

**TECHNICAL WORKING PAPER SERIES**

**MAKING THE MOST OUT OF SOCIAL  
EXPERIMENTS: REDUCING THE  
INTRINSIC UNCERTAINTY IN EVIDENCE  
FROM RANDOMIZED TRIALS WITH AN  
APPLICATION TO THE NATIONAL  
JTPA EXPERIMENT**

Nancy Clements  
James Heckman  
Jeffrey Smith

Technical Working Paper No. 149

NATIONAL BUREAU OF ECONOMIC RESEARCH  
1050 Massachusetts Avenue  
Cambridge, MA 02138  
January 1994

The authors are Research Associate, Center for Social Program Evaluation, The Harris School, University of Chicago (Clements and Smith), and Henry Schultz Professor of Economics and Director of the Center For Social Program Evaluation at the University of Chicago (Heckman). This research was supported by NSF SES-91-11455, a grant from the Russell Sage Foundation, and a grant from the Lynde and Harry Bradley Foundation, Milwaukee, Wisconsin. We thank Suna Barlas, Anders Bjorklund, Tim Conley, Bo Honoré, Hidehiko Ichimura, Derek Neal and Tomas Philipson for helpful comments. We thank Suna Barlas for her programming assistance on the random coefficient model and John Geweke for references. Portions of this paper were presented as the Barcelona Lecture, 1990, which was widely circulated. We have benefited from comments received at the University of Chicago, at the Royal Danish Conference, Kolding, Denmark, May, 1993, the Federal Reserve Bank of Minneapolis, the CEFMI Conference on Evaluation of Training Programs in Madrid, Spain in September 1993 and the NSF Conference on Nonparametric and Semiparametric Inference held at Northwestern University, held in October, 1993. This paper is part of NBER's research program in Labor Studies. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

MAKING THE MOST OUT OF SOCIAL  
EXPERIMENTS: REDUCING THE  
INTRINSIC UNCERTAINTY IN EVIDENCE  
FROM RANDOMIZED TRIALS WITH AN  
APPLICATION TO THE NATIONAL  
JTPA EXPERIMENT

ABSTRACT

This paper demonstrates that even under ideal conditions, social experiments in general only uniquely determine the mean impacts of programs but not the median or the distribution of program impacts. The conventional common parameter evaluation model widely used in econometrics is one case where experiments uniquely determine joint the distribution of program impacts. That model assumes that everyone responds to a social program in the same way. Allowing for heterogeneous responses to programs, the data from social experiments are consistent with a wide variety of alternative impact distributions. We discuss why it is interesting to know the distribution of program impacts. We propose and implement a variety of different ways of incorporating prior information to reduce the wide variability intrinsic in experimental data. Robust Bayesian methods and deconvolution methods are developed and applied. We analyze earnings and employment data on adult women from a recent social experiment. In order to produce plausible impact distributions, it is necessary to impose strong positive dependence between outcomes in the treatment and in the control distributions. Such dependence is an outcome of certain optimizing models of the program participation decision.

Nancy Clements  
Center for Social Program Evaluation  
The Harris School  
University of Chicago  
Chicago, IL 60637

James Heckman  
Henry Schultz Professor of Economics  
Director, Center for Social Program Evaluation  
University of Chicago  
Chicago, IL 60637  
and NBER

Jeffrey Smith  
Center for Social Program Evaluation  
The Harris School  
University of Chicago  
Chicago, IL 60637

*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*

The case for social experimentation implicitly rests on the belief that mean differences in outcomes - comparing those given a treatment with those denied it by randomization - measure something of economic or social interest. Yet, as the quotation from Genesis reveals, even in biblical times features other than mean outcomes were perceived to be of interest. Just as Abraham convinced the Lord to spare Sodom and Gomorrah from destruction if he could find 50 sufficiently righteous persons living there, so many evaluators of contemporary social programs would deem them successful if enough persons reaped sufficient benefits from them even though the average participant did not. Voters might well oppose a social program with a high estimated mean impact arising solely from a large beneficial program effect for a few participants or a program that had substantial negative impacts on even a small fraction of participants. Programs with a negative impact on certain types of participants may need to be retargeted toward those more likely to benefit from them. Features of the distribution of program gains form the basis for all of these judgments.

Social experiments recover two distributions of outcomes - one for the treatment group and one for the control group - but cannot recover the joint distribution of outcomes or the derived gain distribution because no person is observed in both treatment states. From the outcome distributions for participants and non-participants, mean differences can be estimated. However, without invoking additional assumptions one cannot use data from social experiments to estimate the median impact from a program or the gain at the top deciles of the distribution or any other measure that requires the joint distribution for its calculation. (Heckman (1992)).

This problem does not arise in the conventional econometric model that is widely used to evaluate programs. As in all models of program evaluation, in the conventional model,

persons have two possible outcomes corresponding to the treated ( $Y_1$ ) and untreated ( $Y_0$ ) states. We observe only one member of the pair of latent variables ( $Y_1, Y_0$ ). If a person receives treatment we set  $R = 1$ ;  $R = 0$  otherwise. In the conventional econometric approach, it is further assumed that

$$Y_1 - Y_0 = \alpha$$

where  $\alpha$  is the same for everyone. Assuming they exist, the mean outcomes in the control and treatment groups can be used to obtain  $\alpha$ :

$$\alpha = E(Y_1 | R = 1) - E(Y_0 | R = 0).$$

We can derive the joint distribution of ( $Y_0, Y_1$ ) from either marginal distribution since knowledge of  $Y_0$  implies knowledge of  $Y_1$ :

$$F_1(Y_1) = F_0(Y_1 - \alpha).$$

In a more familiar regression setting, follow Heckman and Robb (1985) or Bjorklund and Moffitt (1987) and write outcome  $Y$  in mixture form as

$$Y = RY_1 + (1 - R)Y_0.$$

Assuming a conventional econometric specification,

$$Y_0 = X\beta + U, \quad E(U | X) = 0$$

the conventional identical effect econometric model can be written as

$$\begin{aligned} Y &= R(Y_1 - Y_0) + Y_0 \\ &= R\alpha + X\beta + U. \end{aligned}$$

In an experimental setting,  $R$  is orthogonal to  $U$  and  $X$ . Least squares can be used to consistently estimate  $\alpha$  and  $\beta$  under standard rank conditions. The joint distribution of ( $Y_0, Y_1$ ) is produced by estimating the distribution of  $U$  and adding back  $R\alpha + X\beta$  if  $U$  is independent of  $X$ . Within

the conventional econometric framework, it is possible to answer all of the evaluation questions that require knowledge of the joint distribution.

