and 0.998.
Perfect Positive Percentile Dependence	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.	Yes, the impacts vary between \$0 and \$3250.[2]	\$1857.[2]	100 percent of participants benefit or are indifferent.[2]	Product-moment correlation $\rho = 0.9903$ and Kendall's rank correlation is fixed at 1.00. Both are calculated using the percentiles of the two distributions.[2]
Perfect Negative Percentile Dependence	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.	Yes, the impacts vary between -\$48606 and \$34102.[2]	\$16432.[2]	52 percent positive.[2]	Product-moment correlation $\rho = -0.6592$ and Kendall's rank correlation is fixed at -1.00. Both are calculated using the percentiles of the two distributions.[2]
Positive Percentile Dependence with Rank Correlation $\tau = 0.95$	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.	Yes, the average minimum is -\$14504 and the average maximum is \$48544.[2]	Average standard deviation of \$1857 (with standard deviation of \$480).[2,3]	Average of 93 percent positive (with standard deviation of 3.88).[2,3]	Average product-moment correlation ρ of 0.7885 (with a standard deviation of 0.0402). Kendall's rank correlation τ fixed at 0.95. Both are calculated using the percentiles of the two distributions.[2,3]
Independence of Percentiles of Y^1 and Y^0, Which Implies a Percentile Rank Correlation τ of 0.0	Assumes that the percentiles of Y^1 and Y^0 given $D = 1$ have a rank correlation τ of 0.0, which is implied by independence between them. Estimates are based on a random sample of 50 such permutations.	Yes, the average minimum is -\$44175 while the average maximum is \$60599.[2]	Average standard deviation of \$12879 (with standard deviation of \$259).[2,3]	Average of 54 percent positive (with standard deviation of 1.11).[2,3]	Average product-moment ρ of -0.0147 (with standard deviation of 0.0106). Kendall's rank correlation τ fixed at 0.0. Both are calculated using the percentiles of the two distributions.[2,3]

Table 6 (cont.)
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

	Description of Analysis	Evidence of Heterogeneity?	Standard Deviation of Impacts	Evidence on Voting Criterion	Dependence Between Y^1 and Y^0
Random Coefficient Model	Assumes that $\Delta \perp\!\!\!\perp Y_0 \mid D = 1$.	Yes, see Figure 2.[4]	Standard deviation is \$2271.[4]	If the random coefficient Δ is assumed to be normally distributed then 60.45 percent have positive impacts.[4]	The product-moment correlation $\rho = 0.9595$.[4]
Deconvolution	Assumes that $\Delta \perp\!\!\!\perp Y_0 \mid D = 1$.	Yes, see Figure 2.[4]	Standard deviation is \$1675.[4]	56.35 percent positive.[4]	The product-moment correlation $\rho = 0.9774$.[4]
Self-assessments	Ex post self-evaluations by participants based on a survey question regarding whether or not the program provided a benefit.	Yes, some participants reported a benefit and others did not.	N.A.[5]	Varies from a low of 39.49 percent positive for adult men to a high of 47.55 percent positive for female youth.	N.A.[5]
Dropouts	Attrition decisions after application and acceptance into the program.	Yes. There are non-zero attrition rates and the evidence on the discount rates required to jusitfy a common coefficient model suggests that this model is false.	N.A.[5]	Dropping out ranges from a low of 34.83 percent for male youth to a high of 40.92 percent for adult males.	N.A.[5]

[1] A function $k(x, y)$ is superadditive if $x > x'$ and $y > y'$ implies that $k(x, y) + k(x', y') > k(x, y') + k(x', y)$. Subadditivity reverses the inequality.
[2] Results are for adult women only. Similar results are obtained for adult men and for male and female youth.
[3] The standard deviation is calculated over the random sample of 50 permutations with the indicated value of τ.
[4] Results are for adult women only. For the remaining demographic groups $Var\left(Y^1\right) < Var\left(Y^0\right)$ which indicates that neither the random coefficient model nor deconvolution is appropriate.
[5] N.A. = not applicable.

Table 7

Closed Economy Effects of Alternative Tax Proposals
General Equilibrium (Steady State) and Partial Equilibrium Effects[§]
Percentage Difference from Progressive Case[†]

	Flat Tax[‡]		Flat Cons. Tax[‡]	
	PE	GE	PE	GE
After Tax Interest Rate	0.00	1.96	17.65	3.31
Skill Price College HC	0.00	-1.31	0.00	3.38
Skill Price HS HC	0.00	-0.01	0.00	4.65
Stock of Physical Capital	-15.07	-0.79	86.50	19.55
Stock of College HC	22.41	2.82	-15.77	1.85
Stock of HS HC	-9.94	0.90	1.88	0.08
Stock of College HC per College Graduate	3.04	2.55	-4.08	1.72
Stock of HS HC per HS Graduate	1.84	1.07	-5.23	0.16
Aggregate Output	-0.09	1.15	15.76	4.98
Aggregate Consumption	-0.08	0.16	7.60	3.66
Mean Wage College	3.39	2.60	0.12	6.96
Mean Wage HS	2.44	2.44	0.25	6.82
Standard Deviation Log Wage	4.09	1.56	-1.94	0.69
College/HS Wage Premium at 10 Yrs Exp*	1.92	-0.45	3.10	0.18
Fraction attending college	18.79	0.26	-12.18	-1.92
Type 1: Fraction Attending College	50.29	-1.25	-42.57	2.14
Type 2: Fraction Attending College	28.50	-5.89	-15.60	-7.88
Type 3: Fraction Attending College	14.13	-6.93	-5.20	-9.56
Type 4: Fraction Attending College	15.27	6.13	-11.77	7.50
Type 1: College HC Gain First 10 Years**	5.81	3.12	-7.53	1.51
Type 2: College HC Gain First 10 Years**	5.33	2.86	-6.84	1.38
Type 3: College HC Gain First 10 Years**	5.60	3.10	-6.70	1.61
Type 4: College HC Gain First 10 Years**	6.85	4.17	-6.41	2.56
Type 1: HS HC Gain First 10 Years**	3.42	1.06	-7.79	-0.34
Type 2: HS HC Gain First 10 Years**	4.49	1.97	-7.60	0.46
Type 3: HS HC Gain First 10 Years**	5.36	2.67	-7.62	1.06
Type 4: HS HC Gain First 10 Years**	5.29	2.55	-7.95	0.92

Table continues on next page

Notes:

§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 labor 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 labor income is flat as is calculated to balance the budget in the new GE steady state. This yields a tax rate on labor 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 differences 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 upon entering the labor force.

