Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts

James J. Heckman; Jeffrey Smith; Nancy Clements

The Review of Economic Studies, Volume 64, Issue 4, Special Issue: Evaluation of Training and Other Social Programmes (Oct., 1997), 487-535.

Stable URL:
http://links.jstor.org/sici?sici=0034-6527%28199710%2964%3A4%3C487%3AMTMOOP%3E2.0.CO%3B2-0

Review of Economic Studies (1997) **64**, 487–535
0034-6527/97/00240487$02.00

Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts

JAMES J. HECKMAN
University of Chicago

JEFFREY SMITH
University of Western Ontario

with the assistance of
NANCY CLEMENTS
University of Chicago

First version received November 1993; *final version received May* 1997 (*Eds.*)

The conventional approach to social programme evaluation focuses on estimating mean impacts of programmes. Yet many interesting questions regarding the political economy of programmes, the distribution of programme benefits and the option values conferred on programme participants require knowledge of the distribution of impacts, or features of it. This paper presents evidence that heterogeneity in response to programmes is empirically important and that classical probability inequalities are not very informative in producing estimates or bounds on the distribution of programme impacts. We explore two methods for supplementing the information in these inequalities based on assumptions about participant decision-making processes and about the strength in dependence between outcomes in the participation and non-participation states. Dependence is produced as a consequence of rational choice by participants. We test for stochastic rationality among programme participants and present and implement methods for estimating the option values of social programmes.

And the Lord said, Because the cry of Sodom and Gomorrah is great, and because their sin is very grievous; I will go down now, and see whether they have done altogether according to the cry of it, which is come unto me; and if not, I will know . . . And Abraham drew near and said, Wilt thou also destroy the righteous with the wicked? Peradventure there be fifty righteous within the city; wilt thou also destroy and not spare the place for the fifty righteous that are therein? . . . And the Lord said, If I find in Sodom fifty righteous within the city, then I will spare all the place for their sakes.

Genesis 18: Verses 20–26, King James Version

1. INTRODUCTION

Most evaluations of social programmes focus exclusively on mean impacts. Yet, as the passage from Genesis reveals, features other than the mean are often of interest. Just as Abraham convinced the Lord to spare Sodom and Gomorrah if he could find 50 righteous persons living there, so many persons would judge programmes to be successful if enough persons, or enough of the right kinds of persons, reaped benefits from them even if the average participant did not.

The case for using the mean impact to evaluate a programme rests on two key assumptions: (a) that increases in total output increase welfare; and (b) that undesirable distributional aspects of programmes are either unimportant or are offset by transfers governed by a social welfare function, or its counterpart for families or groups. Both of these assumptions are strong. Many programmes produce outputs that cannot easily be redistributed (e.g. vaccinations or other nontransferable payments in kind). Programme outputs cannot always be valued and summed to produce a measure of total welfare. Appeal to a mythical social welfare function begs fundamental questions of political economy. The distribution of the benefits (and costs) from a programme determines the support for a programme if voters are self-interested or if they are altruistic. In median voter models, the mean is irrelevant unless it coincides with the median. An altruistic voter may wish to see the lot of the worst-off advanced if he adopts Rawls' (1971) maximin criterion for social justice.

Measuring the distribution of impacts across persons is an intrinsically difficult problem. The fundamental aspect of the programme evaluation problem is that one cannot simultaneously observe the same person in a programme and out of it. Thus it is not possible to determine the programme impact for any person. However, one can construct the outcome distribution for participants and the outcome distribution for nonparticipants. If the outcome distribution for nonparticipants coincides with what participants would have experienced had they not participated, or can be adjusted to do so, then the difference in the means of the participant and nonparticipant distributions is the mean impact of the programme. The distribution of impacts, or even its median, is more difficult to obtain. From the two marginal distributions for participants and nonparticipants, it is generally not possible to estimate the joint distribution of outcomes and so it is generally not possible to estimate the distribution of impacts or its median.

One important special case where it is possible dominates the textbook literature in econometrics. This special case is the "common effect" model where the programme is assumed to have the same impact on everyone (or everyone with the same observed characteristics). In this case, the impact distribution is degenerate and is concentrated on the mean impact.[1]

The conventional assumption of identical programme impacts across persons, while convenient, is implausible, and we present evidence against it for a prototypical job training programme. How then can one identify the distribution of programme impacts in the more general case? One way to recover features of joint distributions from marginal distributions is to use the classical probability inequalities of Hoeffding (1940) and Fréchet (1951) that bound joint distributions from known marginals.[2] For our prototypical programme, these inequalities are not very informative in the sense that they do not produce precise estimates or bounds on the quantiles or the other features of the distribution of impacts.

1. The centrality of the common effect model in conventional econometric policy evaluation analysis is stressed in Heckman and Robb (1985, 1986) and Heckman (1992).

2. Lavine, Wasserman and Wolpert (1991) apply these inequalities in a Bayesian framework to study clinical trials and produce negative results similar to the ones reported in this paper.

The failure of a purely statistical approach to determine the distribution of programme impacts leads us to supplement the classical inequalities in two related ways. First, we investigate the value of assumptions about dependence between potential outcomes in the treated and untreated states. Second, we investigate strategies that exploit assumptions about the decision rules governing participation in the programme. These two approaches are linked. If agents follow the decision rules we investigate, dependence is induced in the potential outcomes across the treated and untreated states for programme participants. Evidence on programme participation decision rules justifies certain assumptions about dependence in outcomes that can be used to construct or bound distributions of impacts using the observed marginal distributions of outcomes for participants and for nonparticipants.

Estimates of decision rules can be used to compare participant "subjective" assessments of outcomes with direct "objective" measures of outcomes to see if they agree. Both measures provide useful information for evaluating the welfare state (Heckman and Smith (1997)). If participant self-selection decisions are solely determined by programme outcomes, subjective and objective evaluations agree. In this case it turns out to be possible to extrapolate estimates of impacts obtained in one economic environment to other environments as we demonstrate below.

This paper proceeds in the following way. We present evaluation problems where the distribution of impacts is a central concern and show how the traditional common effect model readily answers those problems. We then consider the more general, and plausible, case where impacts are heterogeneous. We present several different definitions of the option value of a social programme.

We estimate impact distributions using data from a social experiment where there is no selection bias. (Heckman and Smith (1997) consider the more general case of inference from nonexperimental data.) We demonstrate two points using data from a prototypical job training programme. First, classical probability inequalities do not resolve much of the uncertainty about impact distributions. Second, the inequalities, together with our other findings, provide strong evidence of variability in programme impacts. Heterogeneity in the response to treatment is an important feature of the data.

We examine the strength of the dependence between outcomes in the potential outcome states required to produce plausible impact estimates. Very strong dependence is required to produce credible impact estimates in our data. We also present evidence on programme participation decision rules, and test to see if programme participation decisions satisfy stochastic rationality. Many identifying assumptions have testable implications which we examine in this paper.

We demonstrate how information about agent decision rules facilitates extrapolation of estimates of impacts obtained in one evaluation environment to other environments. We examine the benefits of randomization for models with different programme participation decision rules used by agents. We also estimate the option value conferred on persons eligible for social programmes. The paper concludes with a summary of the main results.

2. THE EVALUATION PROBLEM AND THE CRITERIA OF INTEREST IN EVALUATING SOCIAL PROGRAMMES

In the simplest version of the programme evaluation problem, there are two potential states of the world for each individual, (Y_0, Y_1). We treat Y_1 as the outcome obtained given participation in the programme being evaluated and Y_0 as the outcome in the benchmark state of non-participation. Let D denote participation, where $D=1$ if a person participates and $D=0$ otherwise. Coverage of the programme may be partial or universal.

For simplicity, we ignore general equilibrium effects. Our analysis closely approximates a full general equilibrium policy analysis if Y_0 is the outcome in the no-programme state. Heckman and Smith (1997) present precise conditions under which the partial equilibrium "no treatment" state approximates the general equilibrium no-programme state. In general, the nonparticipation outcome state for a given programme is not the same as the no-programme state for any person, so some approximation error is unavoidable in any partial equilibrium evaluation.

If analysts could observe (Y_0, Y_1) for everyone, there would be no evaluation problem. In that case, one could form the traditional measure of gain, $\Delta = Y_1 - Y_0$, for each individual and for various populations of interest, and use this measure to answer a variety of interesting questions. The conventional approach to programme evaluation focuses on estimating mean impacts. The mean that receives the most attention is $E(Y_1 - Y_0 | D=1)$, the effect of "treatment on the treated." This mean is useful in determining the gross gain from an existing programme.[3]

(*a*) *Criteria of interest besides the mean*

Many interesting evaluation questions require knowledge of features of the distribution of programme gains other than the mean. From the standpoint of a detached observer of a social programme, such as the "social planner" of welfare economics, who equates Y_0 with the no-programme state for an individual and hence takes the base state values as those that would prevail in the absence of the programme, it is of interest to know, *inter alia*,

(a) the proportion of people taking the programme who benefit from it,

$$\Pr(Y_1 > Y_0 | D=1) = \Pr(\Delta > 0 | D=1);$$

(b) the proportion of the total population that benefits from the programme,

$$\Pr(Y_1 > Y_0 | D=1) \Pr(D=1) = \Pr(\Delta > 0 | D=1) \Pr(D=1);$$

(c) selected quantiles of the impact distribution,

$$\inf_{\Delta} \{\Delta : F(\Delta | D=1) > q\}, \qquad \text{where } q \text{ is a quantile of the distribution;}$$

(d) the distribution of gains at selected base state values,

$$F(\Delta | D=1, Y_0 = y_0).$$

Each of these measures can be defined conditional on observed characteristics X. Measure (a) is of interest in determining how widely programme gains are distributed among participants. Detached observers with preferences over distributions of programme outcomes or an electorate in a democratic society would be unlikely to assign the same weight to two programmes with the same mean outcome, one of which produced favourable outcomes for only a few persons while the other distributed gains more broadly. This is especially true if programme benefits are not transferrable or if restrictions on feasible

3. Heckman (1997*b*) and Heckman and Smith (1997) discuss the relationship of this parameter to those of conventional cost-benefit analysis and establish conditions under which "treatment on the treated" estimates an economically interesting parameter that can be used in a meaningful cost-benefit analysis.

social redistributions prevent distributional objectives from being attained. It is also interesting to determine the proportion of participants who are harmed as a result of programme participation: $\Pr(Y_1 < Y_0 | D = 1)$. A negative mean impact might be acceptable to many observers if most participants gain from the programme.

Measure (b) is the proportion of the *entire population* that benefits from the programme, assuming that costs are broadly distributed and are not perceived to be related to the specific programme being evaluated.[4] If voters have correct expectations about the joint distribution of outcomes, it is of interest to politicians, and to students of political economy, to determine how widely programme benefits are distributed. At the same time, large programme gains received by a few persons may make it easier to organize interest groups in support of a programme than if the same gains are distributed more widely. In a study of the political economy of interest groups, it is useful to know which groups benefit from a programme and how widely distributed are the programme benefits. Criteria (c) and (d) reveal the distribution of impacts for participants and for subgroups of participants with particular outcomes in the nonparticipation state. These are of interest to persons interested in "social justice." All of these measures require knowledge of features of the joint distribution of outcomes for participants for their computation, and not just the mean.

The traditional literature on programme evaluation focuses on mean impacts, which can be used to measure the effect of a programme on total social output. The theme of that literature is that "a dollar is a dollar," regardless of who receives it. However, an emphasis on efficiency to the exclusion of distribution is not universally accepted in the economic literature on programme evaluation.[5] An emphasis on efficiency is premised on the transferability of outcomes among participants or on the assumption that distributional issues are either irrelevant or are settled by some external distribution mechanism using a family or social welfare function.

Outcomes from health interventions, educational subsidies and training programmes are not transferrable. Moreover, even if all programme outputs can be monetized, the assumption that a family or social welfare function automatically settles distributional questions in an optimal way is questionable. Many programmes designed to supply merit goods are properly evaluated by considering the incidence of their receipt and not the aggregate of the receipts.

Even if distributional issues are ignored and the criteria of cost-benefit analysis are accepted, conventional econometric evaluation estimators do not supply the information required to implement the criteria. Heckman (1997*b*) and Heckman and Smith (1997) demonstrate that the parameters that receive the most attention in the programme evaluation literature, including the mean impact of treatment on the treated, do not provide the information required to compute the gain in GDP that results from a programme. All ignore costs and few answer well-posed economic questions, except under special circumstances.

Social programmes confer options, and it is of interest to assess their option values. Persons offered a subsidized job may take it or opt for their best unsubsidized alternative. The option of having a subsidized alternative job may be worth something. The option may be conferred simply by eligibility or it may be conferred only on participants. If the programme creates an option only for participants, then prior to participating in it their only available option comes from distribution $F_0(y_0)$. Following or during participation

4. A more comprehensive analysis would include costs.
5. See, e.g. Dreze and Stern (1987) and Weisbrod (1968). Kaplow (1993) presents a dissenting view.

in the programme, the individual has a second option Z drawn from distribution $F_Z(\cdot)$. If both options are known prior to choosing between them, and agents are outcome maximizers, then the observed outcome Y_1 is the maximum of the two options, $Y_1 = \max(Y_0, Z)$. The option Z may be available only during the period of programme participation, as in a wage subsidy programme, or it may become a permanent feature of the choice set as when a marketable skill is acquired.[6] It is useful to distinguish the case where the programme offers a distribution F_Z from which new offers are received each period from the case where a permanent Z value is created. Much of the literature on programme evaluation implicitly equates Z with Y_1. This would follow under the assumption that treatment is an irreversible condition that supplants Y_0. Alternatively, this would follow under the assumption that $Z > Y_0$ for all draws of Z and Y_0. In either case, persons always choose Z over Y_0. In either case, $Y_1 = Z$ and the estimated distribution of Y_1 is equivalent to the estimated distribution of Z. For the general case it is useful to determine what a programme offers to potential participants, what the offer is worth to them, and to distinguish the offered option from the realized choice.

The expected value of having a new option Z in addition to Y_0 is

$$\text{(OP-1)} \qquad E(\max(Y_0, Z) \mid D=1) - E(Y_0 \mid D=1),$$

assuming that participants can pick freely between Y_0 and Z. This is the difference in expected outcomes between a two-option world and a one-option world, assuming that it is costless to choose between Y_0 and Z and both are known at the time the choice is made. Assuming that *participants* can choose between realized Z and Y_0 offers, social experiments estimate (OP-1). It is useful to distinguish the opportunities created from the programme, Z, from the options selected. The programme extends opportunities to participants. Providing a new opportunity that may be rejected may improve the average outcome among persons who choose Y_0 over Z in the sense that

$$E(Y_0 \mid Z < Y_0, D=1) - E(Y_0 \mid D=1) > 0,$$

even though the value of Y_0 received by all persons is assumed to be the same irrespective of whether or not they receive draws from the distribution F_Z.[7] This effect is a pure compositional phenomenon arising from self-selection.

If a programme gives participants a second distribution from which they receive a new draw each period, and if realizations of the pair (Y_0, Z) in each future period are statistically independently and identically distributed, then the addition to future wealth of having access to a second option in every period is

$$\text{(OP-2)} \qquad 1/r\,[E(\max(Y_0, Z) \mid D=1) - E(Y_0 \mid D=1)],$$

where r is the interest rate. If Z is available only for a limited time period, as would be the case for a job subsidy, (OP-1) is discounted over that period and expression (OP-2) should be appropriately modified to adjust for the finite life of the first term.[8]

6. Wage subsidy programmes may create lasting skills. See the evidence in Heckman, Lochner, Smith and Taber (1997).

7. In a general equilibrium setting, the existence of F_Z may alter F_0. Thus a new skill may affect the market for old skills. We abstract from these considerations in this paper.

8. Let $g(\tau) = 1 - e^{-r\tau}/r$, where τ is length of life, then (OP-2) is $g(\tau)E(\max(Y_0, Z) \mid D=1) - (1/r)\,E(Y_0 \mid D=1)$.

Returning to the case of a single draw, if the realizations (Y_0, Z) are not known at the time decisions to exercise the option are made, (OP-1) should be modified to

(OP-3) $$\max\,(E(Y_0 \mid D=1), E(Z \mid D=1)) - E(Y_0 \mid D=1).$$

A fourth definition of option value recognizes the value of having uncertainty resolved at the time decisions to choose between Z and Y_0 are made. That definition is

(OP-4) $E(\max\,(Z, Y_0) \mid D=1) - \max\,(E(Z \mid D=1), E(Y_0 \mid D=1)) = \text{(OP-1)} - \text{(OP-3)}.$

Option value (OP-1) is produced from experimental data. The empirical challenge is to recover F_Z or $E(Z \mid D=1)$ from the experimental data, a task attempted in Section 8.

(b) The evaluation problem and the conventional approach to solving it

The evaluation problem poses fundamental limitations on our ability to answer all of the questions posed in subsection (a). The problem arises because we do not observe (Y_0, Y_1) for everyone. Thus it is not possible to estimate the joint distribution of (Y_0, Y_1) or the distribution of gains directly. This problem is explicitly discussed by Fisher (1951), Cox (1958), Roy (1951) and others.[9] From ordinary non-experimental data on participants $(D=1)$ and non-participants $(D=0)$ we can determine the conditional outcome distributions:

$$F_1(y_1 \mid D=1) \quad \text{(participant outcomes)}, \tag{1a}$$

and

$$F_0(y_0 \mid D=0) \quad \text{(non-participant outcomes.} \tag{1b}$$

If programme coverage is universal, there is no information on (1b) because $D=1$ for everyone. Even if coverage is partial, we do not know Y_0 for participants or Y_1 for non-participants, so without additional information it is not possible to construct the counter-factual conditional distributions:

$$F_0(y_0 \mid D=1)$$

(what participant outcomes would have been had they not participated), (1c)

and

$$F_1(y_1 \mid D=0)$$

(what non-participant outcomes would have been had they participated). (1d)

Since we never simultaneously observe both the treated and untreated states either for participants or for nonparticipants, we also do not know the conditional joint outcome distributions:

$$F(y_1, y_0 \mid D=1), \tag{1e}$$

9. The recent statistical literature sometimes calls this the "Rubin" model. In biostatistics, it is called the model of competing risks. This model is known as the switching regression model in econometric theory, see, e.g. Quandt (1972) or Quandt (1988). It is known as the Roy model in labour economics after an early paper by Roy (1951), who expressed the mathematics of the Fisher model verbally and examined the consequences of self-selection on income inequality. See Heckman and Honoré (1990) for an exposition of the Roy model and proofs about its nonparametric identifiability.

and

$$F(y_1, y_0 | D=0). \tag{1f}$$

Unless participation in the programme is random with respect to outcomes, so that $F_0(y_0 | D=1) = F_0(y_0 | D=0)$, it is not possible to use the non-experimental data, (1a) and (1b), to estimate either the conventional parameter of interest, the mean impact of treatment on the treated, $E(Y_1 - Y_0 | D=1)$, or many of the other parameters of economic interest introduced in the preceding discussion. From the population means of programme participants and non-participants, we obtain

$$E(Y_1 | D=1) - E(Y_0 | D=0) = E(Y_1 - Y_0 | D=1) + \{E(Y_0 | D=1) - E(Y_0 | D=0)\}.$$

Only if there is no selection of participants on the basis of Y_0 will the *selection bias* term in braces be zero. Under ideal conditions, social experiments solve the selection bias problem by producing $F_0(y_0 | D=1)$.[10] They do not recover (1d), because one cannot force non-participants to participate. Social experiments unaided by additional assumptions do not recover (1e) and (1f) because one cannot observe both coordinates of (Y_0, Y_1) for anyone. Without invoking further assumptions, it is only possible to determine features of the joint distribution that depend solely on $F_1(y_1 | D=1)$ and $F_0(y_0 | D=1)$, even using experimental data.

One important feature that can be recovered from an ideal social experiment is the mean impact of treatment on the treated, $E(Y_1 - Y_0 | D=1)$. In contrast, medians or other quantiles of the impact distribution cannot be consistently estimated from the marginal distributions provided by experimental data (Heckman (1992)).

However, in the special case where $Y_1 - Y_0 = \alpha$, and α is a function of observed variables X, the distribution of gains is degenerate since everyone with the same X has the same gain. This is the "dummy endogenous variable" model of Heckman (1978). In this case, Heckman (1992) shows that ideal experiments recover the joint distribution of (Y_1, Y_0) since $F_0(y_0 | X) = F_1(y_0 + \alpha | X)$, and hence social experiments can be used to answer all of the evaluation questions.

Assuming no bias is induced by randomization, social experiments determine the mean impact of treatment on the treated.[11] However, the mean does not answer many interesting questions if persons respond differently to treatment.

3. UNCERTAINTY ABOUT IMPACT DISTRIBUTIONS

If the responses to a programme of all persons with identical observable characteristics are identical, the problem of evaluating it simplifies greatly. In this section, we present evidence that variability in impacts is an empirically important phenomenon. We first consider the case where outcomes are continuous random variables.

(*a*) *The continuous case*

Assume access to data from an ideal social experiment. For simplicity, assume samples of N individuals in the treatment state and N in the non-treatment state and that outcomes

10. See Heckman (1992), Heckman (1996) or Heckman and Smith (1993, 1995) for a statement of conditions under which experiments produce this distribution. Heckman, Hohmann, Khoo and Smith (1997) present evidence that questions those assumptions in the context of a prototypical job training programme.