By assuming identical treatment effects for everyone, the conventional model imposes strong restrictions on the data. A person's place in the  $Y_1$  distribution is determined by his/her place in the  $Y_0$  distribution. The best person in the latent  $Y_1$  distribution is the same as the best person in the latent  $Y_0$  distribution. The conventional model is robust to the randomization bias that arises in a more general variable-treatment-impact model when randomization alters the quality of the program's treatment pool. (Heckman (1992)). Since the impact of treatment is the same for everyone in the conventional econometric framework, estimates of program impact are not affected by randomization. The case for social experimentation is made most strongly within the context of this model.

Convenient and familiar though it is, the conventional model is not plausible. People are likely to vary in their response to the same treatment and there are also likely to be important unobserved differences in the treatments received within broadly measured treatment categories. This paper considers how social experiments can be used to estimate or bound the distribution of gains to participating in programs when  $\alpha$  is not a constant, and there is variability in treatment impacts across persons.

The problem of bounding a joint distribution from knowledge of its marginals has been studied by classical and Bayesian statisticians. Hoeffding (1940) and Frechet (1951) present elementary bounds. More recently, Bayesians have drawn on and extended these ideas in constructing bounds on posterior (joint outcome) distributions when only prior marginal distributions are specified. Recent papers by Lavine, Wasserman, and Wolpert (1991), Berger

(1990) and Berger and Moreno (1992) exemplify this literature and provide many additional references.

A considerable body of empirical experience with these bounds suggests that they are often rather wide, especially for continuous outcome measures. (See Lavine et.al for a recent demonstration of this point). Although much of this literature is not relevant to the problem of bounding the distribution of program gains considered in this paper, the empirical evidence reported in this paper is broadly consistent with previous findings. We analyze the outcomes from the recent Job Training Partnership Act (JTPA) experiment conducted by Abt Associates and the Manpower Demonstration Research Corporation for the U.S. Department of Labor. We find that experimental evidence does not settle important questions about the distribution of program gains. It must be supplemented by additional information in the form of plausible features of the distribution of outcomes, plausible dependence relationships among latent outcomes, or plausible selection rules in order to provide meaningful bounds on the distribution of program gains. Fairly strong positive dependence among latent outcomes is required in order to produce plausible impact distributions.

This paper develops and applies methods for incorporating prior information into the analysis of social experiments to produce credible estimates of the distribution of program impacts. Several methods are developed and applied to analyze the distribution of program gains using data from the JTPA experiment. These methods are motivated by intuitions about how closely related persons' treatment outcomes are to their control outcomes. Some of these intuitions emerge from models of the program participation decision. We take as our point of departure the conventional common effect model that specifies a tight deterministic relationship

between  $Y_1$  and  $Y_0$ . All of our methods can be seen as various ways of relaxing this tight relationship and yet still deriving plausible information about the distribution of gains using experimental data. We are unaware of any previous literature that specifies the joint dependence between treatment and control outcomes in the manner we suggest.

The exposition of this paper proceeds as follows. First, we discuss what an experiment does and motivate why it is interesting to know the full distribution of program gains, rather than just its mean. Second, we examine the intrinsic uncertainty regarding the gain distribution present in experimental data. Using JTPA data, we demonstrate the wide variety of distributions of program impacts that are consistent with the experimental evidence. Third, we present two methods for incorporating prior information into the analysis of social experiments and apply them to the JTPA data. Fourth, we use linear model random coefficient methods and nonparametric deconvolution techniques to uniquely recover the distribution of program gains in the case where gains are not known at the time decisions are made about participation in the program. The paper concludes with a summary.

*(1) The Evaluation Problem and How Experiments Partly Solve It*

If analysts could observe  $(Y_0, Y_1)$  for everyone, there would be no evaluation problem.

One could form

$$\Delta = Y_1 - Y_0$$

and compute the gross gain to program participation for various populations of interest. For each person, one could determine whether program participation raised or lowered outcomes. The evaluation problem arises because we do not observe  $(Y_0, Y_1)$  for everyone.

Ordinary program and comparison group data enable determination of the conditional outcome distributions for participants ( $d = 1$ ) and non-participants ( $d = 0$ ):

(1a)  $F(y_1 | d = 1)$  (participant outcomes)

and

(1b)  $F(y_0 | d = 0)$  (non-participant outcomes).

We don't know  $Y_0$  for participants or  $Y_1$  for non-participants. We do not even know the counterfactual conditional distributions:

(1c)  $F(y_0 | d = 1)$  (what participant outcomes would have been had they not participated)

and

(1d)  $F(y_1 | d = 0)$  (what non-participant outcomes would have been had they participated).

Since we don't observe both the treated and untreated states for participants or non-participants, we also do not know the joint distributions for participants and non-participants:

(1e)  $F(y_1, y_0 | d = 1)$

and

(1f)  $F(y_1, y_0 | d = 0)$ .