Source: Heckman, Lochner and Taber, 1999.

Table 8A
Votes for Policy Reform in the Initial State and
Outcomes in Final Steady State

	Movement to Flat Income Tax	Movement to Flat Consumption Tax
% in Favor in Initial State	43%	52%
Final Steady State Utility Gain		
High School Ability 1:	-0.61	0.27
High School Ability 2:	-0.20	0.71
High School Ability 3:	0.09	0.93
High School Ability 4:	-0.13	0.78
College Ability 1:	-0.53	0.35
College Ability 2:	-0.32	0.58
College Ability 3:	-0.18	0.72
College Ability 4:	0.23	1.16
Ability 1	-0.57	0.30
Ability 2	-0.28	0.64
Ability 3	-0.03	0.85
Ability 4	0.11	1.05

Table 8B
Votes for Policy Reform in Initial State
with Introduction of Technical Change
and Outcomes in Final Steady State

	Movement to Flat Income Tax	Movement to Flat Consumption Tax
% in Favor in Initial State	65%	66%
	Final Steady State Utility Gain	
High School Ability 1:	0.33	1.07
High School Ability 2:	0.76	1.54
High School Ability 3:	0.99	1.78
High School Ability 4:	0.83	1.61
College Ability 1:	0.48	1.22
College Ability 2:	0.74	1.50
College Ability 3:	0.89	1.66
College Ability 4:	1.39	2.17
Ability 1	0.39	1.13
Ability 2	0.75	1.52
Ability 3	0.95	1.71
Ability 4	1.25	2.03

Table 9

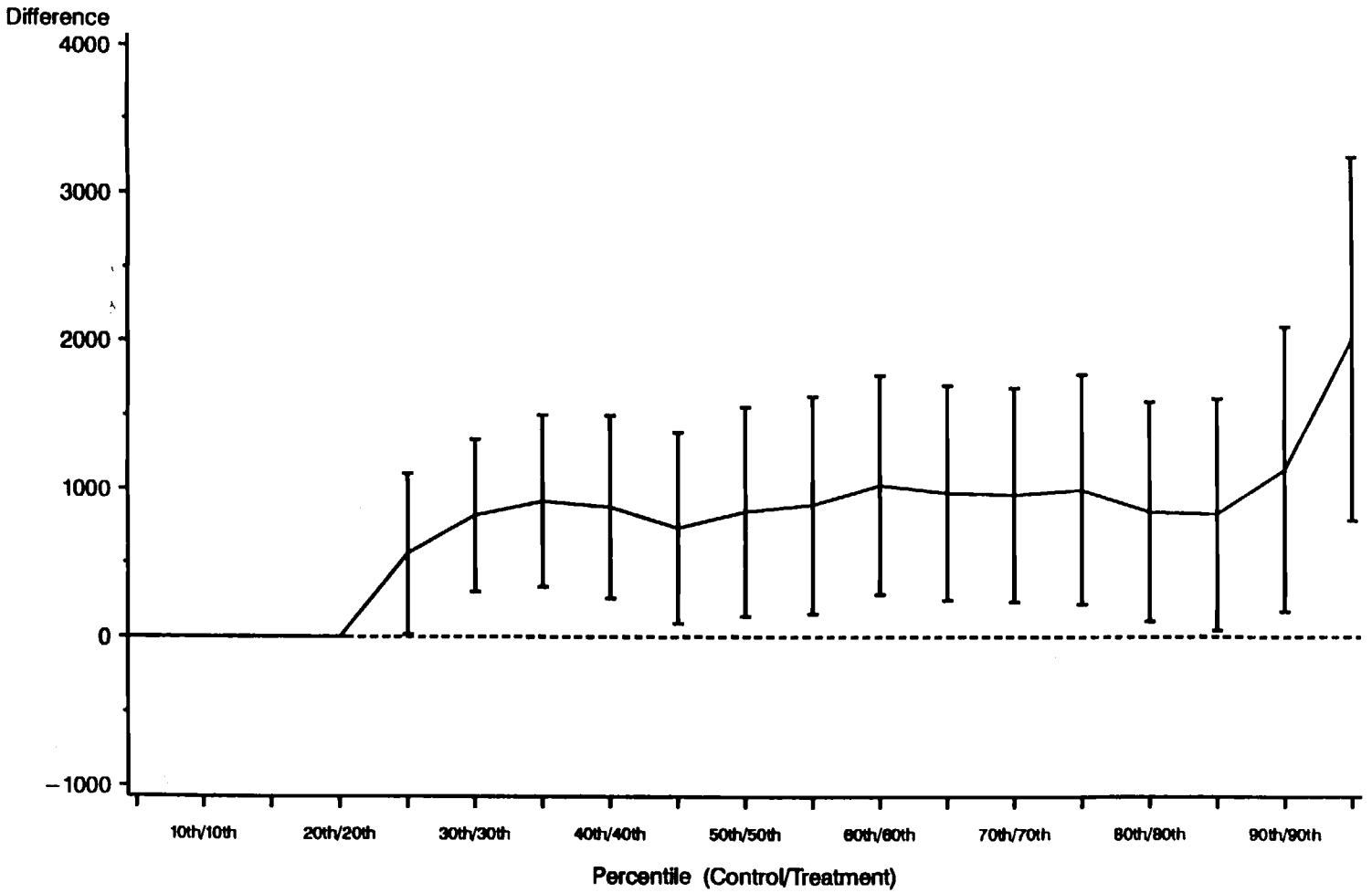
Simulated Effects of $5000 Tuition Subsidy on Different Groups
Steady State Changes in Present Value of Lifetime Wealth
(Thousands of 1995 Dollars)