11. This is true whether randomization is administered at the date of enrollment into the programme or at eligibility. See Heckman and Smith (1993) or Heckman (1996).

are continuously distributed. Ranking individuals in the order of their outcome values from the highest to the lowest, let $Y_j^{(i)}$ be the outcome for the i-th highest-ranked person in empirical distribution j. Ignoring all ties, we obtain two $N \times 1$ vectors of outcomes:

Treatment Outcome: $F_1(y_1 \mid D=1)$ **Non-Treatment Outcome:** $F_0(y_0 \mid D=1)$

$$\underset{\sim}{Y}_1 = \begin{pmatrix} Y_1^{(1)} \\ \vdots \\ Y_1^{(N)} \end{pmatrix} \qquad \underset{\sim}{Y}_0 = \begin{pmatrix} Y_0^{(1)} \\ \vdots \\ Y_0^{(N)} \end{pmatrix},$$

From an ideal social experiment, we can identify the marginal data distributions $F_1(y_1 \mid D=1)$ and $F_0(y_0 \mid D=1)$, but we do not know where person i in the treatment distribution would appear in the non-treatment distribution. Suppose that we have access to a random sample of outcomes with the empirical distributions close to the true distributions. Corresponding to the ranking of the sample non-treatment outcome distribution, there are $N!$ possible patterns of outcomes in the treatment outcome distribution. By considering all possible permutations, we can form a collection C of possible impact distributions, i.e. alternative distributions of

$$\underset{\sim}{\Delta}_l = \underset{\sim}{Y}_1 - \Pi_l \underset{\sim}{Y}_0, \qquad l = 1, \ldots, N!,$$

and associated joint data distributions for (Y_0, Y_1), where Π_l is a particular $N \times N$ permutation matrix in the set of all $N!$ permutations associating the ranks in the Y_1 distribution with the ranks in the Y_0 distribution.[12] By considering all possible permutations, we 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. Taking all possible permutations of the discrete data distribution, as N becomes large, we obtain an increasingly accurate approximation to the convex hull of the space of all admissible joint distributions of outcomes with pre-specified marginals.[13]

To gauge the intrinsic uncertainty about the set of impact distributions consistent with given marginals, assign equal weight to all possible permutations. Using the sample outcome distributions, we can pair $\underset{\sim}{Y}_1$ with each possible permutation of $\underset{\sim}{Y}_0$ and in this way generate all possible permutational contrasts $\underset{\sim}{\Delta}_l$ obtained from the sample distributions. More generally, in place of permutation matrix Π_l, which shifts mass points across distributions, we could allocate the mass more smoothly, but we do not do so in this paper. A more general approach is described in Section 4.

Two complications preclude direct application of the simple idea of constructing all possible permutations of rankings from the empirical distributions. First, in most data sets there are unequal numbers of persons in the two empirical distributions ($N_{Y_1} \neq N_{Y_0}$) and ours is no exception. To circumvent this problem, we permute the *quantiles* of the

12. Each row or column of an $N \times N$ permutation matrix has a single "1" value and all other values are "0".

13. More precisely, Whitt (1976) shows that collection C coincides with the set of extreme points of the set S of all cumulative joint sample distribution functions having the empirical counterparts of $F_1(y_1 \mid D=1)$ and $F_0(y_0 \mid D=1)$ as marginal distributions. If the elements of $\underset{\sim}{Y}_1$ and $\underset{\sim}{Y}_0$ are each distinct, then C corresponds to the set of all $N \times N$ permutation matrices while the set of all cumulative joint distribution functions corresponds to the set of all $N \times N$ doubly stochastic matrices. An $N \times N$ doubly stochastic matrix B has $\sum_{i=1}^{N} B_{ij} = 1$ and $\sum_{i=1}^{N} B_{ij} = 1$ and all elements are non-negative. The data distributions are dense in the space of all probability measures in the topology of weak convergence. In the limit as $N \to \infty$, we can obtain any admissible bivariate distribution that lies in S by operating on C using doubly stochastic matrices. Thus C is the convex hull of S. To simplify the analysis presented in the text we work with the data distributions, passing to the limit as required.

two distributions using the distribution with the smaller number of observations to set the spacing in the distribution with the larger number. Then min (N_{Y_0}, N_{Y_1}) is the maximum number of quantile spacings considered in the distribution with more observations. All elements in a given quantile class determined in this manner are treated as the same by fixing all values at the within-quantile mean or median or by randomizing which elements within each quantile in the larger distribution are associated with elements in the smaller distribution.

Second, for N sufficiently large, it is computationally demanding to consider all possible permutations of the data distribution. To solve this problem, we collapse both distributions down to a small number of quantile classes and use mean values within each quantile class to summarize the class. Permutations are then done with respect to the reduced classes. Distributions of impacts constructed from such permutations obviously understate the full range of values that could be obtained from permuting the quantiles of the true impact distribution.

We obtain estimates using data on self-reported earnings in the eighteen months following random assignment from an experimental evaluation of the employment and training programmes funded under Title II-A of the U.S. Job Training Partnership Act (JTPA). This programme provides classroom training, on-the-job training and job search assistance to the disadvantaged. As the impacts of training for adult women (age 22 or older) are of substantial interest given current attempts to reform welfare in the U.S., we focus our empirical analysis on this group. Appendix A describes the data in greater detail.

Using percentiles as the finest quantile partition, we obtain 100! possible different permutationally-generated impact distributions. Without invoking any prior information connecting outcomes across the two distributions, any one of these permutation patterns is equally likely. To examine the variation consistent with the experimentally-determined marginals, we take a random sample of 100,000 from the population of 100! percentile permutations. Table 1A presents means and selected quantiles of the distributions of the extremes and the 5th, 25th, 50th, 75th and 95th percentiles of the impact distributions corresponding to this sample of permutations of the quantiles. Table 1B presents means and selected quantiles of parameters of interest for the sample of joint outcome distributions generated by the permutations, including the fraction with a positive impact, the impact standard deviation, and several measures of dependence between Y_0 and Y_1.

Tables 1A and 1B demonstrate substantial variability in the quantiles of the impact distributions we generate. For example, the lowest percentile of the medians is −$1999 compared to the highest percentile of $3636. The 5th percentile of the impact distributions has an interquantile range of almost $2500 in this sample. The true variability is even greater since the permutations producing the most extreme values of the impact percentiles—those wherein the best in one distribution are matched with either the best or the worst in the other—are also very few in number. As a result, they appear very rarely in random samples of this size.

Table 2 displays selected percentiles of the impact distribution for the two extreme cases in which either (1) the two marginal distributions are matched in ascending order or (2) the distributions are matched in reverse order. These two special cases reveal wide variation, with the 5th percentile of the impact distribution equal to $0 in the case of perfect positive quantile dependence and −$22,350 in the case of perfect negative quantile dependence. The 95th percentile of the impact distribution equals $2003 in the first case and $23,351 in the second.

Without additional information, the evidence from experimental data is consistent with a broad range of distributions of programme impacts and interpretations of the

TABLE 1A

Percentiles of parameters of the impact distributions implied by a random sample of 100,000 *percentile permutations*
(National JTPA Study 18 month impact sample; adult females)

Statistic	Distn of minimum	Distn of 5th Pctl	Distn of 25th Pctl	Distn of 50th Pctl	Distn of 75th Pctl	Distn of 95th Pctl	Distn of maximum
Mean	−40690·34	−18278·74	−6426·64	272·31	7632·06	18991·88	59516·06
	(6506·09)	(713·35)	(313·41)	(133·26)	(307·82)	(675·32)	(12506·11)
Minimum	−48606·00	−22350·00	−10814·00	−1999·00	3340·50	12205·00	19207·00
	(7986·12)	(818·59)	(443·03)	(333·23)	(406·16)	(548·91)	(6373·10)
5th Percentile	−48606·00	−21348·00	−8114·00	−41·00	6038·00	16173·00	45808·00
	(7986·12)	(913·24)	(319·22)	(119·48)	(305·29)	(532·35)	(9292·33)
25th Percentile	−47551·00	−19512·00	−7055·00	0·00	6935·00	17789·00	54542·00
	(7964·97)	(802·85)	(316·39)	(0·00)	(307·90)	(611·23)	(11791·40)
50th Percentile	−41969·00	−18359·00	−6426·00	0·00	7647·00	19006·00	61318·00
	(7097·89)	(709·75)	(305·08)	(143·40)	(297·03)	(696·56)	(13404·04)
75th Percentile	−35049·00	−17035·00	−5787·00	510·00	8253·00	20275·00	66860·50
	(5199·60)	(641·74)	(325·72)	(226·50)	(321·38)	(733·39)	(13845·22)
95th Percentile	−27450·00	−15409·00	−4777·00	1197·00	9297·00	22088·00	67156·00
	(3498·30)	(574·54)	(290·52)	(256·32)	(322·02)	(976·11)	(13854·83)
Maximum	−15713·00	−11280·00	−2274·00	3636·00	11707·00	23351·00	67156·00
	(1245·41)	(526·96)	(401·04)	(326·56)	(380·13)	(680·61)	(13854·83)

1. The values in this table are calculated using the percentiles of the two distributions. Each of the 100,000 impact distributions is constructed by matching the percentiles of the Y_1 distribution to a random permutation of the percentiles of the Y_0 distribution. The difference between each percentile of the Y_1 distribution and the percentile of the Y_0 distribution associated with it by the random permutation is the impact for that percentile. Taken together, the percentile impacts form the distribution of impacts. It is the mean, minimum, maximum and percentiles of these impact distributions that are reported in the table.
2. Bootstrap standard errors appear in parentheses.

impact of the programme. Additional information is required to narrow down this class in order to obtain more precise answers to the questions posed in Section (1). This paper considers several plausible assumptions that help to narrow the class of admissible distributions. Before turning to the list of candidate assumptions, we first review the standard statistical approach to bounding features of the distribution of programme impacts using only the information in $F_0(y_0|D=1)$ and $F_1(y_1|D=1)$. While many impact distributions are consistent with the data, they all indicate that variability in programme impacts is an essential feature of it.

(*b*) *Results from classical probability inequalities*

The problem of bounding an unknown joint distribution from known marginal distributions is a classical problem in mathematical statistics. Hoeffding (1940) and Fréchet (1951) demonstrate that the joint distribution is bounded by two functions of the marginal distributions.[14] Their inequalities state that

$$\max [F_1(y_1|D=1)+F_0(y_0|D=1)-1, 0] \leqq F(y_1, y_0|D=1) \leqq \min [F_1(y_1|D=1), F_0(y_0|D=1)].$$

14. These inequalities were first applied in the context of the programme evaluation problem in our (1993) paper presented in May of that year. One of us presented the idea of using these bounds in informal discussions at a 1990 conference sponsored by the Institute for Research on Poverty, affiliated with the University of Wisconsin.

TABLE 1B

Percentiles of parameters of the impact distributions implied by a random sample of 100,000 *percentile permutations*
(National JTPA Study 18 month impact sample; adult females)

Statistic	Distn of percent positive	Distn of impact std dev	Distn of outcome correlation	Distn of Kendall's τ	Distn of Spearman's ρ	Distn of Blomquist's Q
Mean	55·07	12767·97	0·0003	0·0001	0·0001	0·0001
	(0·95)	(766·15)	(0·0003)	(0·0002)	(0·0003)	(0·0003)
Minimum	43·00	8972·34	−0·3456	−0·3362	−0·5018	−0·4400
	(1·26)	(404·28)	(0·0205)	(0·0200)	(0·0291)	(0·0251)
5th Percentile	50·00	11638·99	−0·1539	−0·1119	−0·1657	−0·1600
	(1·08)	(690·87)	(0·0034)	(0·0004)	(0·0006)	(0·0000)
25th Percentile	53·00	12381·83	−0·0695	−0·0457	−0·0677	−0·0800
	(0·94)	(753·44)	(0·0004)	(0·0003)	(0·0004)	(0·0000)
50th Percentile	55·00	12821·49	−0·0056	0·0004	0·0004	−0·0000
	(0·95)	(777·61)	(0·0016)	(0·0002)	(0·0003)	(0·0000)
75th Percentile	57·00	13218·09	0·0630	0·0461	0·0683	0·0800
	(0·98)	(793·33)	(0·0019)	(0·0003)	(0·0004)	(0·0000)
95th Percentile	60·00	13724·41	0·1734	0·1119	0·1660	0·1600
	(0·98)	(805·03)	(0·0019)	(0·0004)	(0·0005)	(0·0000)
Maximum	68·00	14810·39	0·5135	0·3051	0·4415	0·4400
	(1·25)	(839·73)	(0·0409)	(0·0199)	(0·0268)	(0·0266)

1. The values in this table are calculated using the percentiles of the two distributions. Each of the 100,000 impact distributions is constructed by matching the percentiles of the Y_1 distribution to a random permutation of the percentiles of the Y_0 distribution. The difference between each percentile of the Y_1 distribution and the percentile of the Y_0 distribution associated with it by the random permutation is the impact for that percentile. Taken together, the percentile impacts form the distribution of impacts. The impact standard deviation and the percent positive are calculated using the percentile impacts. The impact standard deviation is the standard deviation of the percentile differences. The percent positive is the percent of the percentile impacts greater than or equal to zero. The outcome correlation, Kendall's τ, Spearman's ρ and Blomquist's Q are calculated using the matched percentiles of the Y_1 and Y_0 distributions.
2. Bootstrap standard errors appear in parentheses.

Rüschendorf (1981) establishes that these bounds are tight. Mardia (1970) establishes that both the lower bound and the upper bound are proper probability distributions. At the upper bound, Y_1 is a non-decreasing function of Y_0 (almost everywhere). At the lower bound, Y_0 is a non-increasing function of Y_1 (almost everywhere). These inequalities are not helpful in bounding the distribution of $\Delta = Y_1 - Y_0$, although they bound certain features of it.

By a theorem of Cambanis *et al.* (1976), if $k(Y_1, Y_0)$ is superadditive (or subadditive) then extreme values of $E(k(Y_1, Y_0) \mid D=1)$ are obtained from the upper- and lower-bounding distributions.[15] Since $k(Y_1, Y_0) = (Y_1 - E(Y_1))(Y_0 - E(Y_0))$ is superadditive, the maximum attainable product-moment correlation $r_{Y_0Y_1}$ is obtained from the upper bound distribution while the minimum attainable product-moment correlation is obtained at the lower-bound distribution. Thus it is possible to bound $\mathrm{VAR}(\Delta) = (\mathrm{VAR}(Y_1) + \mathrm{VAR}(Y_0) - 2r_{Y_0Y_1}[\mathrm{VAR}(Y_1)\mathrm{VAR}(Y_0)]^{1/2})$ with the minimum obtained from the Fréchet–Hoeffding upper bound.[16] Checking whether the lower bound of $\mathrm{VAR}(\Delta)$ is statistically significantly

15. k is assumed to be Borel-measurable and right-continuous. k is strictly superadditive if $Y_1 > Y_1'$ and $Y_0 > Y_0'$ imply that $k(Y_1, Y_0) + k(Y_1', Y_0') > k(Y_1, Y_0') + k(Y_1', Y_0)$. k is strictly subadditive if the final inequality is reversed.

16. Note that the maximum value of $r_{Y_0Y_1}$ is obtained at the upper bound and that all other components of the variance of Δ are obtained from the marginal distributions. Thus the minimum variance of Δ is obtained from the Fréchet–Hoeffding upper bound distribution.

TABLE 2

Estimated parameters of the impact distribution; perfect positive dependence and perfect negative dependence cases
(National JTPA Study 18 month impact sample; adult females)

Statistic	Perfect positive dependence	Perfect negative dependence
5th Percentile	0·00	−22350·00
	(47·50)	(547·17)
25th Percentile	572·00	−11755·00
	(232·90)	(411·83)
50th Percentile	864·00	580·00
	(269·26)	(389·51)
75th Percentile	966·00	12791·00
	(305·74)	(253·18)
95th Percentile	2003·00	23351·00
	(543·03)	(341·41)
Percent positive	100·00	52·00
	(1·60)	(0·81)
Impact standard deviation	1857·75	16432·43
	(480·17)	(265·88)
Outcome correlation	0·9903	−0·6592
	(0·0048)	(0·0184)

1. The values in this table are calculated using percentiles of the two distributions. The perfect positive dependence case matches the top percentile in the Y_1 distribution with the top percentile in the Y_0 distribution, the second percentile of the Y_1 distribution with the second of the Y_0 distribution and so on. The perfect negative dependence case matches the percentiles in reverse order, so that the lowest percentile of the Y_0 distribution is matched with the highest percentile of the Y_1 distribution and so on.
2. The perfect positive and perfect negative dependence cases are based on the single permutation having this characteristic in the sample.
3. For each case, the difference between each percentile of the Y_1 distribution and the associated percentile of the Y_0 distribution is the impact for that percentile. Taken together, the percentile impacts form the distribution of impacts. It is the percentiles of these impact distributions that are reported in the upper portion of the table. The impact standard deviation, outcome correlation, and the percent positive are calculated using the percentile impacts. The impact standard deviation is the standard deviation of the percentile differences. The outcome correlation is the correlation of the matched percentiles from the two distributions. The percent positive is the percent of the percentile impacts greater than or equal to zero.
4. Bootstrap standard errors in parentheses.

different from zero provides a test of whether or not the data are consistent with the common effect model. If $Y_1 - Y_0 = \alpha$ is a constant, $\text{VAR}(\Delta) = 0$. Tchen (1980) establishes that Kendall's τ and Spearman's ρ also attain their extreme values at the bounding distributions. The bounding distributions produce the cases of perfect negative dependence and perfect positive dependence discussed in the preceding subsection. In general, useful bounds on the quantiles of the $\Delta = (Y_1 - Y_0)$ distribution cannot be obtained from the Fréchet–Hoeffding bounds. Table 3 presents the range of values of $r_{Y_1 Y_0}$, Spearman's ρ

and $[\mathrm{VAR}(\Delta)]^{1/2}$ for the JTPA data. The ranges are rather wide, but it is interesting to observe that the bounds rule out the common effect model, as $\mathrm{VAR}(\Delta)$ is bounded away from zero.[17]

The standard errors for the statistics derived from the bounding distributions are obtained from the bootstrap. If the statistics are asymptotically normal, the bootstrap standard errors are reliable guides to sampling uncertainty. Under these assumptions, the lower bound for $\mathrm{VAR}(\Delta)$ can be used to test the hypothesis of no heterogeneity—$H_0: \mathrm{VAR}(\Delta) = 0$, and the evidence in Table 3 rejects that hypothesis.

TABLE 3

Characteristics of the distribution of impacts on earnings in the 18 months after random assignment at the Fréchet–Hoeffding bounds

(National JTPA Study 18 month impact sample; adult females)

Statistic	From lower bound distribution	From upper bound distribution
Impact standard deviation	14968·76 (211·08)	674·50 (137·53)
Outcome correlation	−0·760 (0·013)	0·998 (0·001)
Spearman's ρ	−0·9776 (0·0016)	0·9867 (0·0013)

1. These estimates differ slightly from those reported in Table 2 because they were obtained using the empirical c.d.f.s calculated at 100 dollar earnings intervals rather than using the percentiles of the two c.d.f.s.
2. Bootstrap standard errors in parentheses.

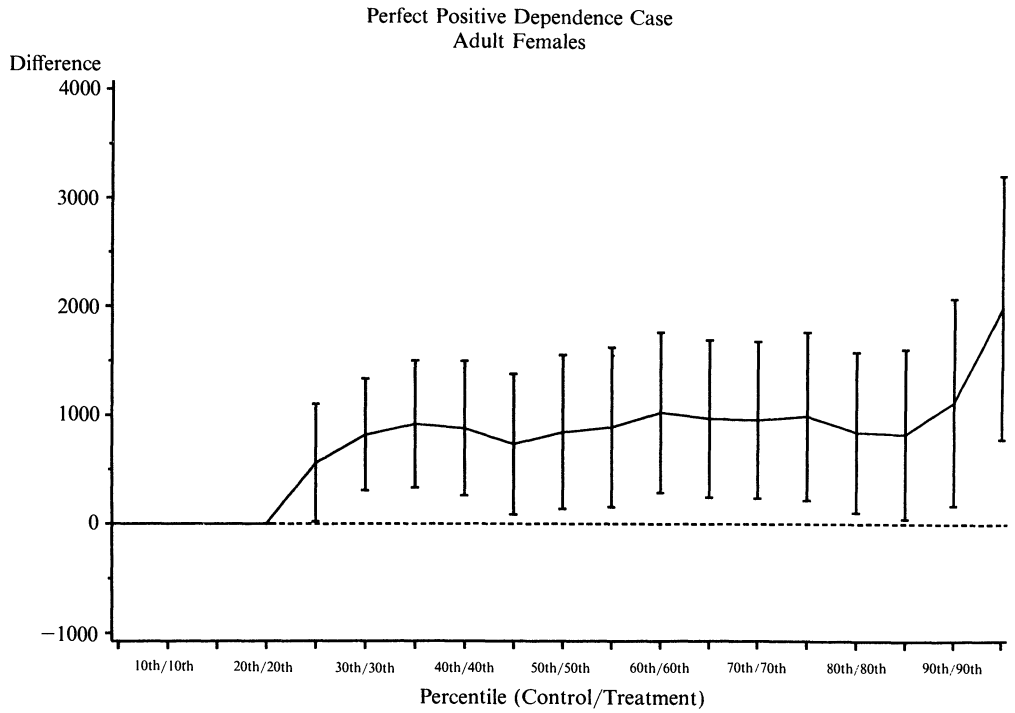
Unfortunately, the statistics are not generally asymptotically normal. The censoring in the lower limit and minimization in the upper limit indicate that the distributions of the test statistics are unlikely to be normal. In Appendix E, we use Monte Carlo methods to investigate the distribution of the statistic and find that it is not normal. Bootstrap confidence intervals centred around the point estimate of Δ have low coverage probabilities even in large samples. We construct Monte Carlo cutoff values for rejecting the null that the true impact standard deviation is zero. Using these values, we reject the null that the true impact standard deviation is zero at the $P = 0{\cdot}0001$ level.