Unless participation is random with respect to outcomes (i.e.  $F(y_0 | d = 1) = F(y_0 | d = 0)$  and/or  $F(y_1 | d = 1) = F(y_1 | d = 0)$ ) it is not possible to use program data, as represented by (1a) and (1b), to estimate either the mean impact of treatment on the treated:

$$E(Y_1 - Y_0 | d = 1)$$

or the mean impact of non-participation on the non-participants:

$$E(Y_0 - Y_1 | d = 0).$$

From populations of program participants and non-participants, we can obtain

$$\begin{aligned} E(Y_1 | d = 1) - E(Y_0 | d = 0) \\ = E(Y_1 - Y_0 | d = 1) + \{E(Y_0 | d = 1) - E(Y_0 | d = 0)\}. \end{aligned}$$

Only if there is no selection of participants on the basis of  $Y_0$  will the term in braces be zero.

Under ideal conditions, randomized trials produce data that can be used to solve some of these problems. The most commonly used point of randomization occurs at the stage of program application and acceptance where a person declares interest in the program and would ordinarily be offered treatment ( $d^* = 1$ ). (We distinguish  $d^*$  from  $d$  to distinguish program participation under random assignment from participation in an environment without random assignment.) As before,  $R = 1$  if a person actually receives treatment;  $R = 0$  otherwise.

Assuming that randomization does not alter the population of accepted applicants, then

$$(A-1) \quad F(y_1 | R = 1, d^* = 1) = F(y_1 | d = 1)$$

where  $d = 1$  is "accepted and admitted in an environment without randomization". A further assumption is that

$$(A-2) \quad F(y_0 | R = 0, d^* = 1) = F(y_0 | d = 1).$$

Throughout this paper we maintain the assumption of no bias induced by randomization noting only in passing that there is much evidence against it. (Heckman (1992)). Under (A-1) and (A-2), we may replace  $d^*$  by  $d$  and we do so for the balance of this paper.

Under assumption (A-2), social experiments produce  $F(y_0 | d = 1)$ . Randomized trials

do not recover (1d), because we cannot force non-participants to participate. Social experiments do not recover (1e) or (1f) because we do not observe  $(Y_1, Y_0)$  together for either participants or non-participants. Without invoking further assumptions, it is only possible to determine features of the joint distribution that depend solely on  $F(y_1 | d = 1)$  and  $F(y_0 | d = 1)$ . One important feature is

$$E(\Delta | d = 1) = E(Y_1 - Y_0 | d = 1) = E(Y_1 | d = 1) - E(Y_0 | d = 1).$$

Medians or other quantiles of the gain distribution cannot be consistently estimated from marginal distributions. However, in the special case

$$Y_1 - Y_0 = \alpha$$

discussed in the introduction, where  $\alpha$  is the same for everyone (or where  $\alpha$  can be generalized to depend on a set of observed variables  $X$ ), the distribution of gains is degenerate since everyone has the same gain. (This is sometimes called the "dummy endogenous variable" model). In this case, ideal experiments recover the full joint distribution of the gains and there is no randomization bias. (Heckman (1992)).

In the absence of randomization bias, social experiments determine the mean impact of treatment on the treated. The current emphasis in the program evaluation literature on means over medians or other quantiles of the impact distribution is largely a cultural artifact bound up with the notion of the "average man" that has dominated thinking in statistics for more than one hundred and thirty years. (See Stigler (1986)). In addition, the widely used regression model of adjusted means presented in the introduction to this paper is the standard framework of analysis for most social scientists. Yet, knowledge of the mean does not suffice to answer many interesting questions if persons respond differently to treatment.

*(2) Why Is It Interesting To Estimate The Distribution  
of Program Gains For Participants?*

Answers to many interesting evaluation questions require knowledge of the distribution of program gains. From the standpoint of a detached observer of a social program (e.g. a "social planner") who takes the base state values (denoted "0") as those that would prevail in the absence of the program, it is of interest to know

- (a) the proportion of people taking the program who benefit from it:

$$\Pr(Y_1 > Y_0 \mid d = 1) = \Pr(\Delta > 0 \mid d = 1);$$

- (b) the proportion of the total population benefitting from the program:

$$\Pr(Y_1 > Y_0 \mid d = 1) \cdot \Pr(d = 1) = \Pr(\Delta > 0 \mid d = 1) \cdot \Pr(d = 1);$$

- (c) selected quantiles of the impact distribution

$$\inf_{\Delta} \{ \Delta : F(\Delta \mid d = 1) > q \}, \text{ where } q \text{ is a quantile of the distribution.}$$

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

$$F(\Delta \mid d = 1, Y_0 = y_0).$$

Each of these measures can be defined conditional on observed characteristics  $X$ . Measure (a) is of interest in determining how widely program gains are distributed among participants. Detached observers with preferences over distributions of program outcomes would be unlikely to assign the same weight to two programs with the same mean outcome, one of which produced favorable outcomes for only a few persons while the other distributed gains more broadly. When considering a proposed program, it is of interest to determine the proportion of participants who are harmed (i.e.  $\Pr(Y_1 < Y_0 \mid d = 1)$ ) as a result of program participation. Negative mean impact results such as those found for certain groups in the national JTPA experiment might be acceptable if most participants receive a positive gain from the program. These features of the

outcome distribution are likely to be of interest to evaluators even if the persons studied do not know both their  $Y_0$  and  $Y_1$  values in advance of participating in the program.

Measure (b), which is derived by multiplying measure (a) by the probability of participation, determines the proportion of the entire population that benefits from the program, assuming that costs are broadly distributed and are not perceived to be related to the specific program being evaluated. If voters have correct expectations about the joint distribution of program gains, it is of interest to students of positive political economy to determine if voting is related to program benefits received by the electorate. Large program gains received by a few persons may make it easier to organize interest groups in support of a program than if the same gains are distributed more widely.