Group(Proportion)†	After-Tax Earnings Using Base Tax‡ (1)	After-Tax Earnings‡ (2)	After-Tax Earnings Net of Tuition‡ (3)
High School-High School (0.5210)	17.520	6.849	6.849
High School-College (0.023)	9.757	-0.372	14.669
College-High School (0.0003)	-37.874	-49.528	-45.408
College-College (0.447)	1.574	-10.233	8.412
Ability Quartile 1			
High School-High School (0.844)	14.696	5.673	5.673
High School-College (0.045)	30.587	21.043	36.179
College-High School (0.000)	0.000	0.000	0.000
College-College (0.111)	1.273	-8.271	10.353
Ability Quartile 2			
High School-High School (0.689)	18.571	7.269	7.269
High School-College (0.033)	-5.356	-15.874	-0.841
College-High School (0.000)	0.000	0.000	0.000
College-College (0.277)	1.308	-9.210	9.428
Ability Quartile 3			
High School-High School (0.446)	20.691	8.181	8.181
High School-College (0.014)	-22.046	-33.156	-18.409
College-High School (0.000)	0.000	0.000	0.000
College-College (0.541)	1.4010	-9.6910	8.946
Ability Quartile 4			
High School-High School (0.139)	19.286	7.633	7.633
High School-College (0.000)	0.000	0.000	0.000
College-High School (0.001)	-37.874	-49.528	-45.408
College-College (0.859)	1.802	-11.152	7.498

(†) The groups denote counterfactual groups. For example, the High School-High School group consists of individuals who would not attend college in either steady state, and the High School-College group would not attend college in the first steady state, but would in the second, etc.

(‡) Column (1) reports the after-tax present value of earnings in thousands of dollars discounted using the after-tax interest rate where the tax rate used for the second steady state is the base tax rate. Column (1) reports just the effect on earnings, column (2) adds the effect of taxes, column (3) adds the the effect of tuition subsidies.

Figure 1
Treatment – Control Differences at Percentiles of the 18 Month Earnings Distribution
Perfect Positive Dependence Case
Adult Females



1. Sample consists of ABT's experimental 18–month study sample.
2. ABT imputed values were used in place of outlying values.
3. Standard errors for the quantiles are obtained using methods described in Csorgo (1993).

Figure 2 - Impact Densities Under Alternative Identifying Assumptions
Adult Females