An alternative test examines whether the quantiles of Y_0 and Y_1, $q(Y_0)$ and $q(Y_1)$ respectively, differ by a common constant: $H_0: q(Y_1) - q(Y_0) = \alpha$ for all $0 \leqq q \leqq 100$. Figure 1 presents the difference in quantiles from the two marginal distributions, along with standard error bands.[18] Over a broad range, the hypothesis of constancy of the impact is consistent with the data but at both extremes it is clearly rejected. Another test is based on the observation that a substantial proportion of persons has zero earnings in both distributions, but the proportions are different and both distributions have substantial mass at zero.[19]

17. The discrepancy in the statistics between the perfect positive and negative dependence cases reported in Table 2 and in Table 3 is due to the approximation arising from using permutations of percentiles in Table 2.

18. The same qualitative features are found when we condition on education levels.

19. The null hypothesis that $q(Y_1) - q(Y_0) = \alpha$ for all q is rejected with a p-value of 0·0467 using percentiles. The p-value for the null hypothesis of equality of the proportion of zero earners is 0·0165. If $\alpha = 0$, the proportions should be equal. If $\alpha \neq 0$, one distribution should have no mass at zero if the other has mass at zero.



1 National JTPA Study 18 month impact sample
2. Standard errors for the quantiles are obtained using methods described in Csorgo (1993)

FIGURE 1
Treatment-control differences at percentiles of the 18 month earnings distribution

In an Appendix available on request, we explore the sensitivity of these estimates to measurement error in earnings. Our basic inferences are not altered, including our major inference bounding the variability in programme impacts away from zero.

(*c*) *The discrete case*

The Fréchet–Hoeffding bounds apply to all bivariate outcome distributions.[20] Variables may be discrete, continuous or both discrete and continuous. In this section, we use the bounding distributions to establish the variability in the distribution of impacts on employment status. The latent distribution underlying this situation is multinomial.[21] Let (E, E) denote the event "employed with treatment" and "employed without treatment" and let (E, N) be the event "employed with treatment, not employed without treatment." Similarly, (N, E) and (N, N) refer respectively to cases where a person would not be employed if treated but would be employed if not treated, and where a person would not be employed in either state. The probabilities associated with these events are P_{EE}, P_{EN}, P_{NE} and P_{NN}, respectively. This model can be written in the form of a contingency table. The columns refer to employment and non-employment in the untreated state. The rows refer to employment and non-employment in the treated state.

20. Formulae for multivariate bounds are given in Tchen (1980) and Rüschendorf (1982).
21. The following formulation owes a lot to the missing cell literature in contingency table analysis. See, e.g. Bishop, Fienberg and Holland (1975).

		Untreated E	Untreated N	
Treated	E	P_{EE}	P_{EN}	$P_{E\cdot}$
	N	P_{NE}	P_{NN}	$P_{N\cdot}$
		$P_{\cdot E}$	$P_{\cdot N}$	

FIGURE 2
2 × 2 Table representation

If we observed the same person in both the treated and untreated states, we could fill in the table and estimate the full distribution. Instead, with experimental data we can estimate combinations of the table parameters

$$P_{E\cdot}=P_{EE}+P_{EN}, \quad \text{(employment proportion among the treated)}, \tag{2a}$$

$$P_{\cdot E}=P_{EE}+P_{NE}, \quad \text{(employment proportion among the untreated)}. \tag{2b}$$

The treatment effect is usually defined as

$$T=P_{EN}-P_{NE}, \tag{3}$$

the proportion of people who would switch from nonemployed to employed as a result of treatment minus the proportion of persons who would switch from being employed to not being employed as a result of treatment. Using (2a) and (2b),

$$T=P_{E\cdot}-P_{\cdot E}, \tag{4}$$

so that T can be estimated without bias by subtracting the proportion employed in the control group ($\hat{P}_{\cdot E}$) from the proportion employed in the treatment group ($\hat{P}_{E\cdot}$).

If we wish to decompose T into its two components, P_{EN} and P_{NE}, the experimental data do not give an exact answer except in special cases. In terms of the contingency table presented in Figure 2, we know the row and column marginals but not the individual elements in the table. In the 2×2 table, the case corresponding to the common effect model for continuous outcomes restricts the effect of the programme on employment to be always positive or always negative, so that either P_{EN} or $P_{NE}=0$, respectively. In this case, the model is fully identified. This is analogous to the continuous case in which the common effect assumption, or more generally, an assumption of perfect positive dependence, identifies the joint distribution.

More generally, the Fréchet–Hoeffding bounds restrict the range of admissible values for the cell probabilities. Their application in this case produces:

$$\max[P_{E\cdot}+P_{\cdot E}-1, 0]\leqq P_{EE}\leqq\min[P_{E\cdot}, P_{\cdot E}]$$

$$\max[P_{E\cdot}-P_{\cdot E}, 0]\leqq P_{EN}\leqq\min[P_{E\cdot}, 1-P_{\cdot E}]$$

$$\max[-P_{E\cdot}+P_{\cdot E}, 0]\leqq P_{NE}\leqq\min[1-P_{E\cdot}, P_{\cdot E}]$$

$$\max[1-P_{E\cdot}-P_{\cdot E}, 0]\leqq P_{NN}\leqq\min[1-P_{E\cdot}, 1-P_{\cdot E}].$$

Table 4 presents the Fréchet–Hoeffding bounds for P_{NE} and P_{EN}. The outcome variable is whether or not a person is employed in the 16th, 17th or 18th month after random assignment. The bounds are very wide. Even without taking into account sampling error, the experimental evidence for adult females is consistent with P_{NE} ranging from 0·00 to 0·36. The range for P_{EN} is equally large. Thus as many as 39% and as few as 3% of adult females may have had their employment status improved by participating in the training

TABLE 4

Fraction employed in the 16th, 17th or 18th months after random assignment and Fréchet–Hoeffding bounds on the probabilities P_{NE} and P_{EN}

(National JTPA study 18 month impact sample; adult females)

Parameter	Estimate
Fraction employed in the treatment group	0·64 (0·01)
Fraction employed in the control group	0·61 (0·01)
Bounds on P_{EN}	[0·03, 0·39] (0·01), (0·01)
Bounds on P_{NE}	[0·00, 0·36] (0·00), (0·01)

1. Employment percentages are based on self-reported employment in months 16, 17 and 18 after random assignment. A person is coded as employed if the sum of their self-reported earnings over these three months is positive.
2. P_{ij} is the probability of having employment status i in the treated state and employment status j in the untreated state, where i and j take on the values E for employed and N for not employed. The Fréchet–Hoeffding bounds are given in the text.
3. Asymptotic standard errors appear in parentheses. See Appendix E for an analysis of the performance of the bootstrap in this context.

programme. As many as 36% and as few as 0% may have had their employment status harmed by participating in the programme. From (3), we know that the net difference $(P_{EN}-P_{NE})=T$, so that high values of P_{EN} are associated with high values of P_{NE}. As few as 25% $[(0{\cdot}64-0{\cdot}39)\times 100]$ and as many as 61% of the women would have worked whether or not they entered the programme ($P_{EE}\in[0{\cdot}25, 0{\cdot}61]$).

Uncritical application of bootstrapping to obtain the standard errors for the bounds is no more justified in the discrete case than it is in the continuous data case. Although the sample proportions are asymptotically normally distributed, the upper bounds are the minima of two normal random variables and the lower bounds are censored normal random variables.

Appendix E investigates this problem. We present the asymptotic distributions for the upper and lower bounds. If the non-negativity constraint in the lower bound is not close to binding, then the assumption of asymptotic normality is innocuous, and coverage probabilities are accurate. The distribution is decidedly non-normal if the constraint binds. The bootstrap coverage probabilities are too high in this case but only by a few percentage points. As the terms in the upper limit approach equality, the distributions become highly non-normal and the coverage probabilities are too low, but again, in samples of our size, and with parameters of our values, the bias is slight. The farther apart are the elements in the upper bound, the more normal the distribution and the more accurate are the coverage probabilities. Overall, the analysis presented in Appendix E supports the use of the bootstrap to compute coverage probabilities for all bounds.

From the evidence presented in Table 4, we cannot distinguish two different stories. The first story is that the JTPA programme benefits many people by facilitating their employment but also harms many people who would have worked if they had not participated in the programme. The second story is that the programme benefits and harms few

people. Conditioning on other background variables (in results available upon request) does not go far in resolving the intrinsic uncertainty in the data. Thus in both the discrete and continuous cases, the experimental data are consistent with a wide variety of impact distributions.

4. HOW FAR CAN WE DEPART FROM PERFECT DEPENDENCE AND STILL PRODUCE PLAUSIBLE ESTIMATES OF PROGRAMME IMPACTS?

The evidence presented in Section 3 suggests that the range of joint outcome and impact distributions consistent with the marginals determined by the JTPA experiment is very wide. In this section, we continue these explorations to see how far from perfect positive dependence we can venture and still produce plausible impact distributions consistent with the marginals. We first generalize the common effect dummy endogenous variable model by preserving perfect positive dependence but allowing the impact of treatment to vary as a function of Y_0. We then consider perturbations away from the case of perfect positive dependence in the ranks of Y_0 and Y_1. The analysis presented in Section 5 demonstrates that positive dependence in outcomes among participants is produced by many plausible models of participant self-selection into programmes.

The dummy endogenous variable model assumes a constant treatment effect for all persons, so that Y_1 and Y_0 differ by a constant at all quantiles of the Y_0 distribution. A generalization of this model assumes that the best in one distribution is the best in the other distribution. Our generalization preserves perfect dependence in the ranks between the two distributions but does not require the impact to be the same at all quantiles of the base state distribution. We obtain the deterministic impact function, $\Delta(y_0) = y_1(y_0) - y_0$, by equating quantiles across the two distributions, forming the pairs

$$\{(y_0, y_1) \mid \inf_{y_1} F_1(y_1 \mid D=1) > q \text{ and } \inf_{y_0} F_0(y_0 \mid D=1) > q,\ 0 \leqq q \leqq 1\}.$$

For the case of absolutely continuous distributions with positive density at y_0, the impact function can be written as $\Delta(y_0) = F_1^{-1}(F_0(y_0 \mid D=1)) - y_0$. We can use experimental data to test non-parametrically for the classical common effect model which implies that $\Delta(y_0)$ is a constant for all y_0. Figure 1 plots the estimated impact function assuming perfect positive dependence in the sense of quantile rankings across the two outcome distributions. Standard errors for the quantiles are obtained from formulae in Csörgo (1983). The estimates are revealing. Over a broad interval the impact is constant, although it turns up at the highest earnings levels and is zero in the lowest levels.

We can form other pairings across quantiles by mapping quantiles from the Y_0 distribution into quantiles from the Y_1 distribution using the map Π, where $\Pi: q_0 \to q_1$. Experimental data are consistent with all admissible transformations including $q_0 = 1 - q_1$, where the best in one distribution is mapped into the worst in the other.

More generally, we could distribute the mass at one quantile in a distribution to the quantiles of the other distribution more continuously. Let $M_\theta(y_1, y_0)$ and $M_\theta^*(y_0, y_1)$ be bounded continuous Markov operators. If the data are continuously distributed, they are, respectively, $f_1(y_1 \mid Y_0 = y_0, D=1)$ and $f_0(y_0 \mid Y_1 = y_1, D=1)$, the conditional densities of Y_1 and Y_0. If the data are discrete, M_θ and M_θ^* are Markov transition matrices that satisfy

$$F_1(y_1 \mid D=1) = M_\theta(y_1, y_0) F_0(y_0 \mid D=1)$$

and

$$F_0(y_0 \mid D=1) = M_\theta^*(y_0, y_1) F_1(y_1 \mid D=1).$$

In the continuous data case with absolutely continuous random variables, consistency of the marginals and conditionals requires

$$f_0(y_0 \mid D=1) = \int M_\theta^*(y_0, y_1) f_1(y_1 \mid D=1) dy_1,$$

$$f_1(y_1 \mid D=1) = \int M_\theta(y_1, y_0) f_0(y_0 \mid D=1) dy_0,$$

where the integrals are assumed to exist. For the discrete data case, we seek all pairs of Markov transition matrices M_θ, M_θ^* that satisfy the pair of equations:

$$F_1(y_1 \mid D=1) = M_\theta(y_1, y_0) M_\theta^*(y_0, y_1) F_1(y_1 \mid D=1),$$
$$F_0(y_0 \mid D=1) = M_\theta(y_0, y_1) M_\theta^*(y_1, y_0) F_0(y_0 \mid D=1).$$

The corresponding equations in the continuous case are

$$f_1(y_1 \mid D=1) = \iint M_\theta(y_1, t) M_\theta^*(t, z) f_1(z \mid D=1) dt dz.$$

$$f_0(y_0 \mid D=1) = \iint M_\theta^*(y_0, t) M_\theta(t, z) f_0(z \mid D=1) dt dz,$$

where the integrals are assumed to exist.

We do not pursue this more general approach in this paper and consider only the extreme points of the set of all admissible joint distributions. These distributions transfer all the mass at one quantile of the Y_0 distribution to a single quantile of the Y_1 distribution. In the discrete case, M_θ and M_θ^* are permutation matrices with $M_\theta^* = M_\theta^{-1} = M_\theta'$. These extreme distributions define the convex hull of all admissible joint distributions with given marginals.[22]

We now generalize from the case of perfect positive dependence to allow some slippage in the quantile ranks between the two distributions. We consider a measure of disarray from perfect dependence in the ranks that characterizes all possible bivariate data distributions. We assume that the data are from absolutely continuous distributions and that there are no ties in the sample distribution. The generalization needed to handle the case of data with mass points is presented in Appendix D.

For a given quantile scale, consider any permutation of the quantiles of the distribution of Y_0 associated with the quantiles of the distribution of Y_1 via a permutation which maps q_0 into q_1. Y_1 and Y_0 are perfectly arrayed if $\Pi = I$ for a fixed definition of the quantile width (e.g. percentiles, deciles, etc.). For other permutations, there is some level of disarray.[23]

22. Strassen (1965) presents conditions for the existence at least one joint distribution consistent with the marginals.

23. An inversion relative to the quantiles for the distribution of Y_0 is said to occur each time, in binary comparisons, an element of the quantiles of the Π-induced Y_1 is bigger than a succeeding element, going down the full Y_1 array from the first element to the last. Thus for a four-element array 2, 3, 1, 4, taken from $\{1, 2, 3, 4\}$, there are two inversions, 2 before 1 and 3 before 1.

For a permutation of the Y_1 associated with the Y_0 that is induced by some Π, we can define the total number of inversions in the array as

$$V=\sum_j \sum_{i\leqq j} h_{ij}, \qquad h_{ij}=\begin{cases}1 & \text{if } Y_1^{(i)}>Y_1^{(j)};\\ 0 & \text{otherwise,}\end{cases}$$

where $Y_1^{(i)}$ is the value of Y_1 associated with the i-th quantile of Y_0. The value of V may range from 0 to $(1/2)I(I-1)$, where I is the pre-assigned number of quantiles. $V=0$ arises in the case of perfect positive dependence in the quantile ranks and $V=(1/2)I(I-1)$ arises when there is perfect inverse ordering in the quantile ranks.

Kendall's rank correlation measure τ may be written as

$$\tau=1-\frac{4V}{I(I-1)}=1-4\frac{\sum_j \sum_{i\leqq j} h_{ij}}{I(I-1)},$$

where τ lies in the interval $[-1,1]$.[24] All possible bivariate distributions for the chosen quantile spacing that are induced by different possible choices of Π for the given marginals are produced by letting τ vary over the entire interval.

Since h_{ij} is a superadditive function (fixing the Y_0 ranking), and since sums of superadditive functions are superadditive, we know from the analysis of Cambanis *et al.* (1976) that τ attains its maximum value at the Fréchet–Hoeffding upper bound and its minimum value at the Fréchet–Hoeffding lower bound. Thus we can characterize the bounding distributions as producing minimal and maximal disarray between Y_1 and Y_0 for a given choice of quantile spacing. By specifying τ, we pick a level of dependence between the two outcomes and hence a level of permutational disarray. Thus τ is a measure of slippage in the ranks of the quantiles across the two distributions with $\tau=1$ corresponding to perfect positive dependence and $\tau=-1$ corresponding to perfect negative dependence. Varying τ between -1 and 1 traces out all possible permutations of the quantiles across the distributions.[25]

Using the sample distributions, we pair each quantile of Y_0 with each possible quantile of Y_1 and generate all possible rearrangements of the quantile ranks for a given choice of quantile spacing. The generated distributions can be used to produce sample gain distributions for different assumed levels of disarray τ. Tables 5A and 5B present estimates of quantiles of the impact distribution and other parameters of interest for various values of τ. Plausible impact distributions require high measures of positive dependence. Values of τ much less than 1·0 produce implausible distributions of impacts. Note the implausibly large gains and losses obtained when τ is 0·5 or less. However, the conclusion that a majority of the adult female participants benefitted from the programme is robust to the choice of τ.[26]

5. THE INDUCED STRUCTURE OF DEPENDENCE AMONG OUTCOMES FROM AGENT DECISION RULES

Many different distributions of impacts are consistent with the data produced from a social experiment. Information about the programme participation decision can sometimes

24. See, e.g. Kendall (1970) and Daniels (1944, 1948).

25. Viverberg (1993) conducts a related sensitivity analysis in an unidentified Roy Model using a normal distribution.

26. In an earlier version of this paper, we emulated robust-Bayesian methods by placing priors over τ (and α to be introduced in Section 6) to perform a non-Bayesian sensitivity analysis that examined the consequences of using different weighted averages of τ. These sensitivity results are available on request from the authors. The results for values of τ below 0·8 are unreasonable to us because they imply negative earnings impacts that are too large to be plausible.

TABLE 5A

Percentiles of parameters of the impact distribution as τ varies based on random samples of 50 permutations with each value of τ

(National JTPA Study 18 month impact sample; adult females)

τ	Minimum	5th Pctl	25th Pctl	50th Pctl	75th Pctl	95th Pctl	Maximum
1·00	0·00	0·00	572·00	864·00	966·00	2003·00	18550·00
	(703·64)	(47·50)	(232·90)	(269·26)	(305·74)	(543·03)	(5280·67)
0·95	−14504·00	0·00	125·50	616·00	867·00	1415·50	48543·50
	(1150·01)	(360·18)	(124·60)	(280·19)	(272·60)	(391·51)	(8836·49)
0·90	−18817·00	−1168·00	0·00	487·00	876·50	2319·50	49262·00
	(1454·74)	(577·84)	(29·00)	(265·71)	(282·77)	(410·27)	(6227·38)
0·70	−25255·00	−8089·50	−136·00	236·50	982·50	12158·50	55169·50
	(1279·50)	(818·25)	(260·00)	(227·38)	(255·78)	(614·45)	(5819·28)
0·50	−28641·50	−12037·00	−1635·50	0·00	1362·50	16530·00	58472·00
	(1149·22)	(650·31)	(314·39)	(83·16)	(249·29)	(329·44)	(5538·14)
0·30	−32621·00	−14855·50	−3172·50	0·00	4215·50	16889·00	54381·00
	(1843·48)	(548·48)	(304·62)	(379·96)	(244·67)	(423·05)	(5592·86)
0·00	−44175·00	−18098·50	−6043·00	0·00	7388·50	19413·25	60599·00
	(2372·05)	(630·73)	(300·47)	(163·17)	(263·25)	(423·63)	(5401·02)
−0·30	−48606·00	−20566·00	−8918·50	779·50	9735·50	21093·25	65675·00
	(1281·80)	(545·99)	(286·92)	(268·02)	(300·59)	(462·13)	(5381·91)
−0·50	−48606·00	−21348·00	−9757·50	859·00	10550·50	22268·00	67156·00
	(1059·06)	(632·55)	(351·55)	(315·37)	(255·28)	(435·78)	(5309·90)
−0·70	−48606·00	−22350·00	−10625·00	581·50	11804·50	23351·00	67156·00
	(1059·06)	(550·00)	(371·38)	(309·84)	(246·58)	(520·93)	(5309·90)
−0·90	−48606·00	−22350·00	−11381·00	580·00	12545·00	23351·00	67156·00
	(1059·06)	(547·17)	(403·30)	(346·12)	(251·07)	(341·41)	(5309·90)
−0·95	−48606·00	−22350·00	−11559·00	580·00	12682·00	23351·00	67156·00
	(1059·06)	(547·17)	(404·67)	(366·37)	(255·97)	(341·41)	(5309·90)
−1·00	−48606·00	−22350·00	−11755·00	580·00	12791·00	23351·00	67156·00
	(1059·06)	(547·17)	(411·83)	(389·51)	(253·18)	(341·41)	(5309·90)

1. The values in this table are constructed from the percentiles of the two distributions. In the cases of $\tau = 1$ and $\tau = -1$, they are based on the single permutation with the indicated value of τ. In the remaining cases, they are the mean of the indicated parameter of the impact distribution over a random sample of 50 permutations having the indicated value of τ. The difference between each percentile of the Y_1 distribution and the percentile of the Y_0 distribution associated with it by the permutation is the impact for that percentile. Taken together, the percentile impacts form the distribution of impacts. It is the minimum, maximum and percentiles of these impact distributions that are reported in the table.
2. Bootstrap standard errors appear in parentheses.

be used to recover the joint distribution of outcomes from experimental or non-experimental data. Such information is also informative about whether personal choices are compatible with stated social objectives. Viewed ex post, do the persons subject to treatment perceive themselves as having benefitted from the programme?