Evaluators interested in the distribution of program benefits would be interested in measure (c). Evaluators with a special interest in the impact of a program on recipients in the lower tail of the base state distribution would find measure (d) of great use. It reveals how the distribution of gains depends on the base state for participants. This measure provides answers to questions such as "do the distributions of gains for the participants who would be among the worse off in the absence of a program stochastically dominate the distributions of gains for the participants who would be among the better off in the absence of the program?" and "does the program reduce inequality among participants?"

All of these measures require for their computation knowledge of features of the joint distribution of outcomes for participants (formula (1e)) which cannot be obtained from data produced by a social experiment unless additional assumptions are invoked.

Information about the joint distribution of outcomes, if available, is also informative

about the structure of decision processes and the information sets used by the agents being studied. Suppose that  $Y_1$  and  $Y_0$  are the net outcomes from participation and non-participation respectively. If agents are uncertain about both their potential  $Y_0$  and  $Y_1$  values, but know the distributions of these variables, then individual rationality combined with a monotonically increasing utility function  $U(y)$  would have agents pick the option with the greatest utility:

$$d = 1 \text{ if } \int U(y)dF_0(y) \leq \int U(y)dF_1(y)$$

$$d = 0 \text{ otherwise.}$$

(We suppress the dependence of  $U$ ,  $F_0$  and  $F_1$  on  $X$  for notational simplicity). If  $U$  is concave, a sufficient condition for  $d = 1$  is that  $Y_1$  second-degree stochastically dominates  $Y_0$  i.e.  $\int_{-\infty}^z F_1(y_1)dy_1 \leq \int_{-\infty}^z F_0(y_0)dy_0$  for all  $z$ . Data from a social experiment do not provide the requisite data for this test of rationality. Without invoking further assumptions social experiments only provide information on  $F_0(y | d = 1)$  and  $F_1(y | d = 1)$ .

However, the distributions produced from an experiment can be used to check if expectations are rational. If persons choose  $d = 1$ , then it follows that for all persons in the program

$$\int U(y)dF_1(y | d = 1) > \int U(y)dF_0(y | d = 1).$$

A necessary and sufficient condition for this to be true for all  $U$  is that  $Y_1$  second order stochastically dominates  $Y_0$  given  $d = 1$ , so that  $\int_{-\infty}^z F_1(y_1 | d = 1)dy_1 < \int_{-\infty}^z F_0(y_0 | d = 1)dy_0$  for all  $z$ . This condition is reversed for a risk-loving agent.

Neither of the preceding tests requires knowledge of the joint distributions  $F(y_1, y_0)$  and  $F(y_1, y_0 | d = 1)$ . This is an intrinsic feature of choice under uncertainty for agents who do not know their position in both the  $Y_0$  and the  $Y_1$  distributions as long as there is no regret in agent's

preferences (i.e. only the realized outcome affects choices). Suppose, however, that in advance of participating in the program, persons know their own  $(Y_0, Y_1)$  values but that observers do not. For such persons, given  $d = 1$ ,

$$Y_1 \geq Y_0$$

is a requirement for rationality. In the population, the requirement becomes

$$\Pr(Y_1 \geq Y_0 \mid Y_0 = y_0, d = 1) = 1.$$

This is a strong form of stochastic dominance. All of the mass of the  $Y_1$  distribution is to the right of  $y_0$ .

More generally, persons may not know  $(Y_0, Y_1)$  but may make unbiased guesses  $(Y_0^*, Y_1^*)$  about them in calculating program gains. In this case

$$Y_0^* = Y_0 + \varepsilon_0$$

and

$$Y_1^* = Y_1 + \varepsilon_1$$

where

$$E(\varepsilon_0, \varepsilon_1) = (0, 0)$$

and

$$(\varepsilon_0, \varepsilon_1) \perp\!\!\!\perp (Y_0, Y_1),$$

and " $\perp\!\!\!\perp$ " denotes independence. In this case, conditioning on realized values produces positive regression dependence (PRD) between  $Y_1$  and  $Y_0$  so that

$$\Pr(Y_1 \leq y_1 \mid Y_0 = y_0, d = 1) \text{ is non-increasing in } y_0 \text{ for all } y_1.$$

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 is  $y_0$ , 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  $Y_1$ ".  $Y_1$  being right tail increasing given  $y_0$  implies that  $Y_1$  and  $Y_0$  (given  $d = 1$ ) are positive

quadrant dependent:  $\Pr(Y_1 \leq y_1 \mid Y_0 \leq y_0, d = 1) \geq \Pr(Y_1 \geq y_1 \mid d = 1)$  and  $\Pr(Y_0 \leq y_0 \mid Y_1 \leq y_1, d = 1) \geq \Pr(Y_0 \leq y_0 \mid d = 1)$ . These implications are strict except in the case where  $Y_0$  and  $Y_1$  are binary random variables. In that case, these notions of dependence are all equivalent. (See, e.g., Tong (1980)). Common measures of dependence like the product-moment correlation, Kendall's tau and Spearman's rho are all positive when there is positive quadrant dependence.<sup>1</sup> Thus rationality imposes a restriction on the nature of the dependence between  $Y_0$  and  $Y_1$  given  $d = 1$ . Evidence against such dependence is evidence against the Roy model. (See, e.g., Heckman and Honore (1990), for an exposition of the Roy model). Even if  $Y_0$  and  $Y_1$  are negatively correlated in the population, they are positively correlated given  $d = 1$  if agents are income maximizers.