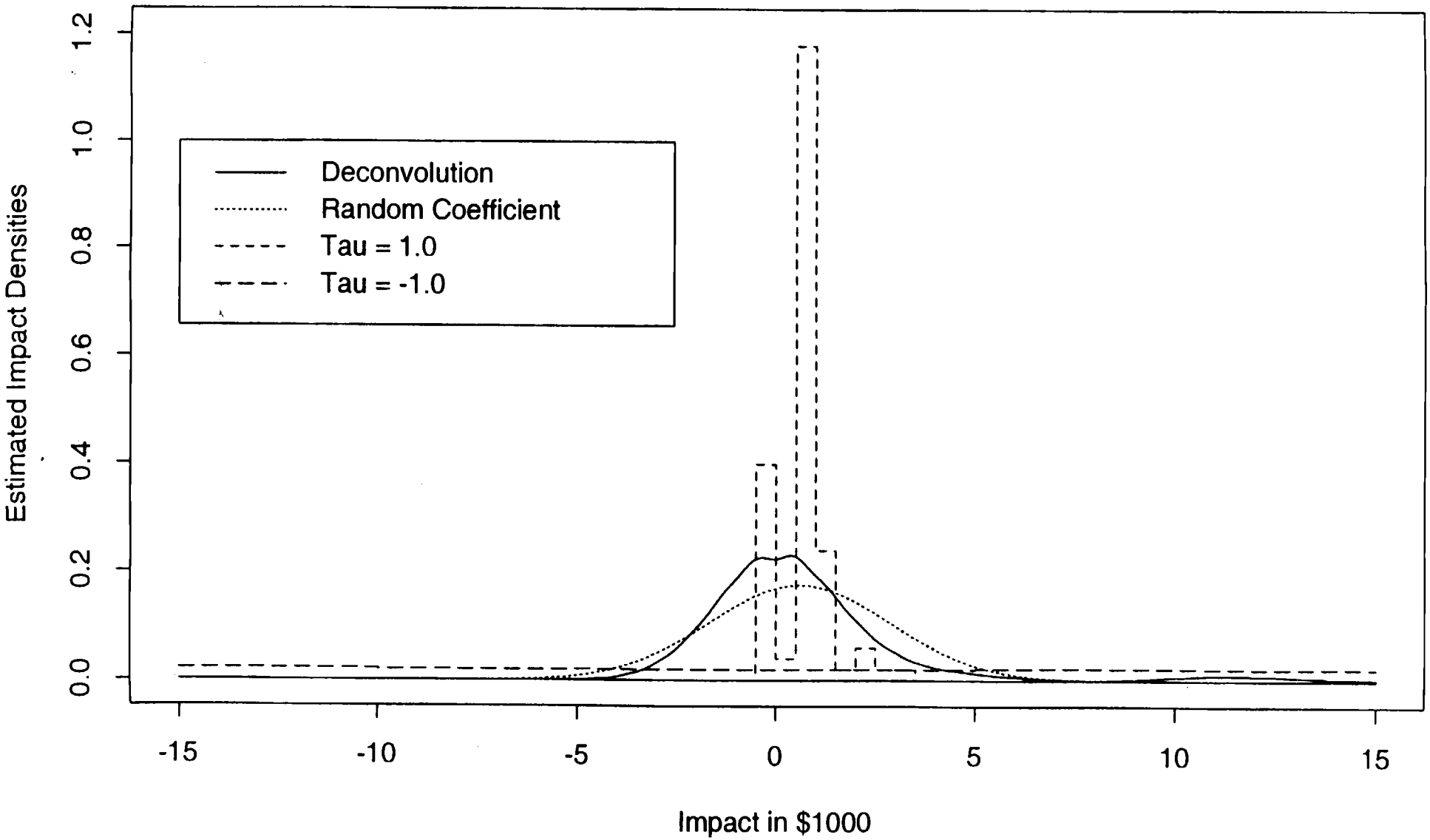


Figure 3A
Changes in Utility from the Reform in the Current Generation: Flat Tax

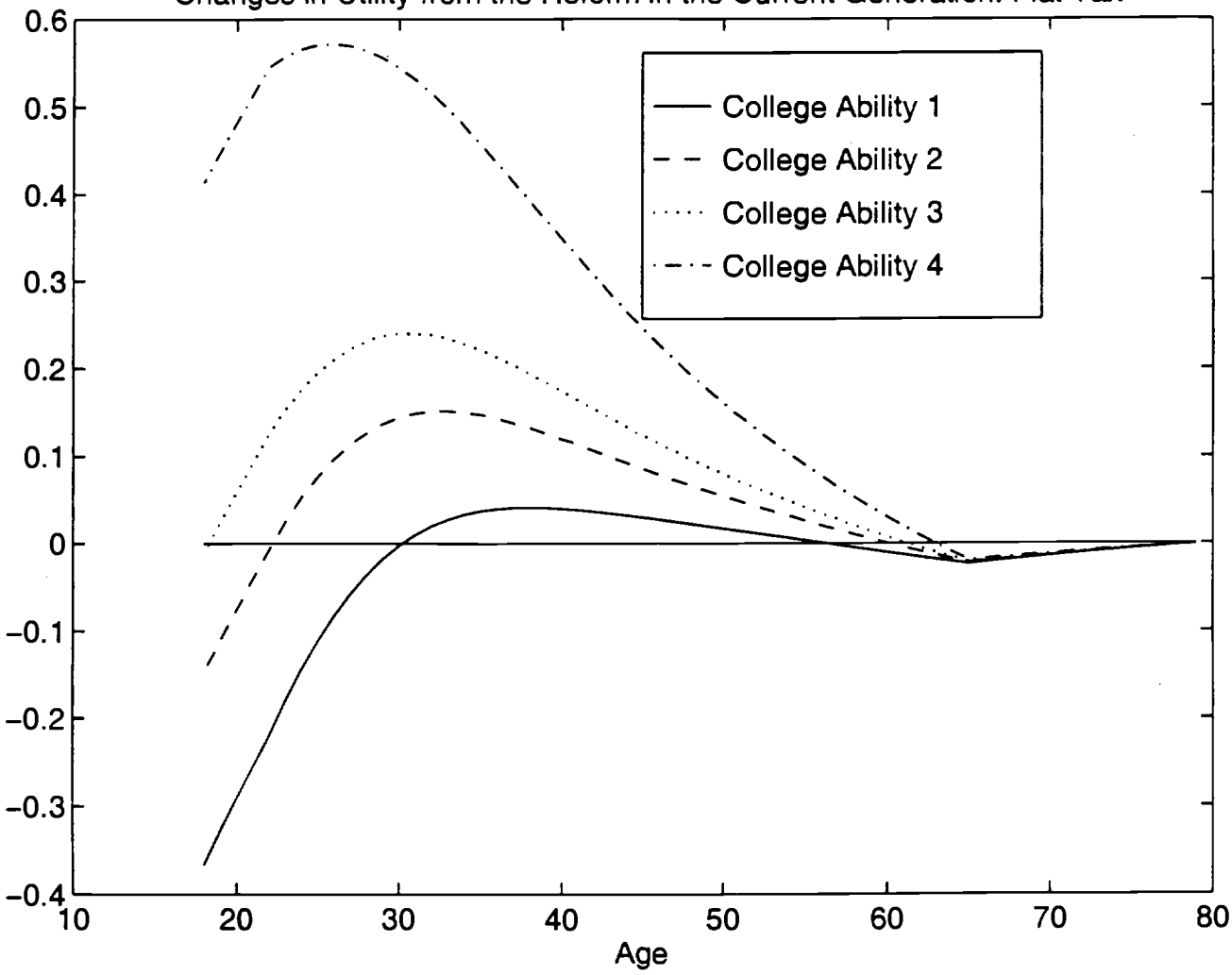


Figure 3B