Suppose that Y_0 and Y_1 are the net potential outcomes from non-participation and participation, respectively. If agents are uncertain about both their potential Y_0 and Y_1 values, but they know the distributions of these variables, then rational individuals concerned only about their own outcomes who have strictly monotonically increasing utility functions $U(y)$ pick the option with the greatest expected utility. Thus,

$$D = \begin{cases} 1 & \text{if } \int U(y_0)dF_0(y_0) \leqq \int U(y_1)dF_1(y_1); \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

TABLE 5B

Percentiles of parameters of the impact distribution as τ varies based on random samples of 50 permutations with each value of τ

(National JTPA Study 18 month impact sample; adult females)

τ	Percent positive	Impact std dev	Outcome correlation	Spearman's ρ	Blomquist's Q
1·00	100·00	1857·75	0·9903	1·0000	1·0000
	(1·60)	(480·17)	(0·0048)	(0·0000)	(0·0000)
0·95	96·00	6005·96	0·7885	0·9676	0·9600
	(3·88)	(776·14)	(0·0402)	(0·0007)	(0·0000)
0·90	88·00	6388·98	0·7591	0·9361	0·9200
	(5·10)	(474·65)	(0·0257)	(0·0015)	(0·0062)
0·70	72·50	8160·36	0·5996	0·7921	0·7600
	(5·30)	(351·67)	(0·0199)	(0·0021)	(0·0117)
0·50	58·00	9475·85	0·4561	0·6170	0·5600
	(4·31)	(327·81)	(0·0161)	(0·0029)	(0·0181)
0·30	57·00	10584·06	0·3185	0·4083	0·3600
	(2·34)	(290·72)	(0·0129)	(0·0026)	(0·0151)
0·00	54·00	12879·21	−0·0147	−0·0093	0·0000
	(1·11)	(259·24)	(0·0106)	(0·0012)	(0·0123)
−0·30	54·00	14550·94	−0·2985	−0·4272	−0·3600
	(0·89)	(267·83)	(0·0093)	(0·0030)	(0·0161)
−0·50	54·00	15294·88	−0·4359	−0·6300	−0·5600
	(0·94)	(274·34)	(0·0122)	(0·0031)	(0·0175)
−0·70	53·00	15852·82	−0·5434	−0·8051	−0·7200
	(0·86)	(267·69)	(0·0153)	(0·0029)	(0·0117)
−0·90	52·00	16265·17	−0·6254	−0·9544	−0·9200
	(0·73)	(267·49)	(0·0174)	(0·0015)	(0·0039)
−0·95	52·00	16376·95	−0·6479	−0·9880	−0·9600
	(0·78)	(267·35)	(0·0180)	(0·0005)	(0·0000)
−1·00	52·00	16432·43	−0·6592	−1·0000	−1·0000
	(0·81)	(265·88)	(0·0184)	(0·0000)	(0·0000)

1. The values in this table are constructed from the percentiles of the two distributions. In the cases of $\tau = 1$ and $\tau = -1$, they are based on the single permutation with the indicated value of τ. In the remaining cases, they are the mean of the indicated parameter of the impact distribution over a random sample of 50 permutations having the indicated value of τ. The difference between each percentile of the Y_1 distribution and the percentile of the Y_0 distribution associated with it by the permutation is the impact for that percentile. Taken together, the percentile impacts form the distribution of impacts. The percent positive, the impact standard deviation, the outcome correlation, Spearman's ρ and Blomquist's Q are calculated using the matched percentiles of the Y_1 and Y_0 distributions.
2. Bootstrap standard errors appear in parentheses.

where we suppress the dependence of U, $F_0(y_0)$ and $F_1(y_1)$ on conditioning variables X to simplify the expressions. If U is concave, a sufficient condition for $D=1$ is that Y_1 second-order stochastically dominates Y_0, so that $\int_{-\infty}^{a} F_1(y_1)dy_1 \leqq \int_{-\infty}^{a} F_0(y_0)dy_0$ for all a. This is a rational self-selection requirement for persons to participate in a social programme. Information from programme participants and nonparticipants does not provide the requisite data to conduct this test of rationality. Even ideal social experiments provide only the conditional distributions $F_0(y_0 \mid D=1)$ and $F_1(y_1 \mid D=1)$. Data from regimes with univeral programme participation generate only one of the two marginal distributions required to perform this test.

The distributions produced from a social experiment can be used to check if expectations are ex ante rational. If this is true, it follows that for each person in the programme

$$\int U(y_1)dF_1(y_1 \mid D=1) > \int U(y_0)dF_0(y_0 \mid D=1), \tag{6}$$

where $U(y)$ is assumed to be common across persons. A necessary and sufficient condition for (6) to be satisfied for all U is that realized Y_1 second-order stochastically dominates realized Y_0 given $D=1$, so that $\int_{-\infty}^{a} F_1(y_1 \mid D=1)dy_1 < \int_{-\infty}^{a} F_0(y_0 \mid D=1)dy_0$ for all a.[27] This test assumes a strong form of the rational expectations hypothesis—that persons base their decisions on the observed ex post outcome distributions.

Using three different tests of the null hypothesis that $Y_1 \mid D=1$ stochastically dominates $Y_0 \mid D=1$, we do not reject the null. As shown in Figure 3, for all values of $y_1 = y_0$,

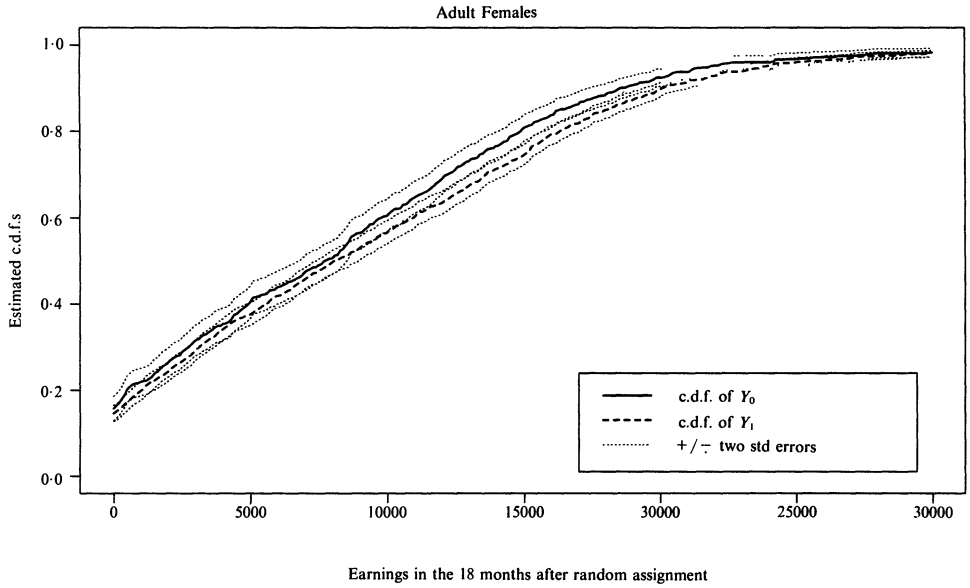


FIGURE 3
C.d.f.s of Y_1 and Y_0

$F_1(y_1 \mid D=1) < F_0(y_0 \mid D=1)$. Since first-order stochastic dominance implies second-order dominance, it is not surprising that the test statistics for hypothesis (6) do not reject that hypothesis. Appendix F describes the tests and reports the results from applying them to our data. There is strong evidence of rational behaviour in the sense of inequality (6). Personal objectives and programme objectives are aligned for adult women. There is ex post regret among randomized-out nonparticipants whatever the shape of their common concave utility function.

Participation condition (5) uses no information about the joint distribution of outcomes and sheds no light on it. Rationality condition (6) may shed light on the joint distribution $F(y_1, y_0)$ or $F(y_1, y_0 \mid D=1)$. Suppose that in advance of participating in a programme, persons know their own (Y_0, Y_1) values but that observing analysts do not. In choosing to participate, $Y_1 \geqq Y_0$ is a requirement for rationality. In the participant

27. The inequality is reversed for a risk-loving agent.

population, the requirement becomes $\Pr(Y_1 \geqq Y_0 \mid Y_0 = y_0, D=1) = 1$. This is a strong form of stochastic dominance. All of the mass of the Y_1 distribution conditional on Y_0 is to the right of y_0. No matter what the population dependence among (Y_0, Y_1), there is a strong positive dependence in potential outcomes among participants.

More generally, persons may not know (Y_0, Y_1) but may base their participation decisions on unbiased guesses (Y_0^*, Y_1^*) about them. Then we may write $Y_0^* = Y_0 + \varepsilon_0$ and $Y_1^* = Y_1 + \varepsilon_1$ where $E(\varepsilon_0, \varepsilon_1) = (0, 0)$, $(\varepsilon_0, \varepsilon_1) \perp\!\!\!\perp (Y_0, Y_1)$ and $\varepsilon_0 \perp\!\!\!\perp \varepsilon_1$, and where "$\perp\!\!\!\perp$" denotes independence.

In this case, if $D = 1(Y_1^* > Y_0^*)$, conditioning on realized values produces positive regression dependence between Y_1 and Y_0 for participants, which means that $\Pr(Y_1 \leqq y_1 \mid Y_0 = y_0, D=1)$ is non-increasing in y_0 for all y_1. This in turn implies that Y_1 is right-tail increasing in y_0. That is, $\Pr(Y_1 > y_1 \mid Y_0 > y_0, D=1)$ is non-decreasing in y_0 for all y_1. Intuitively, the higher the value of y_0, the more the mass in the conditional Y_1 distribution is shifted to the right so that "high values of Y_0 go with high values of Y_1." That is, Y_1 being right-tail increasing given y_0 implies that Y_1 and Y_0 (given $D=1$) are positive quadrant dependent, so that $\Pr(Y_1 \leqq y_1 \mid Y_0 \leqq y_0, D=1) \geqq \Pr(Y_1 \leqq y_1 \mid D=1)$ and $\Pr(Y_0 \leqq y_0 \mid Y_1 \leqq y_1, D=1) \geqq \Pr(Y_0 \leqq y_0 \mid D=1)$.[28] Common measures of dependence like the product-moment correlation, Kendall's τ and Spearman's ρ are all positive when there is positive quadrant dependence.

Rationality under these programme participation rules thus imposes a restriction on the nature of the dependence between Y_0 and Y_1 given $D=1$. With enough structure, one can recover the full distribution of outcomes and extrapolate out of sample, as we show in Section 8. Evidence against such dependence in populations of persons for whom $D=1$ is evidence against the income-maximizing Roy model. Even if Y_0 and Y_1 are negatively dependent in the population, they are positively dependent given $D=1$ in this model if agents are outcome maximizers. We demonstrate below that imposing rationality in this sense helps eliminate some of the uncertainty that is intrinsic in both experimental and non-experimental estimates of programme impacts, and in some cases eliminates it entirely.

6. USING PRIOR INFORMATION TO REDUCE THE INTRINSIC UNCERTAINTY IN DATA FROM SOCIAL PROGRAMMES: THE CASE OF THE 2×2 TABLE

In considering outcomes like employment and earnings, many plausible models of programme participation suggest that outcomes in the treatment state are "positively related" to outcomes in the non-treatment state for persons who self-select into training. There is a widely-held belief that good persons are good at whatever they do. This section applies this notion to the analysis of 2×2 tables.

In order to make this notion operational it is necessary to be more precise about what is meant by dependence among binary outcomes. Notions of dependence in 2×2 tables are presented in Bishop, Feinberg and Holland (1975). In terms of the table in Figure 1, the most commonly used measure of association between the two outcomes is the cross product ratio

$$\eta = \frac{P_{EE}P_{NN}}{P_{EN}P_{NE}}.$$

When $\eta = 1$, the treatment and non-treatment outcomes are independent. This

28. These inequalities are strict except in the case where Y_0 and Y_1 are binary random variables. Tong (1980) shows that these notions of dependence are all equivalent in the case of binary random variables.

measure: (i) is invariant under the interchange of rows and columns; (ii) is invariant to the proportion of persons participating in the programme; (iii) is interpretable and is the ratio of the odds of being employed in the non-participation state conditional on being employed in the participation state (P_{EE}/P_{EN}) and the odds of employment in the non-participation state conditional on not being employed in the participation state (P_{NE}/P_{NN}).

By property (iii), the higher is η, the more likely it is for a person employed in the participation state to be employed in the non-participation state. As the conditional (on employment in the participation state) odds ratio of employment in the non-participation state (P_{EE}/P_{EN}) becomes large and the conditional (on no employment in the participation state) odds ratio of employment in the non-participation state becomes small, η becomes large. In this case, workers in one state are very likely to be workers in the other state, and nonworkers in one state are likely to be nonworkers in the other state. In the case of reverse association, $\eta \to 0$.

In the 2×2 table many apparently diverse notions of positive dependence are equivalent. Positive covariance, association, positive regression dependence, right tail increasing dependence and positive quadrant dependence all describe the same positive "association" of E and N. Thus there is no loss in generality in using η or a monotonic transformation of it.

Given η, the row and column marginals $P_{E\cdot}$ and $P_{\cdot E}$, and the requirement that the probabilities sum to one, we can uniquely determine all of the elements of the 2×2 table.

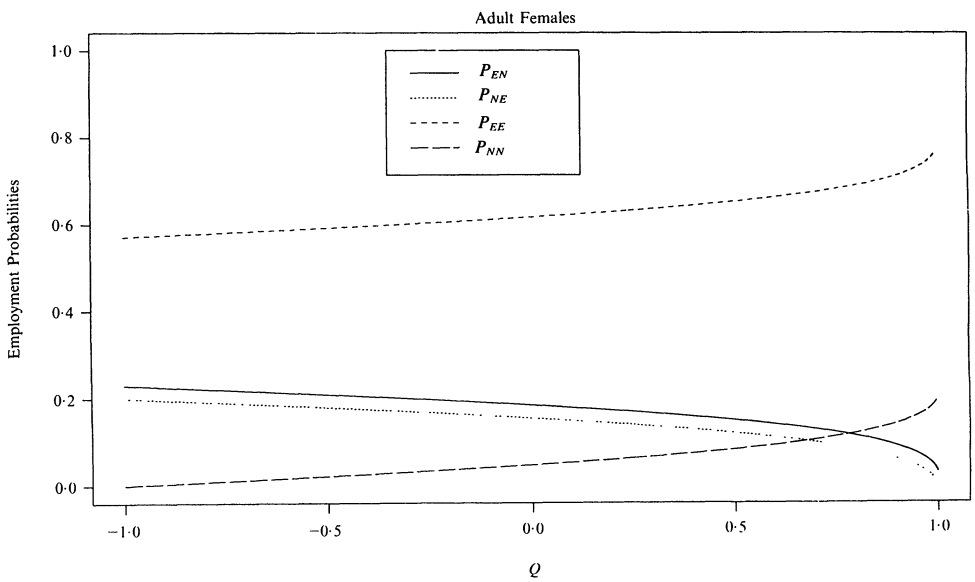


FIGURE 4

Cell employment probabilities as functions of Q

For the employment data analyzed in Section 3(c), Figure 4 presents the relationship between these elements and the measure of association Q,

$$Q = \frac{\eta - 1}{\eta + 1},$$

where $Q \in [-1, 1]$ and $Q=0$ when $\eta=1$ (and the rows and columns are independent). Higher values of Q are associated with higher values of η and thus with greater dependence in outcomes between the two states. As we specify higher values of Q both P_{EN} and P_{NE} decline in absolute value. The difference $P_{EN}-P_{NE}$ is the mean treatment effect T and is constant for all Q. Intuitions that outcomes are strongly positively related across the two states translate into statements that Q is positive and close to one.

In Appendix D, we present a more general approach that combines the methods of this section and Section 4 to allow for mass points at zero in the Y_1 and Y_0 distributions. This case is empirically relevant as there is a significant mass point at zero in both cases in the JTPA data.

7. DECONVOLUTION WHEN GAINS ARE NOT ANTICIPATED AT THE TIME PROGRAMME PARTICIPATION DECISIONS ARE MADE: A NONPARAMETRIC RANDOM COEFFICIENTS MODEL

Another approach to obtaining the joint distribution of outcomes postulates that the gain, Δ, is independent of the base state outcome Y_0 for participants, or

$$Y_0 \perp\!\!\!\perp \Delta \mid D=1. \tag{C-1}$$

If true, $Y_1 = Y_0 + R\Delta$, $R\Delta \perp\!\!\!\perp Y_0$, where $R=1$ if a person is randomized into the programme and $R=0$ otherwise, and where the conditioning on $D=1$ is left implicit.

Condition (C-1) would be satisfied if the gain cannot be forecast at the time decisions are made about programme participation. This case is discussed in Heckman and Robb (1985, p. 181), Heckman (1997*b*) and Heckman and Smith (1997) and produces a model that is intermediate between the common-effect model and the variable-impact model where the impact is anticipated by agents.

Under (C-1), we may write the density of Y_1 as a convolution of Y_0 and Δ

$$f_1(y_1 \mid R=1, D=1) = f_\Delta(\Delta \mid R=1, D=1) * f_0(y_0 \mid R=0, D=1),$$

where "$*$" denotes convolution. Within this context, we may consider "densities" with mass points, such as occurs at zero earnings in our data from the JTPA experiment. Exploiting the independence of Y_0 and Δ, the characteristic function of Y_1 may be written as

$$E(e^{itY_1} \mid D=1) = E(e^{it\Delta} \mid D=1) E(e^{itY_0} \mid D=1).$$

Solving for $E(e^{it\Delta} \mid D=1)$, we obtain

$$\varphi(t) = E(e^{it\Delta} \mid D=1) = E(e^{itY_1} \mid D=1) / E(e^{itY_0} \mid D=1).$$

Then using the inversion theorem[29]

$$F(\Delta \mid D=1) = \frac{1}{2} + \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{[e^{it\Delta}\varphi(-t) - e^{-it\Delta}\varphi(t)]}{it}\, dt. \tag{7}$$

29. The ratio of two characteristic functions need not be a characteristic function. By Bochner's theorem (see, e.g. Gnedenko (1976)), for $\varphi(t)$ to be a characteristic function, it must satisfy $\varphi(0)=1$, $\varphi(t)$ must be continuous for all t and $\varphi(t)$ must be positive definite. This hypothesis could be tested using the methods presented in Heckman, Robb and Walker (1990). The test would consist of checking if the ratio of the two sample characteristic functions is "within sampling variation" of being positive definite.

(See, e.g. Kendall and Stuart (1977), p. 98). We can thus recover the distribution of Δ from the distributions of Y_1 and Y_0 produced by the experiment. Heckman and Smith (1997) extend this procedure to the analysis of nonexperimental data under conditional independence assumptions specified in their paper.

(*a*) *A random coefficient approach*

Setting $Y_0 = X\beta + U$, we obtain a conventional random coefficient model,

$$Y = RY_1 + (1-R)Y_0 = X\beta + R\Delta + U. \tag{8}$$

Using a standard variance components model, we may write $E(\Delta) = \bar{\Delta}$, $\varepsilon = \Delta - \bar{\Delta}$ to obtain

$$Y = X\beta + R\bar{\Delta} + R\varepsilon + U$$

where we assume that U and ε are independent of X ($U, \varepsilon \perp\!\!\!\perp X$). The assumed independence between Δ and Y_0 translates into independence between ε and U. The increase in the variance in the residuals of outcomes for participants can be used to estimate $\mathrm{VAR}(\varepsilon)$. From participant residuals, we can identify $\mathrm{VAR}(\varepsilon + U) = \mathrm{VAR}(\varepsilon) + \mathrm{VAR}(U)$. From nonparticipant residuals, we can identify $\mathrm{VAR}(U)$. Thus we can test an implication of the assumption that $\Delta \perp\!\!\!\perp U \mid D = 1$ by using the empirical analogs of $\mathrm{VAR}(\varepsilon + U)$ and $\mathrm{VAR}(U)$ for participants and non-participants, respectively. A finding that $\mathrm{VAR}(\varepsilon + U) < \mathrm{VAR}(U)$ indicates the failure of independence between ε and U and therefore a failure of the assumption that $\Delta \perp\!\!\!\perp U \mid D = 1$.

TABLE 6

Random coefficient and deconvolution estimates of the 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

1. Estimated standard errors appear in parentheses where available.
2. Random coefficient model includes race/ethnicity, schooling and site indicators. Only the treatment coefficient is treated as random.
3. The estimated impact variance for the random coefficient model is obtained from a regression of the squared residuals from the corresponding fixed coefficient model on the treatment indicator.
4. The estimated percent positive for the random coefficient model assumes that Δ is normally distributed.
5. Mean impact, impact standard deviation and the fraction of positive impacts for the deconvolution case are obtained from the smoothed density. Values for the unsmoothed density differ only slightly from those reported here.

The first row of Table 6 presents estimates based on this approach. There is mild evidence in support of the hypothesis that $\mathrm{VAR}(\Delta) > 0$, suggesting that a more elaborate deconvolution approach to estimating the distribution of Δ is likely to be fruitful.