Finally, consider persons who do not know in advance either  $Y_1$  or  $Y_0$  and who do not guess about these values in the manner suggested by the preceding example, but who know the joint distribution of outcomes for participants ( $F(y_0, y_1 \mid d = 1)$ ). Persons randomized into the program who receive outcome  $Y_1 = y_1$  would express no ex-post regret about participating in the program if

$$U(y_1) \geq \int U(y) dF_0(y \mid y_1, d = 1).$$

Persons randomized out would express regret in being randomized out and receiving  $Y_0 = y_0$  if

$$U(y_0) \leq \int U(y) dF_1(y \mid y_0, d = 1).$$

Information about ex-post regret combined with knowledge of  $F(y_0, y_1 \mid d = 1)$  would enable

---

<sup>1</sup>The result on the correlation coefficient follows trivially from Hoeffding (1940) who proves that  $\text{Cov}(x, y) = \int \int [F(x, y) - F_1(x)F_2(y)] dx dy$  where  $F$  is the joint distribution,  $F_1$  is the marginal distribution for  $X$ ,  $F_2$  is the marginal distribution for  $y$ .

analysts to test for this form of agent rationality. This test requires knowledge of the joint distribution of outcomes.

*(3) Indeterminacy In Social Experiments: The Continuous Case*

Assume access to a sample of N individuals in the treatment state and N in the non-treatment state. Suppose that the outcomes are continuously distributed and that (A-1) and (A-2) are valid so that there is no randomization bias. Ranking the individuals in order of their outcome value from the highest to the lowest, so that  $Y_j^{(i)}$  is the  $i^{\text{th}}$  highest-ranked person in the  $j$  distribution, and ignoring all ties, we obtain two data distributions:

Treatment Outcome:  $F(y_1 | d=1)$

Non-Treatment Outcome:  $F(y_0 | d = 1)$

$$\underline{Y}_1 = \begin{pmatrix} Y_1^{(1)} \\ \vdots \\ Y_1^{(N)} \end{pmatrix}$$

$$\underline{Y}_0 = \begin{pmatrix} Y_0^{(1)} \\ \vdots \\ Y_0^{(N)} \end{pmatrix}$$

We know the marginal data distributions  $F(y_1 | d = 1)$  and  $F(y_0 | d = 1)$  but we do not know where person  $i$  in the treatment distribution would appear in the non-treatment distribution. Corresponding to the ranking of the treatment outcome distribution, there are  $N!$  possible patterns of outcomes in the associated non-treatment outcome distribution. By considering all such possible permutations, we can form a collection  $C$  of possible gain distributions, i.e. alternative distributions of

$$\underline{Y}_1 - \Pi_\ell \underline{Y}_0 \quad \ell = 1, \dots, N!$$

where  $\Pi_\ell$  is a particular  $N \times N$  permutation matrix of  $\underline{Y}_0$  in the set of all  $N!$  permutations associating the ranks in the  $\underline{Y}_1$  distribution with the ranks in the  $\underline{Y}_0$  distribution. 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.

Collection C coincides with the set of extreme points of the set S of all cumulative distribution functions having  $F(y_1 | d = 1)$  and  $F(y_0 | d = 1)$  as marginal distributions. (Whitt (1976)). If the  $Y_1$  and  $Y_0$  are all distinct, then C corresponds to the set of all  $N \times N$  permutation matrices while the set of all cumulative distribution functions corresponds to the set of all  $N \times N$  doubly stochastic matrices. Moreover, the data distributions are dense in the space of all probability measures in the topology of weak convergence. Thus in the limit as  $N \rightarrow \infty$ , we can obtain any admissible bivariate distribution that lies in S by operating on C using doubly stochastic matrices. In other words, C is the convex hull of S. To simplify the analysis we work with the data distributions, passing to the limit as required.

In the case of the "dummy endogenous variable" model or "additive unit treatment model" - the model that assumes a constant treatment effect for all persons - or for a more general model in which the treatment outcome is a deterministic function of observable variables, there is only one admissible permutation:

$$\Pi = I .$$

The best in one distribution is the best in the other distribution. In the additive treatment case,  $Y_1$  and  $Y_0$  differ by a constant for each person. The common effect model assumes that the treatment effect is identical at all quantiles of the gain distribution. A generalization of that model preserves perfect dependence in the ranks between the two distributions but does not require the impact to be the same at all quantiles of the base state distribution. Equating quantiles across the two distributions, form pairs

$$\{(y_0, y_1) \mid \inf_{y_1} F_1(y_1 | d=1) > q \text{ and } \inf_{y_0} F_0(y_0 | d=1) > q, \quad 0 \leq q \leq 1\}$$

and obtain a deterministic gain function:

$$\Delta(y_0) = y_1(y_0) - y_0.$$

For the case of absolutely continuous distributions with positive density at  $y_0$  the gain function can be written as

$$\Delta(y_0) = F_1^{-1}(F_0(y_0 | d = 1)) - y_0.$$

Using standard methods, we can use experimental data to test non-parametrically for the classical common effect model. Observe that we can form other pairings across quantiles by mapping quantiles from the  $Y_1$  distribution into quantiles from the  $Y_0$  distribution using the map  $T$ :

$$T: q_1 \rightarrow q_0.$$

The 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. They cannot reject any of these models or more general models that permit nondegenerate  $(Y_0, Y_1)$  distributions. We present estimates of  $\Delta(y_0)$  after describing our data.

*(a) The Data Used to Estimate  $\Delta(y_0)$*

The data analyzed in this paper were gathered as part of an experimental evaluation of the training programs financed under Title II-A of the Job Training Partnership Act (JTPA). This program provides classroom training in occupational skills, on-the-job training at private firms, job search assistance, and other employment and training services to the economically disadvantaged.

The experiment was conducted at a non-random sample of sixteen of the more than 500 JTPA training sites around the country. 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 site. 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 programs, schooling, and training during the period after random assignment. The response rate for this survey was around 84 percent. The sample used here includes only those adult women who (1) had a follow-up interview scheduled at least 18 months after random assignment, (2) responded to the survey, and (3) had useable earnings information for the 18 months after random assignment. This sub-sample 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 16<sup>th</sup>, 17<sup>th</sup> and 18<sup>th</sup> months after random assignment. A person is defined to be employed if she had any self-reported earnings in these months.