Changes in Utility from the Reform in the Current Generation: Flat Tax

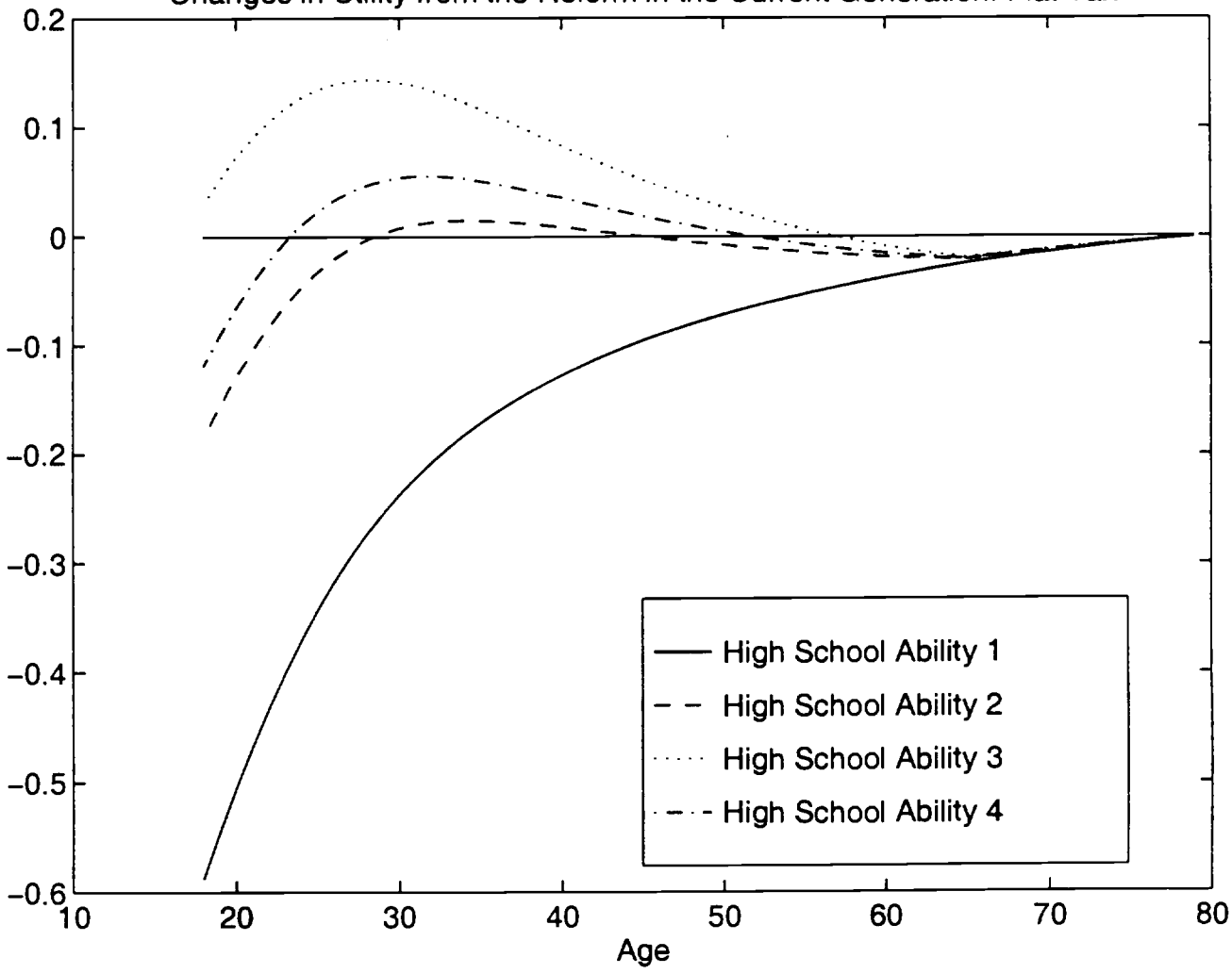


Figure 4A
Changes in Utility from the Reform in the Current Generation: Consumption Tax