(*b*) *Empirical deconvolution*

A more general and robust approach exploits (7) and the empirical characteristic functions for Y_1 and Y_0 to estimate the distribution of Δ. The details needed to implement the

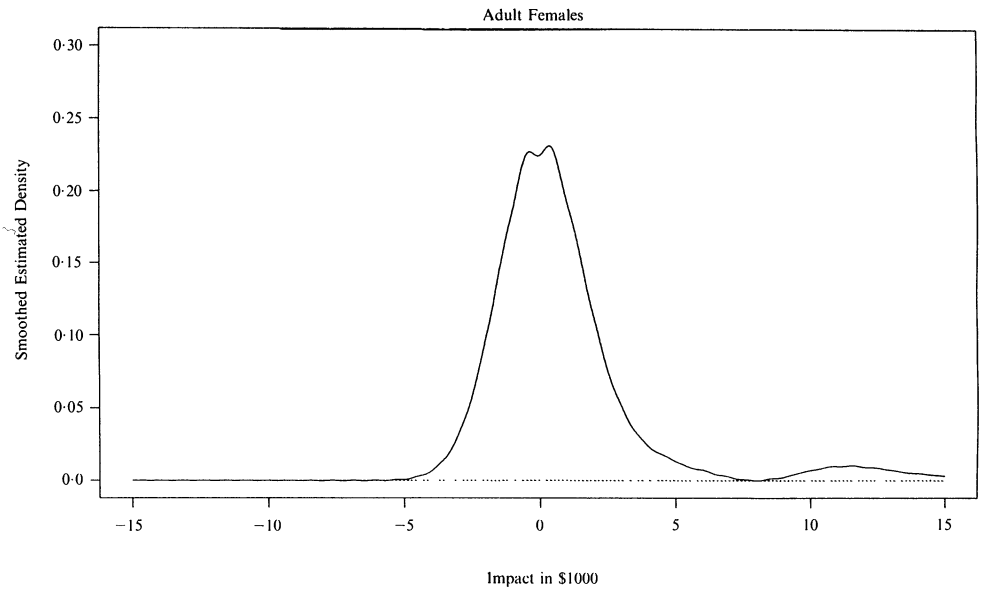


FIGURE 5
Smoothed estimated impact density

deconvolution procedure are given in Appendix C. The bottom row of Table 6 presents parameters calculated from this distribution. The evidence suggests that under assumption (C-1), 43% of adult women were harmed by participating in the programme. The density of Δ is presented in Figure 5. It is clearly non-normal. The estimated variance of the non-parametric gain distribution matches the variance for the gain distribution obtained from the random coefficient model within the range of the sampling error produced from the random coefficient regression model. The fact that we obtain a positive density indicates that (C-1) is not inconsistent with the data. However, in contrast to the results obtained from assuming perfect positive dependence, the average gains are the same at all levels of Y_0 since $Y_0 \perp\!\!\!\perp \Delta \mid D=1$ implies that $E(\Delta \mid Y_0, D=1)$ is the same for all Y_0. This is at odds with the evidence in Figure 2.

Normality is often assumed in implementing the random coefficient model to produce estimated distributions. However inspection of the nonparametric estimate of the density presented in Figure 5 reveals that Δ is not normally distributed. Comparison of the cumulative distribution function estimated by deconvolution, shown in Figure 6, with the CDF for a normal variable having the same variance indicates that the two assumptions are inconsistent. The deconvolution estimate of the distribution of Δ clearly has more mass in the right tail than the estimate based on the normal distribution with the same variance. An appendix available on request documents that these findings are robust to plausible assumptions about measurement error.[30]

8. DECISION PROCESSES AND THE NONPARAMETRIC IDENTIFICATION OF PROGRAMME TREATMENT EFFECTS FROM EXPERIMENTAL AND NON-EXPERIMENTAL DATA

Under other assumptions about the structure of decision processes, and about the variation available in the data, it is possible to use either non-experimental data—the distributions

30. After normal measurement error is removed, the density and distribution of impacts look much more normal, although there is still a spike at zero.

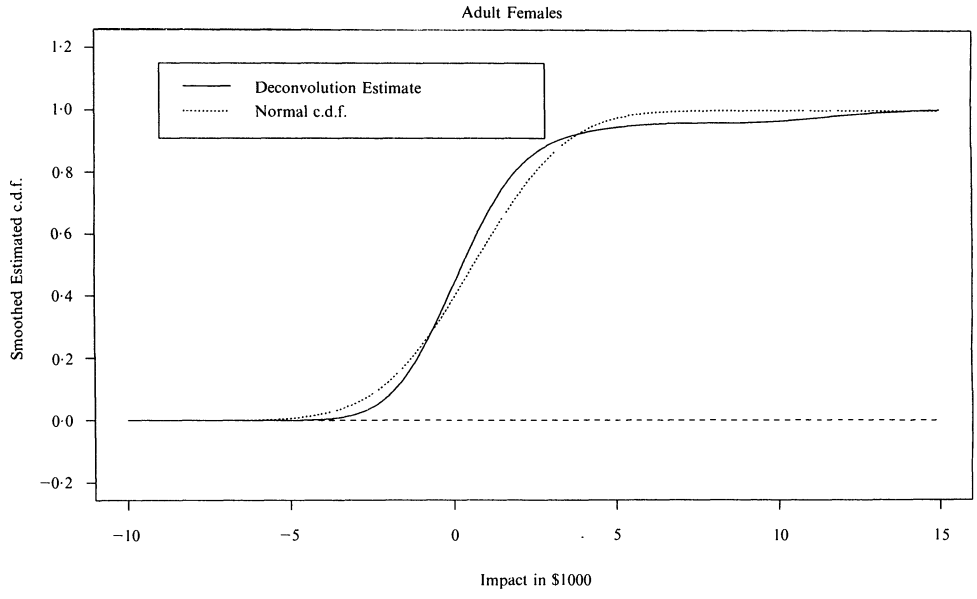


FIGURE 6
Smoothed estimated c.d.f. of impacts and normal c.d.f.

in (1a) and (1b)—or experimental data to recover the joint distribution of outcomes and hence to obtain the distribution of programme impacts. One benefit of such a strategy is that information about self-selection decisions made by participants also provides their revealed preference evaluations of programmes.

Let outcomes Y_1 and Y_0 be written as functions of observed variables X and unobserved variables U_1 and U_0 so that

$$Y_1 = g_1(X_1, X_c) + U_1, \tag{9a}$$

$$Y_0 = g_0(X_0, X_c) + U_0, \tag{9b}$$

where X_1 (a k_1-dimensional vector) and X_0 (a k_0-dimensional vector) are variables unique to g_1 and g_0, respectively, and X_c (a k_c-dimensional vector) includes variables common to the two functions. For notational simplicity, define $X = (X_0, X_1, X_c)$. The variables U_0 and U_1 are unobserved from the point of view of the econometrician.

Let the decision rule for programme participation be given by

$$IN = g_{IN}(X_{IN}, X_c) + U_{IN}, \qquad D = 1(IN \geqq 0), \tag{9c}$$

where X_{IN} consists of observed variables affecting participation, some (or all) of which may appear in X_1 and X_0 and where U_{IN} is unobserved by the econometrician. In this setting, IN is a latent index or net utility. The joint distribution of (U_0, U_1, U_{IN}) is denoted by $F(u_0, u_1, u_{IN})$. These variables are assumed to be statistically independent of X and X_{IN}.

The Roy model is a special case of this framework in which selection into the programme depends only on the gain from the programme. In this case

$$g_{IN} = g_1(X_1, X_c) - g_0(X_0, X_c),$$

and

$$U_{IN} = U_1 - U_0.$$

The joint distribution of (U_1, U_0) is $F(u_1, u_0)$. The following theorem can be proved for the Roy model.

Theorem 1. *Let* $Y_1 = g_1(X_1, X_c) + U_1$ *and* $Y_0 = g_0(X_0, X_c) + U_0$. *Assume*:

(i) $(U_1, U_0) \perp\!\!\!\perp (X_1, X_0, X_c)$;
(ii) $D = 1(Y_1 \geqq Y_0)$;
(iii) (U_1, U_0) *is absolutely continuous with Support* $(U_1, U_0) = R_1 \times R_1$;
(iv) $g_0(X_0, X_c)$: $R_{k_0} \to R_1$ *and* $g_1(X_1, X_c)$: $R_{k_1} \to R_1$ *for each fixed* X_c *and*

$$\textit{Support}\,(g_0(X_0, X_C) | X_C, X_1) = R_1 \quad \textit{for all } X_C, X_1,$$

$$\textit{Support}\,(g_1(X_1, X_C) | X_C, X_0) = R_1 \quad \textit{for all } X_C, X_0,$$

$$\textit{Support}\,(X_0 | X_1, X_C) = \textit{Support}\,(X_0) \quad \textit{for all } X_1, X_0,$$

$$\textit{Support}\,(X_1 | X_0, X_C) = \textit{Support}\,(X_1) \quad \textit{for all } X_0, X_C;$$

(v) *The marginal distributions of* U_1 *and* U_0 *have zero medians.*

Then g_1, g_0 *and* $F(u_1, u_0)$ *are nonparametrically identified from data on programme choices and the outcome distributions, i.e. from* (1*a*), (1*b*) *and* $\Pr(D = 1 | X)$.

Proof. See Heckman and Honoré (1990) or Heckman and Smith (1997).

The content of this theorem is that if there is sufficient variation in X_1, X_0 and X_c, and if we know that programme participation is based solely on outcome maximization, no arbitrary parametric structure on the outcome equations or on the distribution of the unobservables generating outcomes needs to be imposed to recover the full joint distribution of outcomes using ordinary micro data. There is no need to conduct social experiments to answer the distributional questions posed in Section 2. From the information in (1a) and (1b), we can construct the counterfactuals produced from social experiments without running the risk of disruption and randomization bias associated with such experiments that is discussed in Heckman (1992) and Heckman and Smith (1995). Assuming no general equilibrium effects, we can also generalize from a partial coverage distribution to a full coverage distribution.

Social experiments are not required to answer these questions because $F_0(y_0 | D = 1, X)$ is redundant information. Similarly, under the assumptions of a common effect model where $Y_1 = Y_0 + \alpha$ (conditional on X) and $U_0 = U_1$, it is possible to recover the entire joint distribution of outcomes without invoking the explicit income-maximizing assumption in the Roy model.

The assumptions made in Theorem 1 about the supports of $X_1, X_0, g_1, g_0, U_1, U_0$ are made for convenience, in an effort to focus on the main ideas. A version of Theorem 1 can be proved under the following alternative conditions:

(a) The support of (U_0, U_1) is $(\underline{U}_0, \bar{U}_0) \times (\underline{U}_1, \bar{U}_1)$, where $\underline{U}_i$, $\bar{U}_i$, $i = 0, 1$ are, respectively, the finite lower and upper bounds for U_0 and U_1;
(b) The support of (g_0, g_1) is $(\underline{g}_0, \bar{g}_0) \times (\underline{g}_1, \bar{g}_1)$.

Under these conditions, and assuming that all of the other conditions hold, Heckman and Smith (1997) show that it is possible to modify the argument of Theorem 1 and produce a different version of the same basic theorem over a subset of the support.

The Roy model has an unusual structure because the participation rule and the outcome equations are tightly linked. As a consequence, we can recover the full joint

distribution, $F(y_0, y_1 \mid X)$, and the decision rule knowing only the conditional distributions (1a) and (1b) and the participation equation routinely available from cross-section data.

For more general decision rules such as (9c), which break the tight link between outcomes and participation decisions, it is not possible to use (1a) and (1b) to address questions that can only be answered from the full joint distribution of (Y_0, Y_1). Even access to the data obtained from social experiments—distribution (1c)—does not suffice to solve the fundamental evaluation problem that both Y_0 and Y_1 are never observed for the same person. However, a theorem analogous to Theorem 1 can be proved that demonstrates that with sufficient variation in the X variables, it is possible to recover $F_0(y_0 \mid D=1, X)$ from non-experimental data. Thus, it is not necessary to conduct a social experiment to obtain it (see Heckman (1990*a*) and Heckman and Smith (1997)).

Social experiments balance supports

The advantage of social experiments over conventional micro data in the context of Theorem 1 is that they expand the range of the support of the g functions (see Heckman (1996) and the references cited there). Suppose that *Support* $(X \mid D=1) \neq$ *Support* $(X \mid D=0)$. If there are domains of X where there is no common support, Theorem 1 does not apply and the modified version of it presented in Heckman and Smith (1997) must be applied. Randomization guarantees that *Support* $(X \mid D=1, R=1) =$ *Support* $(X \mid D=1, R=0)$, where $R=1$ denotes randomization into the treatment group and $R=0$ indicates randomization into the control group, because persons who would have participated in the programme are now denied access to it. Thus, randomization ensures that the support conditions of Theorem 1 are satisfied for the population of participants. However, it may still happen that the support of X for the population for which $D=1$ is not the same as the support of X for the whole population. Then the models are only identified over the available support. For both experimental and non-experimental data, it may be necessary to sample more widely on X coordinates. Experiments have the advantage that they allow identification of impacts even for persons with values of X such that $\Pr(D=1 \mid X)=1$, which is not possible using non-experimental methods.

9. THE OPTION VALUE OF TRAINING

This section presents estimates from the JTPA data of the option values of training defined in Section 2. Let $F_0(y_0)$ be the distribution of no-training offers in the nonparticipation state. We assume that participants can choose between offers from F_Z and F_0. If participants can inspect their offers from each distribution before choosing between them, $Y_1 = \max(Y_0, Z)$. Four different definitions of the option value under different decision horizons and information structures are given in Section 2.

From the definition of the maximum

$$F_1(y_1 \mid D=1) = F_{0,z}(y_1, y_1 \mid D=1), \tag{10}$$

where $F_{0,z}$ is the joint distribution of Y_0 and Z. We observe Y_1 for persons randomized into the programme, and so we know $F_1(y_1 \mid D=1)$. We observe Y_0 for persons randomized out, so we know $F_0(y_0 \mid D=1)$. We make two different assumptions about the dependence between Y_0 and Z. One case assumes that Y_0 and Z are independent given $D=1$. The other case assumes that $Z = Y_0 + \Delta$, and Δ is independent of Y_0 given $D=1$.

If Y_0 and Z are independent given $D=1$, we can obtain F_z from the formula

$$F_z(z \mid D=1)=\frac{F_1(z \mid D=1)}{F_0(z \mid D=1)} \quad \text{for all values of } Z \text{ such that } F_0(z \mid D=1)\neq 0.$$

In the case of $Z=Y_0+\Delta$ where $(Y_0 \perp\!\!\!\perp \Delta) \mid D=1$, it is convenient to work with the densities, which are assumed to exist. For this case, differentiate (10) to obtain

$$f_1(y_1)=f_0(y_1)F_\Delta(0)+\int_0^{y_1} f_0(y_0)f_\Delta(y_1-y_0)dy_0$$

$$=f_0(y_1)F_\Delta(0)+\int_0^{\infty} f_0(y_1-t)f_\Delta(t)dt.$$

This is an integral equation for f_Δ, which can be easily solved using recursive methods. Note that the second case differs from the apparently similar case analyzed in Section 7 because in this section we assume that Δ is observed before the agent chooses between Z or Y_0. In Section 7, it is assumed that Δ is not observed before programme participation decisions are made.

Table 7 presents estimates of measures of option value expressed in terms of earnings over the 18 months after random assignment. Our data consist of adult women recommended for subsidized jobs at private firms as part of the JTPA programme. The wage subsidy programme gives prospective participants wage offers (Z) that they are free to accept or reject. The final row of Table 7 presents an estimate of the proportion of people who choose to exercise the training option. Standard errors for the estimates in Table 7 are obtained by bootstrapping.

First consider the case where Y_0 and Z are statistically independent. The estimate of option value (OP-1) of \$794 is the impact estimated from the social experiment. The

TABLE 7

Estimated option values from JTPA on-the-job training

(National JTPA Study 18 month impact sample; on-the-job training treatment stream adult females)

Parameter	Y_0 independent of Z given $D=1$	$Z=Y_0+\Delta$ and Δ independent of Y_0 given $D=1$
(OP-1) $E(\max(Y_0, Z) \mid D=1)-E(Y_0 \mid D=1)$	794	794
	(338)	(338)
$E(Y_0 \mid Y_0 \geqq Z, D=1)-E(Y_0 \mid D=1)$	1736	0
	(297)	(0)
(OP-2) $r^{-1}[E(\max(Y_0, Z) \mid D=1)-E(Y_0 \mid D=1)]$	7940	7940
	(3380)	(3380)
(OP-3) $\max(E(Y_0 \mid D=1), E(Z \mid D=1))-E(Y_0 \mid D=1)$	0	N.A.
(OP-4) $E(\max(Y_0, Z) \mid D=1)-\max(E(Y_0 \mid D=1), E(Z \mid D=1))$	794	N.A.
	(338)	
$\Pr(Y_0 \geqq Z \mid D=1)$	0·93	0·93
	(0·04)	(0·10)

1. Bootstrap standard errors appear in parentheses.
2. The "on-the-job training treatment stream" includes persons who were recommended to receive on-the-job training by JTPA case workers prior to random assignment. This group comprises roughly one-third of the adult women in the experiment.
3. "N.A." indicates that the indicated value cannot be calculated because the distribution of Z is not nonparametrically identified.
4. The estimates of (OP-2) are calculated assuming that the interest rate $r=0{\cdot}10$.

second row presents the compositional effect of a rise in the outcome measure among non-participants due to low Y_0 persons accepting the offered Z. The final row reveals that for 93% of the women, unsubsidized jobs give higher earnings than subsidized jobs. However, a small group of women get exceptionally large wage offers from the subsidized job distribution. Option values (OP-1) and (OP-4) are the same because $E(Z \mid D=1) < E(Y_0 \mid D=1)$. The estimate of option value (OP-3) in row four is zero for the same reason. Option value (OP-2) is just the discounted value of (OP-1).

For the second assumption about the joint dependence between Y_0 and Z, (OP-1) is the same as before because it is just the experimental mean. Now, however, the compositional effect reported in the second row is zero because Δ is independent of Y_0 given $D=1$ so $E(Y_0 \mid Y \geqq Z, D=1) = E(Y_0 \mid \Delta < 0, D=1) = E(Y_0 \mid D=1)$. In addition, it is not possible to nonparametrically identify F_Δ for $\Delta < 0$. This makes it impossible to nonparametrically estimate the option values in the fourth and fifth rows. We place much less confidence in

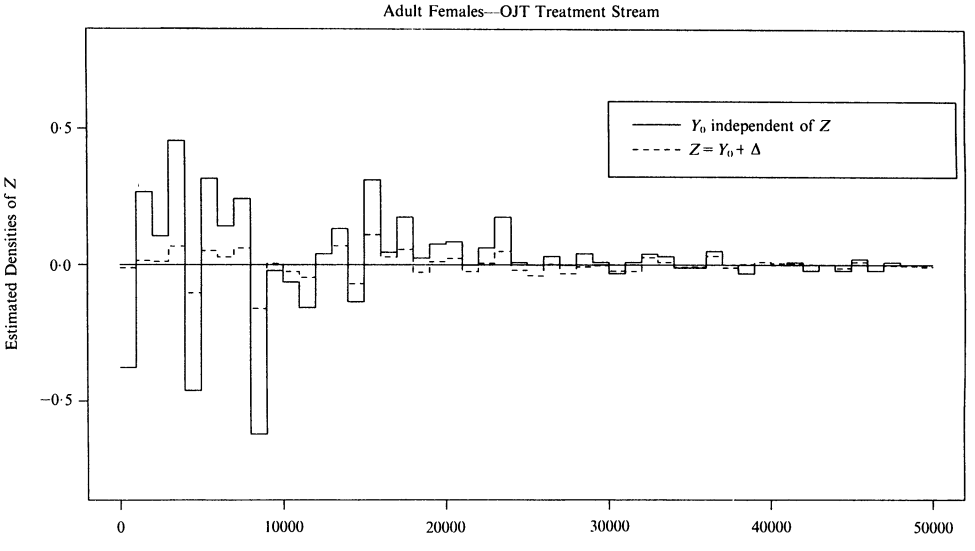


FIGURE 7

Densities of $Z \mid Z > 0$ under option value assumptions

the estimates produced under the second case. As shown in Figure 7, the estimated density of Δ is negative over much of its support, giving little credibility to the identifying assumptions that justify this model. The identifying restriction can be tested and is rejected in our data.[31]

10. SUMMARY AND CONCLUSION

This paper considers the evaluation of a social programme when the responses to the programme differ among otherwise observationally equivalent people. The answers to a variety of important evaluation questions require knowledge of the distribution of programme impacts. Questions of political economy or "social justice" require for their answer

31. Recall the testable restrictions implied by the deconvolution model. Heckman, Smith and Taber (1998) demonstrate that certain estimators used to correct for programme dropouts in social experiments sometimes produce negative outcome densities, which can therefore be used to check the identifying assumptions that support the estimators.

knowledge of the distribution of programme costs and benefits. The conventional focus in the evaluation literature on mean impacts either assumes that distributional issues are irrelevant, or assumes that they are automatically resolved using a social welfare function or a family equivalent. Neither assumption is attractive.

Abstracting from distributional issues, it is also of great interest to examine the impact of social programmes in conferring options on participants, even if all participants do not exercise them. Social programmes may enhance the options of participants. An option may be valued in its own right even if it is an option of limited duration.