*(b) Estimates*

Figure 1A presents empirical evidence on the question of the constancy of the gain effect across quantiles. It displays the estimate of  $\Delta(y_0)$  for adult women assuming that the best persons in the "1" distribution are the best in the "0" distribution. More formally, it assumes that the permutation matrix  $\pi = I$ . Between the 25th and 85th quantiles the assumption of a constant impact is roughly correct. It is grossly at odds with the data at the highest and lowest quantiles. (Standard errors for the quantiles are obtained using standard methods described in Csörgo (1983)). Disaggregating by race, we obtain distinctive patterns. For white females, Figure 1B shows a pattern of increasing levels of the gain at the higher quantiles. There is a similar but less dramatic pattern for black females in Figure 1C. For hispanics in Figure 1D there is little evidence of constancy - indeed there is a sharp decline at the higher quantiles. There are dramatic differences in the gain at different quantiles for different schooling levels (see Figs. 1E-1G). For the less schooled, there is strong evidence of a greater impact of training at the higher base income levels. There is also a sharp rise in the estimated treatment effect in terms of the base state income for the most schooled while for high school graduates the pattern is basically flat.

*(4) A Generalization and A Measure of Intrinsic Uncertainty*

One plausible generalization of the strict preservation of ranks model permits there to be some slippage in the ranks of a person's position in the two distributions. With slippage, the best

in distribution  $F_0(y_0 | d = 1)$  may be near the top of distribution  $F_1(y_1 | d = 1)$ , but need not be exactly at the top. Such positive dependence can arise from the utility-maximizing models of Section (2) when  $Y_1$  and  $Y_0$  are net outcomes (up to some independent costs or error). Formally, consider permutations that are restricted to satisfy

$$\Pi_{ij} = 0 \text{ for } |i - j| > \varphi, \quad \varphi \geq 1$$

where  $\varphi$  is some specified positive integer number of steps away from perfectly matched ranks in the two distributions. By reducing  $\varphi$ , we increase positive dependence and in the limit ( $\varphi = 1$ ) achieve perfect matching - the case that includes the traditional common treatment effect model that guides most evaluation research. Setting  $\varphi = N$  produces all possible distributions including the possibility that the best  $Y_1$  may be the worst  $Y_0$ . We develop this idea below.

To gauge the intrinsic uncertainty in the data, we assign equal weight to all permutations in the data. Using the sample outcome distributions we can pair each  $Y_1$  with each possible  $Y_0$  and in this way generate all possible permutational contrasts. The generated distributions can be used to produce sample gain distributions for different assumed levels of disarray in the matching across the distributions. In practice, two complications preclude the direct application of this idea.

(A) There are unequal numbers in the two distributions ( $N_{Y_1} \neq N_{Y_0}$ ). To circumvent this problem, we propose permutation of quantiles of the two distributions using the distribution with the smallest number of observations to set the spacing in the distribution with the large number. Then  $\min(N_{Y_0}, N_{Y_1})$  is the number of quantile spacings considered in the distribution with more observations. All elements in a given quantile class so determined would be treated as the same, e.g. 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.

(B) For  $N$  sufficiently big, it is computationally demanding to consider all possible permutations. To solve this problem, we propose collapsing both distributions down to a small number of quantile classes and using mean values within each quantile class to summarize the class. Permutations are then done with respect to the reduced classes. Such permutations obviously understate the full range of values that could be obtained from the original distribution.

Using percentiles as the finest quantile partition, we obtain  $100!$  possible different impact or earnings gain distributions. Without any prior information, any one of these permutation patterns is equally likely. To examine the variation present in the experimental data, 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 gain distributions corresponding to this sample of permutations. Table 1B presents means and selected quantiles of the distributions of other parameters of interest for this sample of permutations, including the fraction with a positive impact, the impact standard deviation, and several measures of the dependence between  $Y_0$  and  $Y_1$ . Appendix A describes the construction of these statistics in greater detail.

The numbers in Table 1A and 1B reveal substantial variability in the quantiles of the gain distribution within the sample of permutations. 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 over  $\$2500$  in this sample. This table understates the

true variation in the population 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 permutations in which either (1) the two 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 gain distribution equal to \$0 in the first case and -\$22,350 in the other, and the 95th percentile of the gain distribution equal to \$2003 in the first case and \$23,351 in the second.

Without additional information, the evidence from experimental data is consistent with a broad array of distributions of program impacts. In order to narrow down this class, additional information is required. This paper considers a variety of additional plausible assumptions that help to narrow the class of admissible distributions. Before turning to the list of candidate assumptions, we first review existing statistical approaches to bounding features of the distribution of program gains only using the information in  $F_0(y_0 | d = 1)$  and  $F_1(y_1 | d = 1)$ .

*(5) The Information About The Features of The Distribution of Program Gains  
From An Ideal Randomized Experiment: Results From The Statistics Literature*

The problem of bounding the joint distribution  $F(y_1, y_0 | d = 1)$  from the marginal distributions  $F(y_1 | d = 1)$  and  $F(y_0 | d = 1)$  is a classical problem in mathematical statistics. Hoeffding (1940) and Frechet (1951) demonstrate that the joint distribution is bounded by two functions of the marginal distributions:

$$\begin{aligned} \text{Max}[F_1(y_1 | d=1) + F_0(y_0 | d=1) - 1, 0] &\leq F(y_1, y_0 | d=1) \\ &\leq \text{Min}[F_1(y_1 | d=1), F_0(y_0 | d=1)].^2 \end{aligned}$$

Rüschendorf (1982) establishes that these bounds are tight (exhaust the information in the marginals). 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 bounds are not helpful in bounding the distribution of  $\Delta = Y_1 - Y_0$ , although they bound certain features of it.

Using the upper and lower bounding distributions, the literature establishes that if  $k(Y_1, Y_0)$  is superadditive (or subadditive) then extreme values of

$$E(k(Y_1, Y_0) | d = 1)$$

are obtained from the upper and lower bounding distributions.