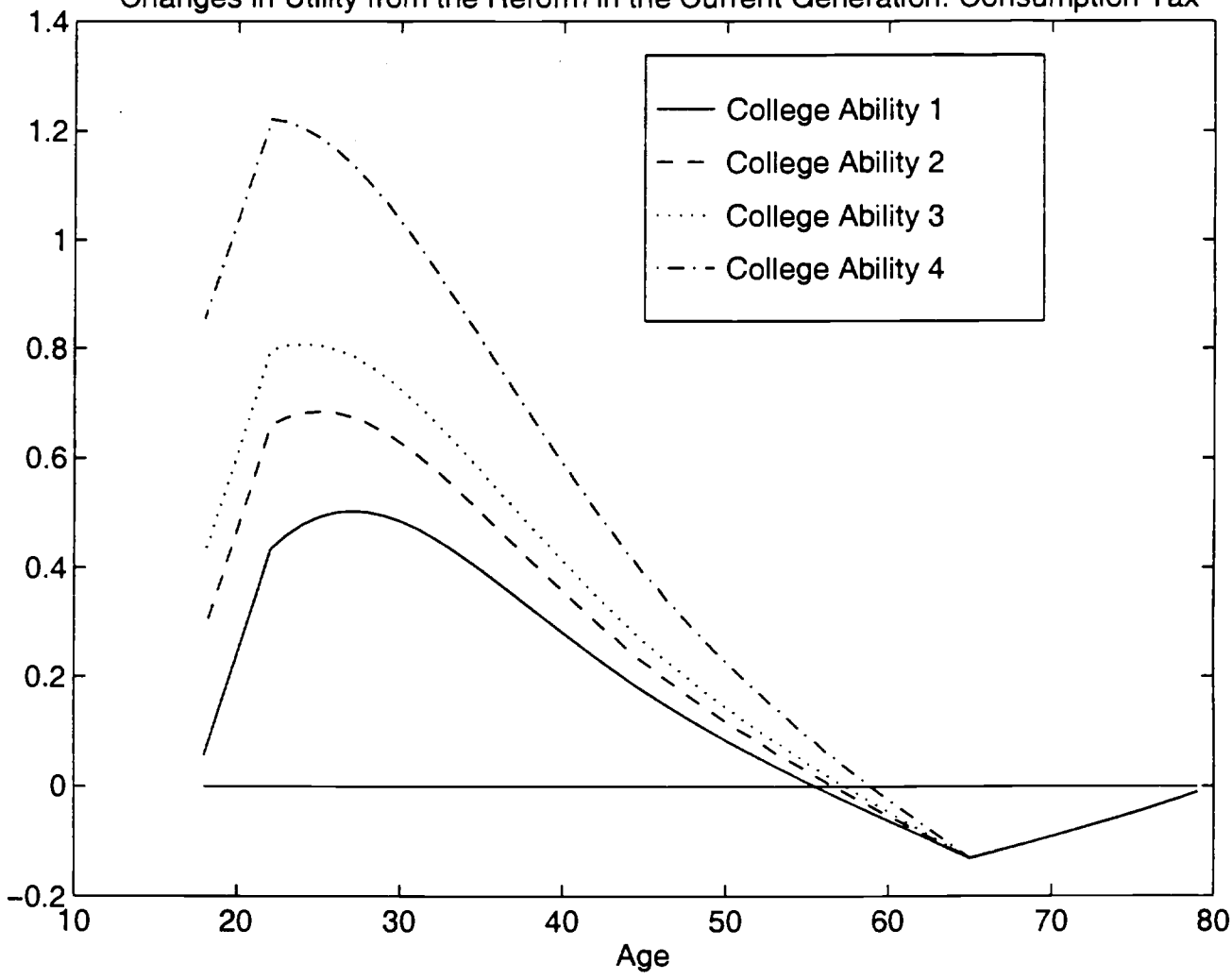


Figure 4B
Changes in Utility from the Reform in the Current Generation: Consumption Tax

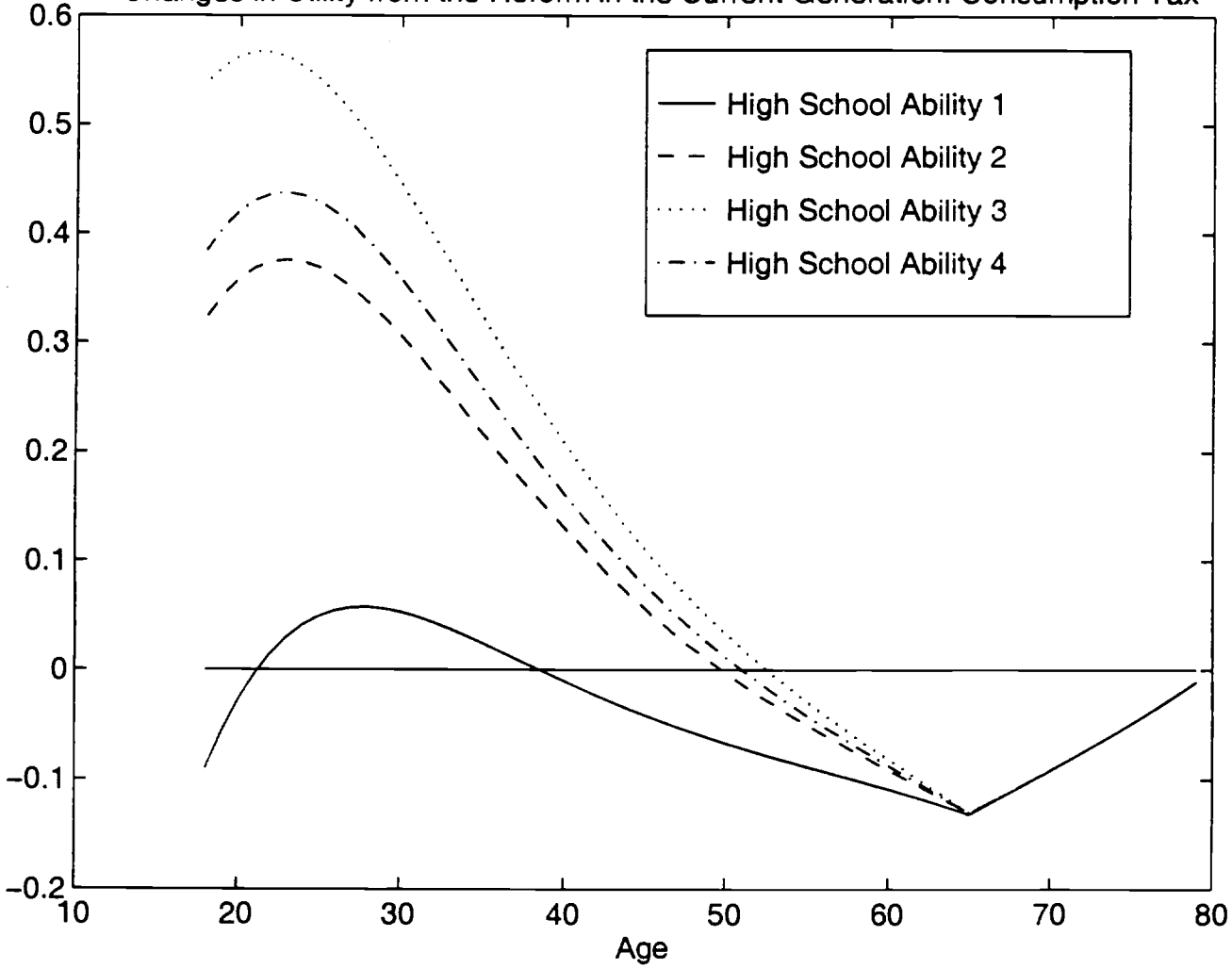


Fig 5A:Utility Changes from Reform in a Generation Experiencing Tech Change:Inc Tax