Using data from a social experiment, we estimate the distribution of programme impacts. By using experimental data, we can abstract from selection problems that plague nonexperimental data and so can focus on the main issues of this paper. However, even with such data, we cannot escape the fundamental evaluation problem that persons cannot simultaneously participate and not participate in a programme. This precludes direct estimation of impact distributions and requires an indirect procedure using information on the distributions of outcomes for participants and randomized-out nonparticipants. Many distributions of programme impacts are consistent with the available data. Widely used classical probability inequalities are not very informative about the distribution of programme impacts, but they reveal that heterogeneity is an essential feature of impact distributions in our data.

We supplement the information supplied by classical probability theory in two ways. One method makes assumptions about the dependence between potential outcomes in the treatment and non-treatment states. The second method builds models of programme participation that produce, as a by-product, implications about dependence among potential outcomes, supplying the missing information required to identify the impact distributions or to narrow down the uncertainty about them. The second approach also enables us to investigate the question of whether participant perceptions of programme impact are in agreement with other objective measures of programme gains. (This approach is pursued extensively in Heckman and Smith (1997).) Knowledge of the programme participation decision rule is also essential in generalizing the findings from one evaluation to other economic environments.

By analyzing the distribution of programme impacts, we present a more informative evaluation of a programme compared to what is obtained by confining attention solely to mean impacts. We present strong evidence that heterogeneity is an important feature of impact distributions. We define and estimate several distinct measures of the option values produced by a social programme and find some of them to be sizeable while others are negligible. Programmes enhance the choice sets of participants even if the participants do not always choose to exercise the options they provide.

Programme impacts are not uniform across outcomes in the non-participation state. Most women benefit from the programme. Our results on stochastic rationality provide additional corroborating evidence that the programme is beneficial. Randomized-out women have ex post regret. Our more comprehensive consideration of programme impacts produces a more nuanced interpretation of the prototypical training programme we evaluate.

Table 8 summarizes our evidence on the distribution of programme impacts for participants under different identifying assumptions. In all cases, the proportion of adult female participants who benefit is greater than half. Departures from high levels of dependence across potential outcomes produce implausible estimated impacts at the top and bottom quantiles and implausibly large estimated variability. The estimated median impact is also sensitive to alternative identifying assumptions. Our analysis rejects the widely-used common effect model but suggests that substantial departures from our generalization

TABLE 8

Summary of characteristics of the distribution of impacts under alternative assumptions[1]; *adult females*
(National JTPA Study 18 month impact sample)

Identifying assumption	25th Percentile of the impact distribution	Median impact	75th Percentile of the impact distribution	Percent with positive impact ($Y_1 \geqq Y_0$)	Impact standard deviation	Outcome correlation (product-moment correlation of Y_1 and Y_0)
Percentiles of Y_1 and Y_0 perfectly positively dependent (Kendall's rank correlation $\tau = 1{\cdot}00$)	\$572·00 (232·90)[2]	\$864·00 (269·26)[2]	\$966·00 (305·74)[2]	100·00% (1·06)[2]	\$1857·75 (480·17)[2]	0·9903 (0·0048)[2]
Percentiles of Y_1 and Y_0 positively related with Kendall's rank correlation $\tau = 0{\cdot}95$	\$125·50 (124·60)[2]	\$616·00 (280·19)[2]	\$867·00 (272·60)[2]	96·00% (3·88)[2]	\$6005·96 (776·14)[2]	0·7885 (0·0402)[2]
Percentiles of Y_1 and Y_0 independent (Kendall's rank correlation $\tau = 0{\cdot}00$)	−\$18098·50 (630·73)[2]	\$0·00 (163·17)[2]	\$7388·50 (263·25)[2]	54·00% (1·11)[2]	\$12879·21 (259·24)[2]	−0·0147 (0·0106)[2]
Percentiles of Y_1 and Y_0 perfectly negatively dependent (Kendall's rank correlation $\tau = -1{\cdot}00$)	−\$11755·00 (411·83)[2]	580·00 (389·51)[2]	12791·00 (253·18)[2]	52·00% (0·81)[2]	16432·43 (265·88)[2]	−0·6592 (0·0184)[2]
Random coefficient model; $Y_1 = Y_0 + \Delta$ with Δ normally distributed and independent of Y_0 given $D = 1$	−\$919·83	\$601·74	\$2123·31	60·45%	2271·00 (1812·90)[3]	0·9595
Empirical deconvolution; $Y_1 = Y_0 + \Delta$ with Δ independent of Y_0 given $D = 1$	−\$850·00	\$280·00	\$1530·00	56·35%	1675·00	0·9774

[1] Note that the mean impact is the same regardless of the assumptions made to identify the joint distribution because it depends only on the known marginal distributions of Y_1 and Y_0.
[2] Bootstrap standard error in parentheses.
[3] Estimated standard error in parentheses.

of it—perfect positive dependence across potential outcome distributions—produces estimates of impact distributions that are not credible.[32]

APPENDIX

A. Data

The data analyzed in this paper were collected as part of an experimental evaluation of the training programmes financed under Title IIA of the Job Training Partnership Act (JTPA). The experiment was conducted at a sample of sixteen JTPA training centres around the United States. Data were gathered on JTPA applicants randomly assigned to either a treatment group allowed access to JTPA training services or to 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 centre. A total of 20,601 persons were randomly assigned. In this paper we only present results for women age 22 or more at the time of random assignment.

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 programmes, schooling, and training during the period after random assignment. The response rate for this survey was around 84%. 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 subsample includes 5725 adult women.

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 here as they were not available at the time this paper was written. The employment measure used in this paper is based on the 16th, 17th and 18th months after random assignment. A person is defined to be employed if she had any self-reported earnings in these months.

B. Description of algorithms used for permutations

This appendix describes the algorithms used to produce Tables 1A, 1B, 5A and 5B. These tables involve samples of impact distributions obtained by permuting the percentiles of the control and treatment outcome distributions.

Tables 1A and 1B

The total number of possible impact distributions obtained by percentile permutation consistent with the observed control and treatment group outcome distributions is 100. As it is computationally infeasible to construct all of these permutations, in Tables 1A and 1B we instead report results from a random sample of 100,000 of these permutations. Each permutation is obtained by taking a set of uniform random deviates, indexing them, sorting them, and then using the order of the sorted indices to permute the percentiles of the control group outcome distribution. The permuted control group percentiles are then subtracted from the treatment group percentiles to produce the impact distribution corresponding to the given permutation.

The percentiles of each impact distribution generated in this manner are then retained until the full sample of 100,000 has been generated. The mean and percentiles of the generated percentiles and other parameters of the impact distributions corresponding to the sample of permutations are reported in Tables 1A and 1B. The bootstrap standard errors are generated by repeating this process 50 times with 50 different random samples of 100,000 permutations and 50 bootstrap samples of earnings observations and then calculating the empirical standard deviations of the generated parameters.

Tables 5A and 5B

For Tables 5A and 5B, the algorithm of Hibbard (1963), cited in Knuth (1973), was used to draw random samples of permutations conditional on a particular value of τ (number of inversions). Fifty permutations were generated for each value of τ shown in the tables, except for $\tau = 1{\cdot}0$ and $\tau = -1{\cdot}0$, where only the single

32. Heckman and Robb (1985, 1986), Heckman (1997*b*) and Heckman and Smith (1997) consider the implications of heterogeneity of programme impacts on conventional nonexperimental econometric evaluation methods.

permutation with the indicated value of τ was used. The tables report the median parameter value among these 50 permutations. Bootstrap standard errors were obtained by repeating the process 50 times, each time with a new sample of permutations and a bootstrap sample of earnings values.

C. Construction of impact distribution through deconvolution

This appendix describes how to implement condition (C-1) and equation (7) to estimate the distribution of impacts through deconvolution. Deconvolution is discussed in Kendall and Stuart (1977), Stefanski and Carroll (1990) and Carroll and Hall (1988). In a useful paper, Horowitz and Markatou (1996) extend Stefanski and Carroll (1990) by accounting for estimation of β and $\bar{\Delta}$ and present an economic application. Heckman (1990*b*) discusses the use of deconvolution in evaluating social programmes.

We deconvolve the sum of self-reported earnings in the 18 months after random assignment. The Y values are divided by 1000 in order to reduce the loss of precision during exponentiation. The empirical characteristic functions for ε and ξ, where $\xi=\Delta+\varepsilon$, were obtained from the control and treatment group earnings values, respectively. The standard formulae are $\hat{\phi}_\varepsilon(\lambda)=(N_c)^{-1}\sum_j \exp(i\lambda Y_j)$ for the control group and $\hat{\phi}_\xi(\lambda)=(N_t)^{-1}\sum_j \exp(i\lambda Y_j)$ for the treatment group, where N_t is the number of observations in the treatment group, N_c is the number of observations in the control group, and where estimates of the value of the characteristic function were obtained at 10001 equally spaced frequencies λ between $-5{\cdot}00$ and $5{\cdot}00$. These limits on the frequencies were determined by examining a number of possible choices involving both wider and narrower ranges, and trading off the numerical problems induced by including higher frequencies against the loss of accuracy due to narrowing the range of included frequencies.

Estimates of the value of the characteristic function of Δ, $\phi_\Delta(\lambda)$, were then obtained at the same 10001 values of λ by taking the ratio of the characteristic functions for ξ and ε: $\hat{\phi}_\Delta(\lambda)=\hat{\phi}_\xi(\lambda)/\hat{\phi}_\varepsilon(\lambda)$. These estimates were used to obtain estimates of $F(\Delta)$, the c.d.f. of the impact distribution, using the relation

$$F(\Delta)=0{\cdot}5+(\pi)^{-1}\int_0^\infty [\mathrm{Re}\,(\phi_\Delta(\lambda))\cos(\Delta\lambda)-\mathrm{Im}\,(\phi_\Delta(\lambda))\cos(\Delta\lambda)]d\lambda,$$

where Re ( ) denotes the real portion of a complex argument, Im ( ) denotes the imaginary portion of a complex argument, and where the numerical integration is carried out over the 10001 values of λ using the trapezoidal rule. Estimates of $F(\Delta)$ were obtained for 2001 evenly spaced values of Δ between -50 and 50. The estimated p.d.f. obtained from this procedure appears in Figure 5 and the corresponding c.d.f. appears in Figure 6. The p.d.f. shown in the figure has been smoothed using standard kernel smoothing techniques. We employ the empirical characteristic functions and the c.d.f. rather than estimating the characteristic functions over smoothed data distributions and obtaining the estimated impact density directly because the earnings data contain an important point mass at zero.

Our experience with this procedure is that marginal increases or decreases in the number of frequencies, λ, used to approximate the characteristic functions have little effect on the substance or stability of the estimates. Using half as many points in each case produces about the same estimate. We test the sensitivity of our estimates to the weighting of the frequencies, λ, used in the estimation by implementing the smoothing function proposed in Horowitz and Markatou (1996). In our implementation of their smoother, the empirical characteristic function corresponding to the impact density is estimated using the method previously presented. We then select a smoothing function, ϕ_K, in this case the characteristic function of a standard normal random variable. This function is a bounded, real characteristic function with support $[-1, 1]$. We estimate the c.d.f. of the impacts using a modified version of the formula given earlier

$$F(\Delta)=0{\cdot}5+(\pi)^{-1}\int_0^\infty [\mathrm{Re}\,(\phi_\Delta(\lambda)\phi_K(h_n\lambda))\cos(\Delta\lambda)-\mathrm{Im}\,(\phi_\Delta(\lambda)\phi_K(h_n\lambda))\cos(\Delta\lambda)]d\lambda,$$

where h_n is a sequence of positive constants satisfying $h_n\to 0$ as $n\to\infty$ where $n=\min(N_t, N_c)$. Table C-1 shows the estimates of the impact mean and standard deviation obtained for various values of h_n when $F(\Delta)$ was evaluated at 5001 evenly spaced values of Δ between -50 and 50. While the estimated mean is quite stable across values of h_n, the estimated impact standard deviation is somewhat sensitive to the value of the smoothing parameter, with the estimates differing by almost a factor of three between the cases where $h_n=1{\cdot}00$ and $h_n=3{\cdot}50$.

Finally, we find that applying our original procedure to the empirical characteristic function of one of the data distributions, rather than to the ratio of the two empirical characteristic functions, produces estimates that are both reliable and stable. Problems in the form of negative estimated densities and/or substantial instability

TABLE C-1

Estimated impact mean and standard deviation empirical deconvolution with weighted frequencies for various values of h_n

(National JTPA Study 18 month impact sample; adult females)

Value of h_n	Estimated mean impact	Estimated impact std dev
$h_n = 1{\cdot}00$	651·63	1490·31
$h_n = 1{\cdot}25$	645·55	1252·18
$h_n = 1{\cdot}50$	642·17	1100·10
$h_n = 1{\cdot}75$	639·96	993·23
$h_n = 2{\cdot}00$	638·19	908·67
$h_n = 2{\cdot}25$	636·54	834·41
$h_n = 2{\cdot}50$	634·89	765·17
$h_n = 2{\cdot}75$	633·20	699·58
$h_n = 3{\cdot}00$	631·48	638·16
$h_n = 3{\cdot}50$	628·10	532·43

1. The characteristic function of a standard normal random variable was used as the weighting function in generating these estimates.

to choice of the range of λ only arise when using the ratio of the characteristic functions from the two data distributions to estimate the characteristic function of the impact distribution. These problems may result either from the failure of the independence assumption to hold, in which case the negative density estimates may be correct, or from numerical factors associated with taking nonlinear functions of the ratio of two relatively small numbers. Numerical problems of the latter type are commonly reported in the literature on deconvolution (see, e.g. Jansson (1984)).

D. Allowing for mass points at zero in the population

In many cases, it is plausible that there are mass points at zero for Y_1 and Y_0. This is the case with the JTPA earnings data used in this study. (Obviously the mass points may be at some place other than zero, may be different for Y_1 than for Y_0, and there may be multiple mass points. We consider only the simplest case in this appendix). Our analysis of this case combines the analyses in Sections 4 and 6. However, a new result is required because it is necessary to match the zeros for one outcome measure with the continuous outcome components for the other.[33]

Define the following notation: Let $\Pr(Y_1=0, D=1)=P_{0\cdot}>0$ and $\Pr(Y_0=0, D=1)=P_{\cdot 0}>0$. The density of Y_1 for $Y_1>0$ is

$$f(y_1 \mid Y_1>0, D=1)$$

while the density of Y_0 for $Y_0>0$ is

$$f(y_0 \mid Y_0>0, D=1).$$

In constructing bounds for the joint distribution of (Y_0, Y_1) we must allow for $Y_0=0$ to be paired with continuous Y_1, for $Y_1=0$ to be paired with continuous Y_0, and for Y_1 and Y_0 to be either both discrete or both continuous. The following three step method generalizes the procedures used in Sections 4 and 6 in the text:

Step 1. Using the methods of Section 6, bound the joint distribution of the indicators of positive earnings. Let $E_0=1$ if earnings $Y_0>0$; $E_0=0$ otherwise. Let $E_1=1$ if $Y_1>0$; $E_1=0$ otherwise. Then define:

$$P_{11}=\Pr(Y_1>0, Y_0>0 \mid D=1)=\Pr(E_1=1 \text{ and } E_0=1),$$

$$P_{10}=\Pr(Y_1>0, Y_0=0 \mid D=1)=\Pr(E_1=1 \text{ and } E_0=0),$$

$$P_{01}=\Pr(Y_1=0, Y_0>0 \mid D=1)=\Pr(E_1=0 \text{ and } E_0=1),$$

$$P_{00}=\Pr(Y_1=0, Y_0=0 \mid D=1)=\Pr(E_1=0 \text{ and } E_0=0).$$

33. To do this, we use well-known ideas in the probability literature. See Rachev (1985).

We know the left-hand sides of the following equations but the available population information does not afford a further resolution into the components on the right-hand side: $P_{1\cdot}=P_{10}+P_{11}=\Pr(Y_1>0\mid D=1)$, $P_{0\cdot}=1-P_{1\cdot}$, $P_{\cdot 1}=P_{01}+P_{11}=\Pr(Y_0>0\mid D=1)$ and $P_{\cdot 0}=1-P_{\cdot 1}$.

Following the procedure outlined in Section 6, we can represent all of the possible 2×2 tables with fixed marginals by varying Q over the interval $[-1, 1]$. Each value of Q produces unique values for P_{ij}, $i,j=0, 1$.

Step 2. Next derive bounds on

$$f(y_1\mid Y_1>0, Y_0=0, D=1) \quad \text{and} \quad f(y_0\mid Y_1=0, Y_0>0, D=1).$$

We know the left-hand sides of the following equations:

$$f(y_1\mid Y_1>0)=f(y_1\mid Y_1>0, Y_0=0, D=1)\frac{P_{10}}{P_{11}+P_{10}}$$
$$+f(y_1\mid Y_1>0, Y_0>0, D=1)\frac{P_{11}}{P_{11}+P_{10}}, \tag{D-1}$$

and

$$f(y_0\mid Y_0>0)=f(y_0\mid Y_0>0, Y_1=0, D=1)\frac{P_{01}}{P_{11}+P_{01}}$$
$$+f(y_0\mid Y_0>0, Y_1>0, D=1)\frac{P_{11}}{P_{11}+P_{01}}, \tag{D-2}$$

where the weights on the densities are given by specifying Q in Step 1.

We can construct $f(y_1\mid Y_1>0, Y_0=0, D=1)$ by weighting $f(y_1\mid Y_1>0, D=1)$:

$$f(y_1\mid Y_1>0, Y_0=0, D=1)=f(y_1\mid Y_1>0, D=1)w_1(y_1\mid Y_1>0),$$

where $w_1(y_1\mid Y_1>0)\geqq 0$ and

$$1=\int_0^\infty f(y_1\mid Y_1>0, D=1)w_1(y_1\mid Y_1>0)dy_1,$$

must be satisfied. Similarly we can construct

$$f(y_0\mid Y_1=0, Y_0>0, D=1)=f(y_0\mid Y_0>0, D=1)w_0(y_0\mid Y_0>0),$$

with the requirements that $w_0(y_0\mid Y_0>0)\geqq 0$ and

$$1=\int_0^\infty f(y_0\mid Y_0>0, D=1)w_0(y_0\mid Y_0>0)dy_0.$$

To ensure consistency with (D-1) and (D-2), it is required that the weights satisfy:

$$f(y_1\mid Y_1>0, Y_0>0)=\frac{f(y_1\mid Y_1>0)(1-w_1P_{10}/(P_{11}+P_{10}))}{(P_{11}/(P_{11}+P_{10}))}, \tag{D-3}$$

$$f(y_0\mid Y_1>0, Y_0>0)=\frac{f(y_0\mid Y_0>0)(1-w_0P_{01}/(P_{11}+P_{01}))}{(P_{11}/(P_{11}+P_{01}))}. \tag{D-4}$$

It is easy to verify that the left-hand sides integrate to one over the full supports of Y_1 and Y_0, respectively. For them to be proper densities, it is required for (D-3) that for $P_{10}>0$

$$\frac{P_{11}}{P_{10}}+1\geqq w_1\geqq 0,$$

for all y_1 in the support of Y_1 and for (D-4) that for $P_{01}>0$

$$\frac{P_{11}}{P_{01}}+1\geqq w_0\geqq 0,$$

for all y_0 in the support of Y_0. When $P_{10}=P_{01}=0$, (D-3) and (D-4) simplify in an obvious way. These conditions bound the amount of the mass that can be transferred to one part of the distribution from the other parts.

Moreover, the pairs of weights $(w_1, 1-w_1P_{10}/(P_{11}+P_{10}))$ and $(w_0, 1-w_0P_{01}/(P_{11}+P_{01}))$ bear a reciprocal relationship within each pair. For example, weighting $f(y_1 \mid Y_1>0)$ by placing more mass at the low values of y_1 to obtain $f(y_1 \mid Y_1>0, Y_0=0)$, so that zero values of Y_0 are associated with low values of Y_1, necessitates placing more mass at the high values of y_1 to obtain $f(y_1 \mid Y_1>0, Y_0>0, D=1)$. Independence is captured by selecting $w_0=1$ and $w_1=1$.

Two weighting schemes can be ordered in terms of their positive dependence by the amount of mass they transfer near the origin. Thus for all values of $y_1 \in (0, \varepsilon)$, w_1^* induces more positive dependence in the interval than w_1^{**} if $w_1^* > w_1^{**}$. This ordering can be defined more generally by noting that w_1^* induces more positive dependence than w_1^{**} if

$$\int_0^{\varepsilon} f(y_1 \mid Y_1>0, D=1)w_1^*(y_1 \mid Y_1>0)dy_1 \geqq \int_0^{\varepsilon} f(y_1 \mid Y_1>0, D=1)w_1^{**}(y_1 \mid Y_1>0)dy_1.$$

If this relationship is true for all $\varepsilon \in Support(Y_1)$, then w_1^* is a uniformly more positively dependent weighting scheme than w_1^{**}. In that case the random variable induced by w_1^* is stochastically smaller than the random variable induced by w_1^{**}.

Step 3. Use (D-3) and (D-4) as the marginals for the permutation procedure presented in Section 4.[34]

Proceeding in this fashion, we can vary Q, (w_1, w_0), and the inversion classes of τ for $Y_1>0$, $Y_0>0$ to produce ranges of values on the joint distribution of (Y_1, Y_0). Table D-1 (available on request) presents the joint distribution of (E_1, E_0) for adult women for selected values of Q. This table gives results for the first step of the three step procedure. As Q increases, the probability of a favourable outcome from the programme increases. For all values of Q, the programme produces net earnings gains in the sense that $P(Y_1>0, Y_0=0)>P(Y_1=0, Y_0>0)$.