---

<sup>2</sup>These inequalities are easy to establish. The upper inequality arises from the observation that the probability of the joint event

$$\Pr(Y_1 \leq y_1, Y_0 \leq y_0 | d = 1)$$

can be no more than the probability of each single event,  $\Pr(Y_1 \leq y_1 | d = 1)$  and  $\Pr(Y_0 \leq y_0 | d = 1)$ , and hence is less than or equal to the minimum of these two probabilities.

The lower bound on the distribution is obtained from the following argument. The probability space can be partitioned into the following probability associated with the region  $(Y_1 \leq y_1, Y_0 \leq y_0)$  and its complement:

$$\begin{aligned} 1 &= \Pr(Y_1 \leq y_1, Y_0 \leq y_0 | d = 1) \\ &+ \Pr(Y_1 > y_1 | d = 1) + \Pr(Y_0 > y_0 | d = 1) \\ &- \Pr(Y_1 > y_1, Y_0 > y_0 | d = 1). \end{aligned}$$

The final three terms are the probability of the set complementary to  $(Y_1 \leq y_1, Y_0 \leq y_0)$ . The final term in this expression corrects for the overlap in the two sets  $(Y_1 > y_1)$  and  $(Y_0 > y_0)$  in evaluating the complementary set. Substituting  $\Pr(Y_1 \leq y_1, Y_0 \leq y_0 | d = 1) = F(y_1, y_0 | d = 1)$ ,  $1 - F(y_0 | d = 1) = \Pr(Y_1 > y_1 | d = 1)$ ,  $1 - F(y_1 | d = 1) = \Pr(Y_0 > y_0 | d = 1)$  and recognizing that  $\Pr(Y_1 > y_1, Y_0 > y_0 | d = 1) \geq 0$  and that probability distributions cannot become negative, we obtain the lower bound.

A function is superadditive if for all  $Y_1 \leq Y'_1, Y_0 \leq Y'_0$

$$0 \leq k(Y_1, Y_0) + k(Y'_1, Y'_0) - k(Y_1, Y'_0) - k(Y'_1, Y_0)$$

and subadditive if the inequality is reversed.<sup>3</sup>

Examples of superadditive functions are

$$\begin{aligned} k(Y_1, Y_0) &= Y_1 Y_0 \quad \text{or} \\ &= (Y_1 + Y_0)^2 \quad \text{or} \\ &= \text{Min}(Y_1, Y_0) \quad \text{or} \\ &= f(Y_1 - Y_0) \quad \text{where } f \text{ is concave and continuous.} \end{aligned}$$

Examples of subadditive functions are

$$\begin{aligned} k(Y_1, Y_0) &= |Y_1 - Y_0|^p \quad \text{for } p \geq 1 \quad \text{or} \\ &= \text{Max}(Y_1, Y_0) \quad \text{or} \\ &= f(Y_1 - Y_0) \quad \text{where } f \text{ is convex and continuous.} \end{aligned}$$

The indicator function  $1(Y_1 \geq Y_0)$  and the quantiles of the impact distribution are neither superadditive nor subadditive.

Cambanis, Simons and Stout (1976) have established the following theorem which demonstrates the usefulness of the bounding distributions for producing bounds on certain subadditive or superadditive statistics produced from the joint distribution.

**Theorem:** Let  $E_+$  denote the expectation with respect to the upper bound distribution and let  $E_-$  denote the expectation with respect to the lower bound distribution. Then if  $k(Y_1, Y_0)$  is superadditive (subadditive), the bound on  $E(k(Y_1, Y_0) | d = 1)$  is given by

---

<sup>3</sup>k is assumed to be Borel-measurable and right-continuous.

$$E_-(k(Y_1, Y_0) | d = 1) \leq E(k(Y_1, Y_0) | d = 1) \leq E_+(k(Y_1, Y_0) | d = 1)$$

or

$$[E_+(k(Y_1, Y_0) | d = 1) \leq E(k(Y_1, Y_0) | d = 1) \leq E_-(k(Y_1, Y_0) | d = 1)], \text{ respectively,}$$

if either

(i)  $k(Y_1, Y_0)$  is symmetric and the expectations  $E(k(Y_1, Y_1) | d = 1)$  and

$E(k(Y_0, Y_0) | d = 1)$  are finite; or

(ii) the expectations of  $E(k(Y_1, y_0) | d = 1)$  and  $E(k(y_0, Y_1) | d = 1)$  are finite

for some  $y_0, y_1$ , and at least one of  $E_+(k(Y_1, Y_0) | d = 1)$  and

$E_-(k(Y_1, Y_0) | d = 1)$  are finite. ■

**Proof:** See Theorem 2, p. 291 of Cambanis et al (1976). Tchen (1980) establishes these and other results under different conditions. ■

Expectations of superadditive or subadditive functions of  $(Y_1, Y_0)$  attain extreme values at the boundary distributions and so can be computed from the marginal distributions. Since  $k(Y_1, Y_0) = Y_1 Y_0$  is superadditive, the maximum attainable product-moment correlation  $r_{Y_0 Y_1}$  is obtained from the upper bound distribution while the minimum attainable product moment correlation is obtained at the lower bound distribution. Since  $\text{VAR}(\Delta)$  is a subadditive function, it is possible to bound the variance of  $\Delta (= \text{VAR}(Y_1) + \text{VAR}(Y_0) - 2r_{Y_0 Y_1} [\text{VAR}(Y_1)\text{VAR}(Y_0)]^{1/2})$  and thus determine if the data are consistent with  $Y_1 - Y_0 = \alpha$ , a constant, in which case  $\text{VAR}(\Delta) = 0$ . Tchen (1980) establishes that Kendall's  $\tau$  and Spearman's  $\rho$  also attain their extreme values at the bounding distributions. However, in general it is not the case that useful bounds on the quantiles of the  $(Y_1 - Y_0)$  distribution can be derived from the Frechet-Hoeffding

bounds. Only the extreme high and extreme low quantile values are obtained from the Frechet bounds of the joint distribution. Table 3 presents the range of values of  $r_{Y_1Y_0}$ , Kendall's tau, Spearman's  $\rho$  and  $[\text{VAR}(\Delta)]^{\text{th}}$  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  $\text{VAR}(\Delta)$  is bounded away from zero. They obviously do not rule out the deterministic case  $\Delta(y_0)$  as long as  $\Delta$  is not a constant. Before turning to methods for adding information so as to narrow the class of admissible distributions, it is instructive to consider the case of a discrete - data  $2 \times 2$  contingency table where the case for application of the Frechet-Hoeffding bounds is the most favorable.