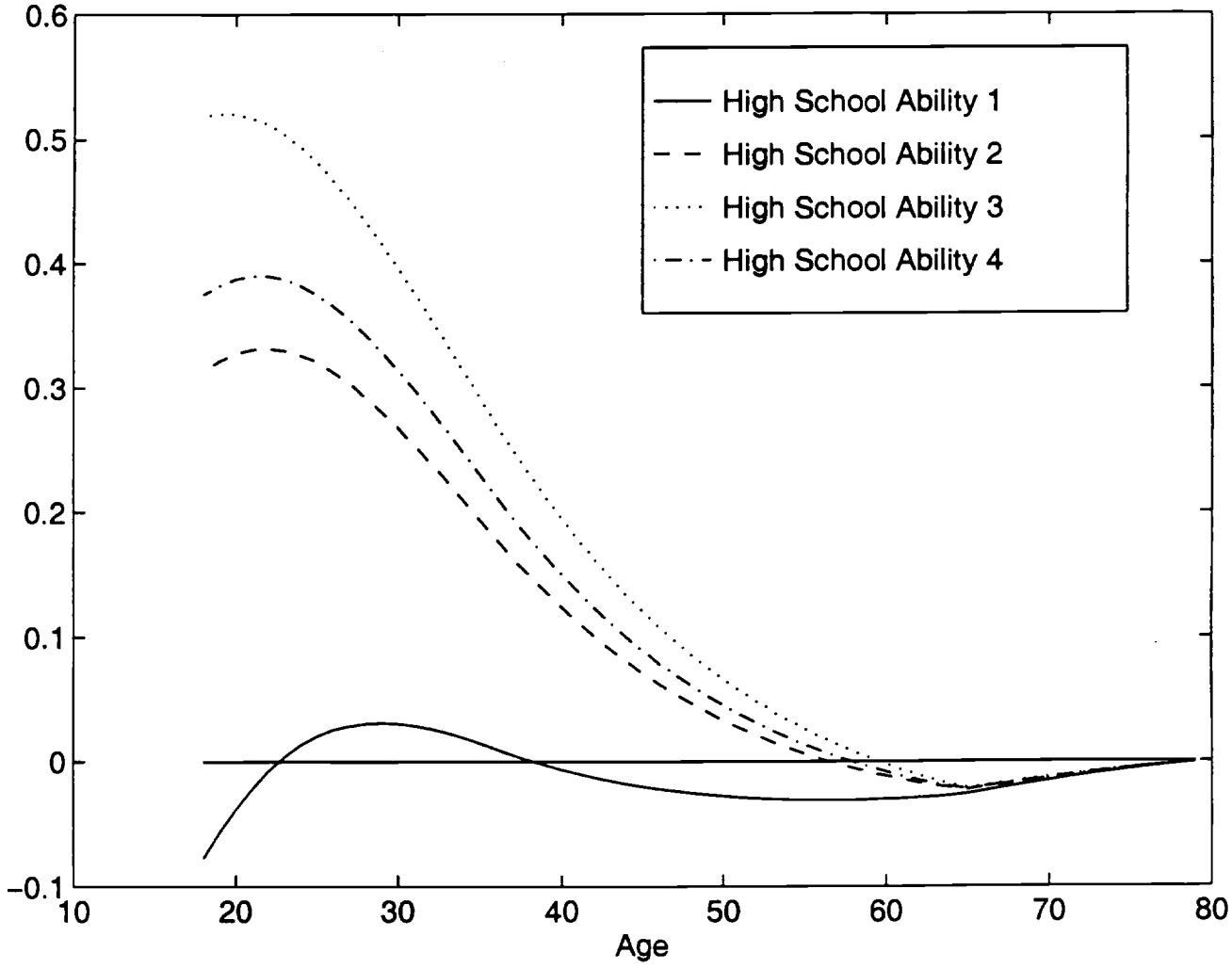


Fig 5B:Utility Changes from Reform in a Generation Experiencing Tech Change:Inc Tax