Table D-2 (available on request) presents the empirical results from Steps 2 and 3. We parameterize the weighting function for stage two in the following different ways:

w_-: point mass placed at opposite extremes (i.e. for $Y_0=0$, place as much mass as possible at the extreme upper values of Y_1; for $Y_1=0$, place as much mass as possible at extreme upper values of Y_0).
w_p: independence, with $w_0=w_1=1$.
w_p: $w_0, w_1 \propto a+bq$ with $a=1$, $b=3$. (Denoted w_p in the table.)
w_+: point mass placed at the same extremes (i.e. for $Y_0=0$, place as much mass as possible near $Y_1=0$; for $Y_1=0$ place as much mass as possible near $Y_0=0$).

A summary of Table D-2 is as follows. As the weights range from w_- to w_+, more of the mass of the Y_0 given $Y_1=0$ and Y_1 given $Y_0=0$ distributions is concentrated near zero. The mean values of Y_1 and Y_0 conditional on $(Y_1>0, Y_0>0)$ must rise as a consequence of (D-3) and (D-4). As $w_1(y_1 \mid Y_1>0)$ decreases for higher values of y_1, the mass of $f(y_1 \mid Y_1>0, Y_0>0)$ necessarily increases in the upper tail. Similar remarks apply to the behaviour of $f(y_0 \mid Y_0>0)$ as higher values are downweighted. For virtually the entire range of dependence parameters, the median impact is positive. At the same time, the median impact is never greater than \$1100 for the full eighteen month period. Second, unless a very high positive value is specified for the continuous outcome measure of dependence across potential outcomes—the τ parameter—the interquartile range on the impact distribution is very large and not credible. This finding supports the conclusions for the more restrictive analysis reported in the text. Third, for most configurations of the dependence parameters, more persons benefit from participation than nonparticipation.

Summary statistics of the overall impact distribution are presented in Table D-3 (available on request). This distribution is formed by combining the three types of conditional distributions. For no combination of values of the dependence parameters are a majority of women harmed by participating in the programme. Yet for some isolated values, a majority do not gain. For some configurations of the dependence parameters, as many as 20% of the women do not change their status by participating in the programme. The interquartile range is plausible only for high values of Q and τ and for weighting functions w_p and w_+. The median gain ranges from −455 to 714. For virtually all configurations with positive dependence parameters, the median programme impact is positive.

E. Distributions of the bounds, statistics derived from the bounding distributions, and the behaviour of the bootstrap

This appendix considers two topics: (i) The derivation of the asymptotic distribution of the bounds for the 2×2 table analyzed in Section 3(c) and a Monte Carlo study of coverage probabilities for the Fréchet-Hoeffding

34. A more general scheme, which we do not use in this paper, employs all Markov operators that satisfy the conditions stated in the text rather than using the permutation matrix.

bounds based on bootstrapped standard errors; and (ii) a Monte Carlo analysis of the distribution of VAR(Δ) as estimated from the bounds in the continuous case, and the coverage performance of bootstrapped standard errors. We consider each topic in turn.

(i) 2 × 2 Table

Upper Bound for $P_{\cdot E}$

Under very general conditions, the sample counterparts to $P_{\cdot E}$ and $P_{E\cdot}$ are asymptotically normal. Using the experimental data, the two estimates are independently distributed because they come from independent samples. Let N_1 and N_0 be the sample sizes in the treatment and control distributions and $\hat{P}_{E\cdot}$ and $\hat{P}_{\cdot E}$ be, respectively, the proportions employed in the treatment and control samples. Let $X_1 = \hat{P}_{E\cdot}$, $X_0 = \hat{P}_{\cdot E}$, $\mu_1 = P_{E\cdot}$, and $\mu_0 = P_{\cdot E}$. In large samples, $X_1 - \mu_1 \sim N(0, \sigma_1^2)$ and $X_0 - \mu_0 \sim N(0, \sigma_0^2)$, where $\sigma_1^2 = (P_{E\cdot})(1 - P_{E\cdot})/N_1 = \mu_1(1-\mu_1)/N_1$ and $\sigma_0^2 = (P_{\cdot E})(1 - P_{\cdot E})/N_0 = \mu_0(1-\mu_0)/N_0$.

Let $Z = \min(X_1, X_0)$. Then because X_1 and X_0 are independent, the distribution of Z is the distribution of the minimum of two independent normal random variables. The distribution of Z is the distribution for the upper bound. Routine calculations reveal that

$$g(z) = \frac{1}{\sigma_0}\varphi\left(\frac{z-\mu_0}{\sigma_0}\right)\Phi\left(\frac{\mu_1 - z}{\sigma_1}\right) + \frac{1}{\sigma_1}\varphi\left(\frac{z-\mu_1}{\sigma_1}\right)\Phi\left(\frac{\mu_0 - z}{\sigma_0}\right), \tag{E-1}$$

where φ is the density of a standard normal and Φ is its c.d.f.

As $\mu_1 \to 1$, the distribution of Z converges to the distribution of X_0. As $N_1, N_0 \to \infty$, σ_0 and σ_1 both converge to zero, and if $\mu_1 > \mu_0$, $P \to 1$, $E(Z) \to \mu_0$ and $\text{VAR}(Z) \to \sigma_0^2$. Since a parallel analysis can be conducted for $\mu_1 < \mu_0$, a normal approximation to the distribution of Z becomes better as N becomes larger unless $\mu_1 = \mu_0$.

Lower bound for $P_{E\cdot}$

We seek the distribution of max $(\hat{P}_{E\cdot} + \hat{P}_{\cdot E} - 1, 0)$. Using the previous notation, define

$$J = (\hat{P}_{E\cdot} + \hat{P}_{\cdot E} - 1 - (P_{E\cdot} + P_{\cdot E} - 1)).$$

In large samples $J \sim N(0, \sigma_0^2 + \sigma_1^2)$. Let $\sigma_q^2 = \sigma_0^2 + \sigma_1^2$ and let $T = \max(J, 0)$. Then T has the density

$$g(t) = f(t \mid J > 0)\ \Pr(J > 0)1(J > 0) + \Pr(J < 0)1(J < 0)$$

$$= \frac{1}{\sigma_q}\varphi(t/\sigma_q)1(J > 0) + \Phi\left(\frac{1 - \mu_0 - \mu_1}{\sigma_q}\right)1(J < 0). \tag{E-2}$$

Observe that as μ_0 and μ_1 become large, or as N_0, $N_1 \to \infty$ when $\mu_0 + \mu_1 > 1$, the density closely approximates a normal density.

We now present a Monte Carlo analysis of the validity of bootstrap standard errors as a guide to summarizing the sampling variability of the Fréchet–Hoeffding bounds for the 2 × 2 table. We conduct an analysis for two sets of population parameter values. In the first set, the population cell probabilities are $P_{EE} = P_{EN} = P_{NE} = P_{NN} = 0{\cdot}25$. In the second set, the population cell probabilities are $P_{EE} = P_{NN} = 0{\cdot}25$, $P_{EN} = 0{\cdot}45$, $P_{NE} = 0{\cdot}05$. In the first case, the population values of T would be zero even without imposing the non-negativity condition $T = \max(J, 0)$. An analysis of this case explores how the coverage probabilities and distribution of the bootstrap standard errors are affected by imposition of the constraint even though it does not bind. In the second case the non-negativity constraint is binding. The constraint at zero is binding for the lower bound of P_{NE}, while the population value of the lower bound for the P_{EN} cell is 0·40.

For both cases, 1000 samples were drawn. Each sample consists of a treatment sample and a control sample, with the probabilities of employment in each sample given by the population probabilities from the 2 × 2 table. Each sample has 1500 control group members and 3000 treatment group members. These samples are roughly the same size as those used in the empirical work described in the text.

For each of the 1000 samples, 250 bootstrap samples are drawn. The bootstrap samples are the same size as the original sample of data but are created using the employment proportions from the sample as estimates of the probability of employment, rather than the population employment probabilities. In addition, 250 Monte Carlo samples are drawn from each data sample. The Monte Carlo samples are the same size as each of the data samples and are based on the population employment probabilities. Bootstrap and Monte Carlo standard errors are calculated for each of the 1000 data samples. Mimicking the practice that is rigorously justified when asymptotic normality is justified, bootstrap and Monte Carlo coverage indicators for each sample are calculated by constructing an interval centred at the sample estimate of each bound and extending 1·96 times the estimated

standard error on each side. For each interval, an indicator variable is set to one if it contains the true parameter value and zero if it does not.

Table E-1 displays the results of the Monte Carlo analysis. The upper panel displays results for the first set of population cell probabilities and the lower panel displays the results for the second set. The first column lists the parameter being bounded and the second column gives the population bound. The third column gives the mean and, in parentheses, the *standard deviation* of the Monte Carlo standard errors over the 1000 data samples. The fourth column gives the mean and, in parentheses, the *standard deviation* of the bootstrap standard errors. The fifth column gives standard errors calculated using the mean Monte Carlo standard errors from 1000 "large" samples containing 5000 control and 10,000 treatment observations, but adjusting the variance upward by the square root of the ratio of the sample sizes. These adjustments assume that the large sample standard errors are correct, and that when scaled up they provide a reliable guide to the smaller sample standard errors. The evidence presented in the tables suggests that this assumption is valid. The sixth column presents standard errors under the assumption that the bound is normally distributed, thus ignoring the nonnegativity requirement for the lower bound and the presence of a second proportion in the upper bound. The final column gives the asymptotic standard errors derived earlier in this appendix.

The main lessons from Tables E-1 are as follows. First, in general, the mean of the Monte Carlo standard errors is very close to the mean of the bootstrap standard errors. In addition, for the lower bounds, where the asymptotic distribution is normal away from the constraint and censored normal close to the constraint, both the bootstrap and Monte Carlo standard errors are very close to the asymptotic standard errors. In both cases, they are also very close to the standard errors calculated using the "large" sample standard errors as asymptotic standard errors. Thus, on average, the bootstrap performs quite well.

Second, the standard deviations for both the bootstrap and Monte Carlo standard errors depend on whether or not the true parameter value is near the constraint in the case of the lower bound and on whether or not the two arguments in the bound formulae are the same or different in the case of the upper bound. Consider each case in turn. In situations where the population lower bound is zero, as in the upper panel of Table E-1 and for the lower bounds on P_{EE} and P_{NN} in the lower panel, the *standard deviations* of the bootstrap standard errors are six or seven times as large as the standard deviations of the Monte Carlo standard errors. In situations where the constraint is strongly binding, as with the lower bound on P_{NE} in the lower panel, both sets of standard errors have a point mass at zero, as the constraint also binds in every bootstrap and Monte Carlo sample. In the case where the population value of the bound is distant from the constraint, as in the lower bound on P_{EN} in the lower panel of Table E-1, the bootstrap standard errors are no more variable than the Monte Carlo ones. With regard to the upper bound estimates, the relative variability depends on whether or not the two arguments in the bound formula are the same. When they are, the bootstrap standard errors are relatively larger, while the variabilities are equal when $P_{\cdot E}=\mu_0 \neq P_{E\cdot}=\mu_1$. This can be seen by comparing the upper bounds on P_{EN} and P_{NE} in the lower panel of Table E-1, where $P_{E\cdot}=P_{\cdot E}$, to the upper bounds on P_{EE} and P_{NN}, where they are different. Finally, in results not shown here, increasing the sample size does not change the relative performance of the bootstrap and Monte Carlo standard errors.

Evidence on the coverage probabilities of the bootstrap is presented in Table E-2 for the case when the 1·960 rule is used to approximate a 95% confidence interval. For the lower bounds, the performance of both the bootstrap and Monte Carlo confidence intervals depends on where the population value of the parameter lies relative to the constraint. When the population value of the bound lies on the constraint, as in the upper panel of Table E-2 and for P_{EE} and P_{NN} in the lower panel, the bootstrap coverage probabilities are about 2% too high and the Monte Carlo coverage probabilities are about 2% too low. When the constraint is strongly binding, as for P_{NE} in the lower panel of Table E-2, both coverage probabilities are one, again because the constraint binds on every bootstrap and Monte Carlo sample. When the population parameter value is distant from the constraint, as for P_{EN} in the lower panel of Table E-2, both coverage probabilities are very close to the value of 0·95 that is assumed in standard applications of the bootstrap.

As is true for the variability in the standard errors, the coverage probabilities for the upper bound parameters depend on the distance between $P_{E\cdot}$ and $P_{\cdot E}$. When they are the same, as in the upper panel of Table E-2 and for P_{EN} and P_{NE} in the lower panel, both the bootstrap and Monte Carlo coverage probabilities are too low: the former by 1 to 3% and the latter by 5 or 6%. In contrast, when the two arguments in the bound formula differ, both coverage probabilities are very close to 0·95.

It is also of interest to consider the shapes of the distributions of the bootstrap and Monte Carlo standard errors. In all cases, the distribution of the Monte Carlo standard errors is approximately normal in shape. This is true regardless of where the population parameter lies relative to the constraint for the lower bounds and regardless of whether the two arguments in the upper bound formula are the same or different. In contrast, the shape of the distribution of bootstrap standard errors depends strongly on these two factors. There are three cases. When the population parameter value lies on the constraint, as for all four lower bounds in the case of equal population cell probabilities, the distribution is what might be described as uniform with rounded corners.

TABLE E-1

Monte Carlo analysis of standard errors for Frechet–Hoeffding bounds

(Estimates based on 1000 data samples; 250 Monte Carlo and 250 bootstrap samples per data sample; sample size of 1500 controls and 3000 treatments)

Bound	Population value of bound	Mean Monte Carlo std. err.[2]	Mean bootstrap std. err.[3]	Large N std. err.[4]	Uncorrected normal std. err.[5]	Correct asymptotic std. err.[6]
Population cell probabilities are $P_{EE}=P_{EN}=P_{NE}=P_{NN}=0{\cdot}25$						
Lower bound on P_{EE}	0·0	0·0092 (0·0006)	0·0088 (0·0043)	0·0092	0·0158	0·0092
Lower bound on P_{EN}	0·0	0·0092 (0·0006)	0·0088 (0·0042)	0·0092	0·0158	0·0092
Lower bound on P_{NE}	0·0	0·0092 (0·0006)	0·0090 (0·0042)	0·0092	0·0158	0·0092
Lower bound on P_{NN}	0·0	0·0092 (0·0006)	0·0091 (0·0042)	0·0092	0·0158	0·0092
Upper bound on P_{EE}	0·5	0·0092 (0·0004)	0·0097 (0·0014)	0·0092	0·0092	0·0092
Upper bound on P_{EN}	0·5	0·0092 (0·0004)	0·0098 (0·0013)	0·0092	0·0092	0·0092
Upper bound on P_{NE}	0·5	0·0092 (0·0004)	0·0098 (0·0014)	0·0092	0·0092	0·0092
Upper bound on P_{NN}	0·5	0·0092 (0·0004)	0·0098 (0·0013)	0·0092	0·0092	0·0092
Population cell probabilities are $P_{EE}=P_{NN}=0{\cdot}25$, $P_{EN}=0{\cdot}45$, $P_{NE}=0{\cdot}05$						
Lower bound on P_{EE}	0·0	0·0083 (0·0006)	0·0085 (0·0038)	0·0084	0·0145	0·0085
Lower bound on P_{EN}	0·4	0·0144 (0·0006)	0·0144 (0·0006)	0·0144	0·0145	0·0145
Lower bound on P_{NE}	0·0[1]	0·0000 (0·0000)	0·0000 (0·0000)	0·0000	0·0145	0·0000
Lower bound on P_{NN}	0·0	0·0084 (0·0006)	0·0081 (0·0038)	0·0084	0·0145	0·0085
Upper bound on P_{EE}	0·3	0·0118 (0·0005)	0·0118 (0·0005)	0·0118	0·0118	0·0118
Upper bound on P_{EN}	0·7	0·0085 (0·0004)	0·0090 (0·0013)	0·0085	0·0084	0·0085
Upper bound on P_{NE}	0·3	0·0084 (0·0004)	0·0088 (0·0012)	0·0084	0·0084	0·0085
Upper bound on P_{NN}	0·3	0·0083 (0·0004)	0·0083 (0·0004)	0·0084	0·0084	0·0084

1. The non-negativity constraint on the lower bound is binding in this case.
2. Monte Carlo standard errors are calculated for each data sample by drawing 250 additional samples using the population cell probabilities and calculating the standard deviation of each parameter across these samples. The table reports the mean and, in parentheses below the mean, the *standard deviation* of these Monte Carlo standard errors.
3. Bootstrap standard errors are calculated for each data sample by drawing 250 bootstrap samples using the estimated row and column probabilities. $\hat{P}_{E\cdot}$ and $\hat{P}_{\cdot E}$, from the data sample and calculating the standard deviation of each parameter across these bootstrap samples. The table reports the mean and, in parentheses below the mean, the *standard deviation* of these bootstrap standard errors.
4. The "Large N" standard error is based on the mean Monte Carlo standard errors from samples containing 5000 control and 10,000 treatment group members. This mean is treated as correct and adjusted upward by the square root of the ratio of the two sample sizes.
5. The "Normal" standard errors assume that the distribution of the estimated bound is normal. Thus, they ignore the non-negativity constraint for the lower bounds and the fact that the upper bound is the maximum of two random variables. These standard errors are calculated using the population values of the cell probabilities.
6. The "Correct Asymptotic" standard errors are based on the distributions of the estimated bounds given in equations (F-1) and (F-2). For the lower bound, they take account of the non-negativity constraint, while for the upper bound they take account of the fact that the bound is the maximum of two asymptotically normal random variables. These bounds are calculated using the population values of the cell probabilities.

TABLE E-2

Monte Carlo analysis of standard errors for Fréchet–Hoeffding bounds; coverage probabilities

(Estimates based on 1000 data samples; 250 Monte Carlo and 250 bootstrap samples per data sample; sample size of 1500 controls and 3000 treatments)

Bound	Population value of bound	Bootstrap CI coverage probability[2]	Monte Carlo CI coverage probability[3]
Population cell probabilities are $P_{EE}=P_{EN}=P_{NE}=P_{NN}=0{\cdot}25$			
Lower bound of P_{EE}	0·0	0·9720	0·8800
Lower bound of P_{EN}	0·0	0·9730	0·8750
Lower bound of P_{NE}	0·0	0·9830	0·8680
Lower bound of P_{NN}	0·0	0·9680	0·8610
Upper bound of P_{EE}	0·5	0·9280	0·8850
Upper bound of P_{EN}	0·5	0·9450	0·9010
Upper bound of P_{NE}	0·5	0·9320	0·8850
Upper bound of P_{NN}	0·5	0·9440	0·8990
Population cell probabilities are $P_{EE}=P_{NN}=0{\cdot}25$, $P_{EN}=0{\cdot}45$, $P_{NE}=0{\cdot}05$			
Lower bound of P_{EE}	0·0	0·9710	0·8690
Lower bound of P_{EN}	0·4	0·9530	0·9510
Lower bound of P_{NE}	0·0[1]	1·0000	1·0000
Lower bound of P_{NN}	0·0	0·9710	0·8700
Upper bound of P_{EE}	0·3	0·9500	0·9530
Upper bound of P_{EN}	0·7	0·9360	0·9010
Upper bound of P_{NE}	0·3	0·9260	0·8860
Upper bound of P_{NN}	0·3	0·9460	0·9440

1. The non-negativity constraint is binding in this case as the population value of the bound is zero. Because the expected value of the other term in the bound formula is −0·4, the estimated bound is also zero in all of the Monte Carlo and bootstrap samples, with the result that all of the confidence intervals equal [0, 0]. Since the population value is zero, the coverage probability equals 1·0 in this case.
2. Bootstrap confidence intervals are obtained for each data sample by adding and subtracting 1·96 times the bootstrap standard error to the estimated bounds for the data sample. Coverage probabilities are the fraction of data samples for which the bootstrap confidence interval constructed in this way contains the population value of the bound.
3. Monte Carlo confidence intervals are obtained for each data sample by adding and subtracting 1·96 times the Monte Carlo standard error to the estimated bounds for the data sample. Coverage probabilities are the fraction of data samples for which the Monte Carlo confidence interval constructed in this way contains the population value of the bound.

The distributions in these cases are asymmetric as well, being somewhat skewed left. In the case of upper bounds where the population values of the two arguments in the bound formula are equal, the bootstrap standard errors have a highly skewed distribution with a thick right tail. In both cases, increasing the sample size does not decrease the non-normality. In the remaining cases, the distributions of the bootstrap standard errors closely resemble those of the Monte Carlo standard errors.

We draw two main conclusions from this analysis. First, for the practical purpose of establishing statistical significance and sampling variability, the bootstrap standard errors perform well. Our analysis supports the use of the bootstrap standard errors. Second, as our theoretical analysis reveals, the quality of the bootstrap standard errors, as indicated by their variability, the coverage probabilities of confidence intervals generated from them, and the normality of their distribution, is strongly affected by the proximity of the population parameter value to the boundary of the parameter space (as represented by the constraints in the bounds formulae) and, in the case of the upper bounds, by whether or not the two arguments in the bounds formula have the same population values.

(ii) Monte Carlo Analysis of Bootstrap Standard Errors for the Impact Standard Deviation for the Case of Continuous Data

This portion of the appendix summarizes a Monte Carlo analysis of the performance of the bootstrap standard errors used to account for the variability in the estimated impact standard deviation, $[\text{VAR}(\Delta)]^{1/2}$. We find that the bootstrap standard errors are inaccurate when the population impact standard deviation is zero. Bootstrap confidence intervals centred around the point estimate of the impact standard deviation have very low coverage probabilities even for samples substantially larger than those used in the empirical analysis in this paper. In view of this evidence, we construct Monte Carlo cutoff values for rejection of the null hypothesis that the population impact standard deviation is zero. Using these cutoff values, we find that we can reject the null hypothesis that the population impact standard deviation is zero at the $P=0{\cdot}0001$ level.

The upper and lower Fréchet bounds are, respectively, the minimum of two empirical processes and the maximum of an empirical process above zero. Informative analytical expressions for the distribution of $\text{VAR}(\Delta)$ are difficult to obtain, so we use Monte Carlo methods. Our Monte Carlo analysis of the performance of the bootstrap standard errors proceeds in the following way. For each of the four cases we consider, we generate 250 samples of data by drawing observations at random from the distribution of earnings in the 18 months after random assignment for adult female controls in the National JTPA Study. For each sample, we first draw a control group sample of the size indicated in the first row of Table E-3. We then draw a "synthetic treatment group" sample of twice the size, in agreement with the 2:1 random assignment ratio used in the experimental data analyzed in the text. The Monte Carlo treatment group samples are also drawn from the control earnings distribution. For columns where the true impact standard deviation is set to zero, \$500 is added to each treatment group observation. In columns where the true impact standard deviation is set to \$500, the treatment observations are sorted and \$1000 is added to every other observation. In results not shown here, the same qualitative results are obtained from adding \$1000 to the upper half of the treatment group earnings distribution in each synthetic treatment group sample.

The percentiles of the control and synthetic treatment group earnings distributions for each sample are then obtained. Differencing across percentiles of the two distributions (i.e. subtracting the first percentile of the control distribution from the first percentile of the treatment group distribution and so on) yields the distribution of impacts for each data sample. The standard deviation of these differences is the estimated impact standard deviation for a given sample.

We draw 250 bootstrap samples for each sample of controls and synthetic treatments by random sampling from the control and treatment group earnings distribution of the sample. Each bootstrap sample is equal in size to the original sample. The standard deviation of impacts is constructed for each bootstrap sample in the manner just described. The standard deviation of the estimated impact standard deviations in the bootstrap samples is the bootstrap standard error for the corresponding data sample.

Table E-3 presents various statistics from the four cases we consider. The first three columns all have the population impact standard deviation set at zero, but vary the sample sizes from 1500 controls and 3000 treatments up to 5000 controls and 10,000 treatments. The final column returns to a sample size of 1500 controls and 3000 treatments, but sets the population impact standard deviation at \$500. The third row of Table E-3 reports the mean of the estimates of the impact standard deviation over the samples. The key comparison here is between the first and fourth columns, which reveal that the mean of the estimated impact standard deviations is \$431 in the case where the population value is \$0 and \$500 in the case where the population value is \$500. Thus, the estimates are, on average, right on target for the case where the population value is away from zero, but strongly biased when the population value is actually zero.

The second and third columns reveal that this bias declines with sample size, but even for samples of 5000 controls and 10,000 treatments the average estimate in the data samples is still over \$200 when the population standard deviation is zero. The source of this bias is clear. Any random variation that results in the differences in the percentiles of the control and treatment earnings distributions being other than the common impact of \$500 shows up as upward bias in the estimated impact standard deviation since the random variation is squared in the calculation of the impact variance. Since the percentiles are more precisely estimated as the sample sizes increases, the random variation decreases and so does the mean bias. The fourth row of Table E-3 shows that the variance of the estimated impact standard deviations is not affected very much by the population value of the parameter—compare again the first and fourth columns—and declines strongly with the sample size, as expected. In terms of shape, the distributions of the estimated impact standard deviations from the data samples in the cases corresponding to both the first and fourth columns are skewed right, but more strongly so in the case where the population impact standard deviation is zero.

The fifth row of Table E-3 presents the mean of the bootstrap standard errors in each case. In general, the mean is about 15% larger than the variation in the data sample estimates. For example, in the second column the variation in the data sample estimates is \$110 while the mean of the bootstrap standard errors is \$130. This

TABLE E-3

Performance of bootstrap standard errors for the estimated standard deviation of impacts

(Perfect positive dependence case; 250 data samples and 250 bootstrap samples per data sample; samples drawn from the National JTPA Study 18 month impact sample; adult females)

Parameter	Sample 1	Sample 2	Sample 3	Sample 4
Sample size	1500 control 3000 treatment	3000 control 6000 treatment	5000 control 10,000 treatment	1500 control 3000 treatment
Population impact standard deviation	0·0	0·0	0·0	500·0
Mean of data sample estimates	431	290	220	500
Std. dev. of data sample estimates	166	110	77	145
Mean of bootstrap standard errors	189	130	98	178
Std. dev. of bootstrap standard errors	39	29	23	38
Bootstrap CI coverage probability	0·352	0·412	0·3480	0·984

1. Estimates are based on 250 data samples of the indicated size. Data samples are drawn from the control group data on earnings in the 18 months after random assignment for adult females. Treatment group data are drawn from the control group distribution, with the known impact distribution then added to the treatment group observations. For the case of zero impact variance, 500 is added to each treatment observation in the data samples. For the case of an impact standard deviation of 500, impacts of 0 and 1000 are added to alternating treatment observations. Concentrating the non-zero impacts on one end of the distribution (results not shown here) does not affect the performance of the bootstrap.
2. The third row of the table reports the mean of the estimated impact standard deviations across the 250 data samples in each case. The fourth row reports the standard deviation of these estimates.
3. A total of 250 bootstrap samples are drawn from each data sample. For each data sample, a bootstrap standard error for the estimated standard deviation of impacts is obtained by taking the square root of the variance of the impact standard deviation estimates in the bootstrap samples. The fifth row of the table reports the mean of these bootstrap standard errors over the 250 data samples, while the sixth row reports their standard deviation.
4. The final row of the table reports the coverage probability that results from constructing a bootstrap confidence interval centred on the estimated impact standard deviation for each data sample and extending to 1·96 times the bootstrap standard error on either side. The coverage probability is the fraction of the data samples for which the interval so constructed contains the population value of the impact standard deviation.

additional variation arises from the fact that the bootstrap standard error is based on sampling from the data sample rather than from the population.

The variability in the bootstrap standard errors shown in the sixth row of Table E-3 is modest by comparison with the variability in the estimates of the impact standard deviation across data samples. In the cases corresponding to both the first and fourth columns, the distributions of bootstrap standard errors are modestly skewed right, but concentrated almost entirely within the range from $100 to $300. The shapes of the two distributions are roughly similar, suggesting that the shape is not affected by the population value of the impact standard deviation.

The final row of Table E-3 presents coverage probabilities for bootstrap confidence intervals. These confidence intervals are centred on the data sample estimate of the impact standard deviation for each sample, extending 1·96 times the bootstrap standard error on either side. The coverage probabilities give the fraction of the 250 data samples in each column for which the bootstrap confidence interval contained the population value of the impact standard deviation. The coverage probabilities are very low in the columns where the population impact standard deviation is zero. In contrast, the coverage probability is within 0·03 of 0·95 for the case where the population impact standard deviation is $500. The poor performance in the case where $\mathrm{VAR}(\Delta)=0$ results from the upward bias in the data sample estimates that centre the confidence intervals. This poor performance suggests that reliance on the bootstrap standard errors produces misleading statistical inferences regarding the null hypothesis that the population impact standard deviation is zero.

An alternative to relying on the bootstrap standard errors to test the null hypothesis of a zero impact standard deviation is to construct the distribution of estimates under the null using Monte Carlo methods. The quantiles of the simulated distribution of estimates under the null provide cutoff values that may be used to assign a p-value for a test of the null. Such cutoff values appear in Table E-4. These cutoff values are based on 100,000 Monte Carlo samples of the indicated size. For example, the cutoff corresponding to a p-value of 0·60 is the 60th percentile of the simulated distribution of estimates under the null. The sample size used in the paper is a bit larger than that in the first column, yet the estimate of the impact standard deviation in Table 2 is $1858. Thus, the first column of Table E-4 indicates that we can reject the null of a zero impact standard deviation at

TABLE E-4

Monte Carlo cutoff values for probabilities of type I error under the null hypothesis that the population impact standard deviation is zero

(Perfect positive dependence case; cutoff values based on 100,000 Monte Carlo samples; samples drawn from the National JTPA Study 18 month impact sample; adult females)

P-value	Cutoff value for 1500 controls and 3000 treatments	Cutoff value for 5000 controls and 10,000 treatments
0·50	449	218
0·40	466	229
0·30	516	251
0·20	577	282
0·10	668	327
0·05	756	367
0·01	957	458
0·001	1163	580
0·0001	1335	744

1. Estimates are based on 100,000 random samples of the indicated size drawn from the real control sample for adult females. Treatment group samples are created from the control group samples by drawing at random and then adding 500 to all treatment group observations.
2. The impact standard deviation is calculated for each Monte Carlo sample by collapsing the control and treatment group distributions into percentiles, taking differences across percentiles (thereby imposing perfect positive dependence) and then calculating the standard deviation of the percentile differences.

the 0·0001 level. Indeed, our estimate exceeds the estimates for all 100,000 of the Monte Carlo samples, providing very strong evidence in support of our conclusion in the text that the experimental data bound the population impact standard deviation away from zero for adult women.

F. Tests of first and second order stochastic dominance

This appendix describes the tests of stochastic dominance discussed in Section 5 of the text. We implement three different tests of stochastic dominance. The tests compare the c.d.f.s, or integrated c.d.f.s, of the distributions of earnings in the 18 months after random assignment for adult women in the treatment and control groups in the JTPA experiment. The sample and earnings measure are the same as used in the other empirical analyses we present, and the two c.d.f.s are shown in Figure 3. When applied to the JTPA data for adult women, all of the tests are consistent with the hypothesis that the treatment group earnings distribution first-order (and therefore also second-order) stochastically dominates the control earnings distribution.

(1) We implement tests due to Anderson (1996). Using multiple comparisons of the c.d.f.s and integrated c.d.f.s of the treatment and control distributions at a finite set of points, we reject the null that the two distributions are equal in favour of the alternatives that the treatment distribution first- and second-order stochastically dominates the control distribution at the 1% significance level.

(2) We also use the test statistics $D^{+}=\max_{x}\{\hat{F}_1(x)-\hat{F}_0(x)\}$ and $D^{-}=\min_{x}\{\hat{F}_1(x)-\hat{F}_0(x)\}$, where "^" denotes an estimated c.d.f. The null that the treatment distribution first-order stochastically dominates the control distribution is rejected for large values of D^{+}, while the null that the control distribution first-order stochastically dominates the treatment distribution is rejected for large values of D^{-}. Applying this test to the JTPA data and using the approximate asymptotic standard errors provided by Stata, we reject the null that the control distribution first-order stochastically dominates the treatment group distribution at the 1% significance level, and fail to reject the null that the treatment distribution first-order stochastically dominates the control distribution.

(3) We also utilize a test due to Klecan, McFadden and McFadden (1997). They implement the two tests just described but use Monte Carlo methods to obtain the standard errors. In practice, their standard errors are larger, which leads to weaker inferences. They also consider the statistic $D^* = \min \{|D^+|, |D^-|\}$. They use this statistic to test the null that one of the distributions first-order dominates the other. We fail to reject this null with a p-value of 0·81.

All three sets of test statistics are thus consistent with the hypothesis that the treatment group distribution first-order and second-order stochastically dominates the control group distribution.

Acknowledgements. This paper previously circulated under the title "Making The Most Out of Social Experiments: Reducing the Intrinsic Uncertainty in Evidence From Randomized Trials With An Application to The National JTPA Experiment" with the names in alphabetical order. This research was supported by NSF SBR 91-11-455, NSF SBR 93-21-048, a grant from the Russell Sage Foundation, and a grant from the Lynde and Harry Bradley Foundation, Milwaukee, Wisconsin. We thank Anders Björklund, Richard Blundell, Tim Conley, Bo Honoré, Joel Horowitz, Hidehiko Ichimura, Matt Kahn, Tom MaCurdy, Robert Moffitt, Derek Neal, Jose Scheinkman, and Ed Vytlacil, the editor Manuel Arellano and three anonymous referees for helpful comments. We thank John Geweke for references. Substantial portions of this paper were presented as the Barcelona Lecture, 1990, which was widely circulated. We have benefitted from comments received at the University of Chicago, at the Royal Danish Conference, Kolding, Denmark, May 1993, the Federal Reserve Bank of Minneapolis, October 1993, the CEMFI Conference on Evaluation of Training Programmes in Madrid, Spain in September 1993, the NSF Conference on Nonparametric and Semiparametric Inference held at Northwestern University in October 1993, the University of California, Berkeley, 1994, and the University of Western Ontario in March 1994.

REFERENCES

ANDERSON, G. (1996), "Nonparametric Tests of Stochastic Dominance in Income Distributions", *Econometrica*, **64**, 1183–1193.

BISHOP, Y., FEINBERG, S. and HOLLAND, P. (1975) *Discrete Multivariate Analysis: Theory and Practice* (Cambridge: MIT Press).

BLOOM, H., ORR, L., CAVE, G., BELL, S. and DOOLITTLE, F. (1993) *The National JTPA Study: Title IIA Impacts on Earnings and Employment at 18 Months* (Bethesda: Abt Associates).

CAMBANIS, S., SIMONS, G. and STOUT, W. (1976), "Inequalities for $E(k(X, Y))$ When the Marginals Are Fixed", *Zeitschrift Für Wahrscheinlichskeitstheorie*, **36**, 285–294.

CARROLL, R. and HALL, P. (1988), "Optimal Rates of Convergence for Deconvolving a Density", *Journal of the American Statistical Association*, **83**, 1184–1186.

COX, D. R. (1958) *The Planning of Experiments* (New York: Wiley).

CSÖRGO, M. (1983) *Quantile Processes with Statistical Applications* (Philadelphia: Society for Industrial and Applied Mathematics).

DANIELS, H. E. (1944), "The Relationship Between Measures of Correlation in the Universe of Sample Permutations", *Biometrika*, **33**, 129–135.

DANIELS, H. E. (1948), "A Property of Rank Correlations", *Biometrika*, **35**, 416–447.

DREZE, J. and STERN, N. (1987), "The Theory of Cost-Benefit Analysis", in Auerbach, A. and Feldstein, M. (eds.) *Handbook of Public Economics, Volume 2* (Amsterdam: North Holland), 909–989.

FISHER, R. A. (1951) *The Design of Experiments, 6th Edition* (London: Oliver and Boyd).

FRÉCHET, M. (1951), "Sur Les Tableaux de Corrélation Dont Les Marges Sont Données", *Annals University Lyon: Series A*, **14**, 53–77.

GNEDENKO, B. (1976) *The Theory of Probability* (Moscow: Mir).

HECKMAN, J. (1978), "Dummy Endogenous Variables in a Simultaneous Equations System", *Econometrica*, **46**, 931–961.

HECKMAN, J. (1990*a*), "Varieties of Selection Bias", *American Economic Review*, **80**, 313–318.

HECKMAN, J. (1990*b*), "Alternative Approaches To The Evaluation of Social Programs: Econometric and Experimental Methods" (Barcelona Lecture, World Congress of the Econometric Society).

HECKMAN, J. (1992), "Randomization and Social Policy Evaluation", in Manski, C. and Garfinkel, I. (eds.) *Evaluating Welfare and Training Programs* (Cambridge: Harvard University Press), 201–230.

HECKMAN, J. (1996), "Randomization as an Instrumental Variable", *Review of Economics and Statistics*, **78**, 336–341.

HECKMAN, J. (1997*a*), "Constructing Econometric Counterfactuals Under Different Assumptions" (University of Chicago, mimeo).

HECKMAN, J. (1997*b*), "Instrumental Variables: A Study of Implicit Behavioral Assumptions in One Widely-Used Estimator", *Journal of Human Resources*, **1**, 1–40.

HECKMAN, J., HOHMANN, N., KHOO, M. and SMITH, J. (1997), "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment" (University of Chicago, mimeo).

HECKMAN, J. and HONORÉ, B. (1990), "The Empirical Content of the Roy Model", *Econometrica*, **58**, 1121–1149.
HECKMAN, J., LOCHNER, L., SMITH, J. and TABER, C. (1997), "The Effects of Government Policy on Human Capital Investment and Wage Inequality", *Chicago Policy Review*, **1**, 1–40.
HECKMAN, J. and ROBB, R. (1985), "Alternative Methods for Evaluating the Impact of Interventions", in Heckman, J. and Singer, B. (eds.) *Longitudinal Analysis of Labor Market Data* (Cambridge: Cambridge University Press), 156–246.
HECKMAN, J. and ROBB, R. (1986), "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes", in Wainer, H. (ed.), *Drawing Inference From Self-Selected Samples* (Berlin: Springer-Verlag), 63–107.
HECKMAN, J., ROBB, R. and WALKER, J. (1990), "Testing the Mixture of Exponentials Hypothesis and Estimating the Mixing Distribution by the Method of Moments", *Journal of the American Statistical Association*, **85**, 582–589.
HECKMAN, J. and SMITH, J. (1993), "Assessing the Case for Randomized Evaluation of Social Programs", in Jensen, K. and Madsen, P. K. (eds.), *Measuring Labour Market Measures: Evaluating the Effects of Active Labour Market Policies* (Copenhagen: Danish Ministry of Labour).
HECKMAN, J. and SMITH, J. (1995), "Assessing the Case for Social Experiments", *Journal of Economic Perspectives*, **9**, 85–110.
HECKMAN, J. and SMITH, J. (1997), "Evaluating the Welfare State", in Strom, S. (ed.), *Frisch Centenary* (Cambridge: Cambridge University Press).
HECKMAN, J., SMITH, J. and TABER, C. (1998), "Accounting for Dropouts in Evaluations of Social Programs", *Review of Economics and Statistics* (forthcoming).
HIBBARD, T. (1963), "An Empirical Study of Minimal Storage Sorting", *Communications of the ACM*, **6**, 206–213.
HOEFFDING, W. (1940), "Masstabinvariate Korrelationstheorie", *Schriften des Mathematischen Instituts und des Institutes Für Angewandte Mathematik der Universität Berlin*, **5**, 179–233.
HOROWITZ, J. and MARKATOU, M. (1996), "Semiparametric Estimation of Regression Models for Panel Data", *Review of Economic Studies*, **63**, 145–168.
JANSSON, P. (1984) *Deconvolution* (Orlando: Academic Press).
KAPLOW, L. (1993), "Should the Government's Allocation Branch Be Concerned About the Distortionary Cost of Taxation and Distributive Effects?" (National Bureau of Economic Research, Working Paper No. 4566).
KENDALL, M. G. (1970) *Rank Correlation Methods, Fourth Edition* (London: Griffen).
KENDALL, M. G. and STUART, A. (1977) *The Advanced Theory of Statistics, Volume 1, Fourth Edition* (London: Griffen).
KLECAN, D., McFADDEN, D. and McFADDEN, R. (1997), "A Robust Test for Stochastic Dominance", *Econometrica* (forthcoming).
KNUTH, D. (1973) *The Art of Computer Programming, Volume 3, Sorting and Searching* (Reading: Addison-Wesley).
LAVINE, M., WASSERMAN, L. and WOLPERT, R. (1991), "Bayesian Inference with Specified Prior Marginals", *Journal of the American Statistical Association*, **86**, 964–971.
MARDIA, K. V. (1970) *Families of Bivariate Distributions* (London: Griffen).
QUANDT, R. (1972), "A New Approach to Estimating Switching Regressions", *Journal of the American Statistical Association*, **67**, 306–310.
QUANDT, R. (1988) *The Econometrics of Disequilibrium* (Oxford: Blackwell).
RACHEV, S. (1985), "The Monge-Kantorovich Mass Transfer Problem and Its Stochastic Applications", *Theory of Probability and Its Applications*, **29**, 147–671.
RAWLS, J. (1971) *A Theory of Justice* (Cambridge: Harvard University Press).
ROY, A. (1951), "Some Thoughts on the Distribution of Earnings", *Oxford Economic Papers*, **3**, 145–146.
RÜSCHENDORF, L. (1981), "Sharpness of Fréchet Bounds", *Zeitschrift Für Wahrscheinlichskeitstheorie*, **41**, 293–302.
STEFANSKI, L. and CARROLL, R. (1990), "Deconvoluting Kernel Density Estimators", *Statistics*, **2**, 169–184.
STRASSEN, V. (1965), "Existence of Probability Measures Given Marginals", *Annals of Mathematical Statistics*, **33**, 423–439.
TCHEN, A. (1980), "Inequalities for Distributions With Given Marginals", *Annals of Probability*, **8**, 814–827.
TONG, Y. L. (1980) *Probability Inequalities in Multivariate Distributions* (New York: Academic Press).
VIVERBERG, W. (1993), "Measuring the Unidentified Parameter of the Roy Model of Selectivity", *Journal of Econometrics*, **57**, 69–90.
WEISBROD, B. (1968), "Income Redistribution Effects and Cost-Benefit Analysis", in Chase, S. (ed.), *Problems in Public Expenditure Analysis* (Washington: Brookings Institution), 41–55.
WHITT, W. (1976), "Bivariate Distributions With Given Marginals", *The Annals of Statistics*, **4**, 1280–1289.