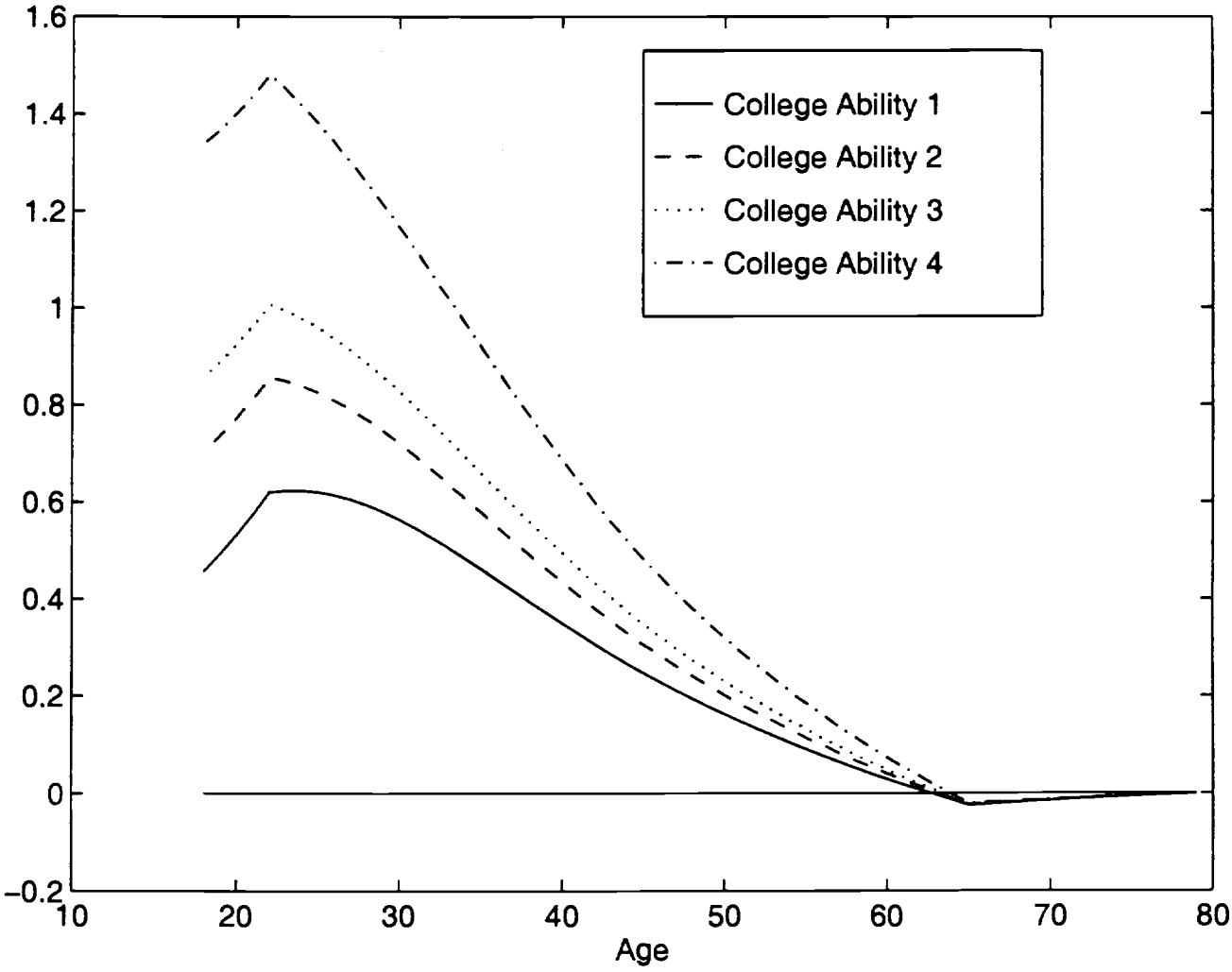


Fig 6A:Utility Changes from Reform in a Generation Experiencing Tech Change:Cons Tax

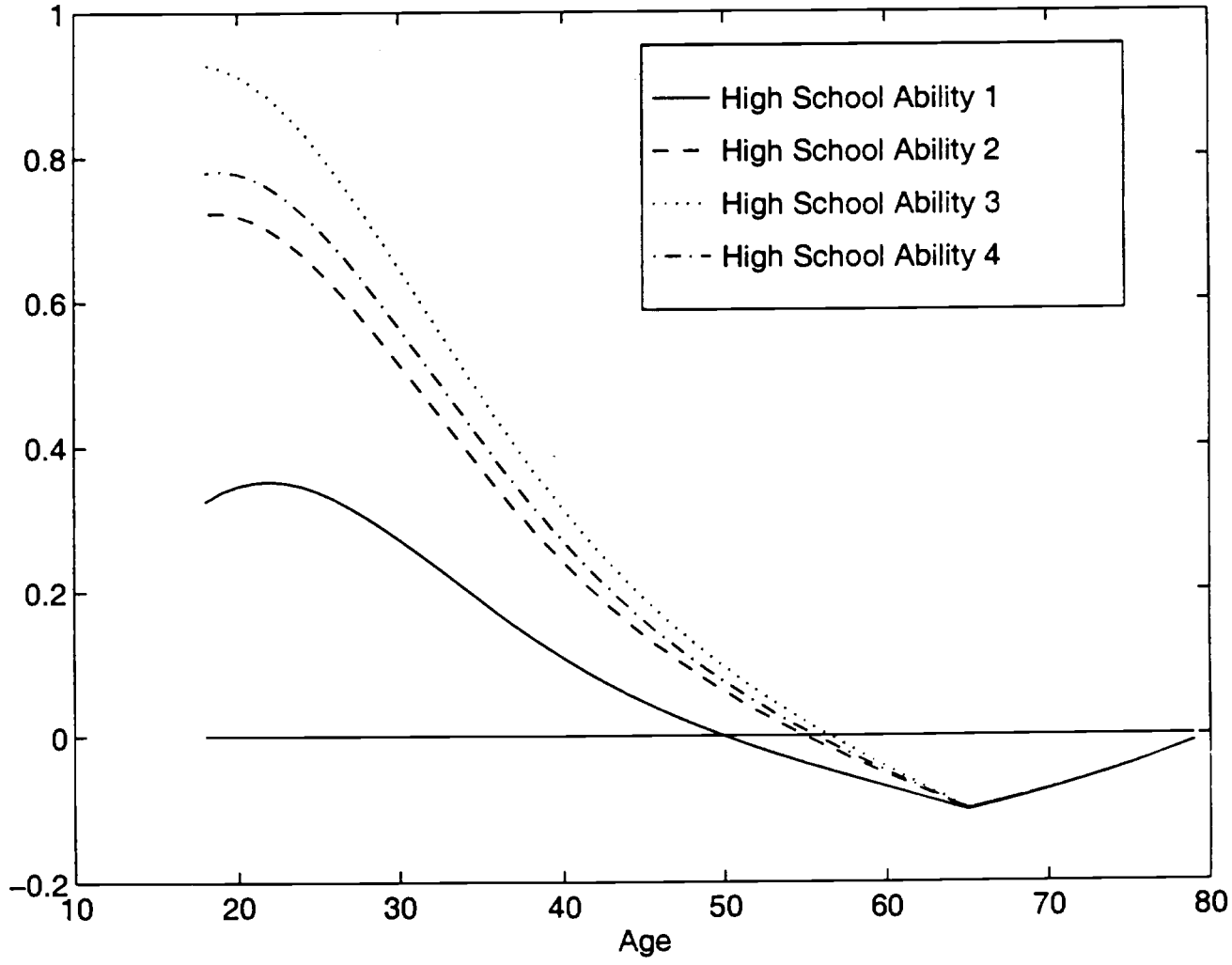


Fig 6B:Utility Changes from Reform in a Generation Experiencing Tech Change:Cons Tax