#### *(6) The Discrete Case*

The Frechet-Hoeffding bounds apply to all bivariate outcome distributions.<sup>4</sup> Variables may be discrete, continuous or both discrete and continuous. In order to fix ideas, it is helpful to consider the most elementary case: that of a discrete outcome variable such as employment. It turns out to be the case most favorable for the application of the Frechet-Hoeffding bounds, and so in this sense is misleading. Yet an analysis of the simple case is a fruitful point of departure for the analysis used in the rest of this paper.

Those who enroll in a program may be employed or not employed after completing it. Those who are randomized out of a program may also be employed or not employed in the evaluation period. The following analysis is a simple application of the missing-cell literature in contingency table analysis such as that given in Bishop, Fienberg and Holland (1975).

The latent distribution underlying this situation is bivariate binomial. Let  $(E, E)$  denote

---

<sup>4</sup>Formulae for multivariate bounds are given in Tchen (1980) or Rüschendorf (1982).

the event "employed with treatment" and "employed without treatment". (E,N) is the event "employed with treatment, not employed without treatment". Similarly, (N,E) and (N,N) refer respectively to cases where (a) a person would not be employed if treated but would be employed if not treated, and (b) the person would not be employed in either case. 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 nonemployment in the untreated state. The rows refer to employment and non-employment in the treated state.

		Untreated		
		E	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

The evaluation problem arises from the fact that we do not observe the same person in both the treated and untreated states. If we did, we could fill in the table and estimate the full distribution. Instead, with data from randomized trials we can estimate combinations of the table parameters

(2a)  $P_{E\cdot} = P_{EE} + P_{EN}$  (Employment Proportion Among Treated)

(2b)  $P_{\cdot E} = P_{EE} + P_{NE}$  (Employment Proportion Among Untreated).

The treatment effect is usually defined as

$$(3) \quad T = P_{EN} - P_{NE},$$

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 receiving the treatment.

T is easily seen to be equal to

$$T = P_{E\bullet} - P_{\bullet B}$$

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

If we wish to decompose T into its two components, 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. The case in the  $2 \times 2$  table corresponding to the common effect model for continuous outcomes restricts the discrete outcomes to be either positive or negative so that either  $P_{EN}$  or  $P_{NE} = 0$ . In this case, the model becomes fully identified just as the common effect assumption in the continuous case fully identifies the joint distribution.

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

$$\text{Max}[P_{E\bullet} + P_{\bullet B} - 1, 0] \leq P_{EE} \leq \text{Min}(P_{E\bullet}, P_{\bullet B})$$

$$\text{Max}[P_{E\bullet} - P_{\bullet B}, 0] \leq P_{EN} \leq \text{Min}(P_{E\bullet}, 1 - P_{\bullet B})$$

$$\text{Max}[-P_{\bullet B} + P_{\bullet B}, 0] \leq P_{NE} \leq \text{Min}(1 - P_{E\bullet}, P_{\bullet B})$$

$$\text{Max}[1 - P_{E\bullet} - P_{\bullet B}, 0] \leq P_{NN} \leq \text{Min}(1 - P_{E\bullet}, 1 - P_{\bullet B}).$$

Table 4 presents the Frechet bounds for  $P_{NE}$  and  $P_{EN}$ . They are very wide. Even without

taking into account sampling error, the experimental evidence for adult females is consistent with  $P_{NE}$  ranging from .00 to .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 prospects improved by participating in the training program. As many as 36% and as few as 0% may have had their employment prospects diminished by participating in the program. 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%  $[(.64 - .39) \times 100]$  and as many as 61% of the women would have worked whether or not they entered the program ( $P_{EE} \in (0.25, .61)$ ).

From this evidence, we cannot distinguish between two stories. The first story is that the JTPA program benefits many people by facilitating their employment but it also harms many people is that they are less likely to work than if they had not participated. The second story is that the program benefits and harms few people. Conditioning on other background variables (results not shown) does not go far in resolving the fundamental uncertainty intrinsic in the data.

*(7) Using Prior Information To Reduce The Intrinsic Uncertainty in Social Experiments: The Case of The 2x2 Table*

While disappointing, it is not surprising that classical probability theory unaided by prior information only weakly restricts the range of admissible estimates of the distribution of program gains. Previous studies by Lavine, Wasserman and Wolpert and others demonstrate that the Frechet-Hoeffding bounds are rather wide in the analysis of data from clinical trials.

In many problems - and certainly in the problem studied in this paper - we have additional information and intuition about the likely relationship between program and non-program outcomes for each person. The Bayesian paradigm provides a convenient way to

formalize this intuition and apply it to an analysis of the experimental data. In this paper we are influenced by the robust Bayesian paradigm but we adopt a new class of priors. We consider ranges of priors that capture the loose information provided by our intuition. The goal is to partially reduce the uncertainty about the distribution of program gains, and features of the distribution of program gains, without imposing false prior information onto the data and without distorting the sample information revealed by the experiment. (See, e.g. Hartigan (1983) or Berger (1985, 1990) for discussions of robust Bayesian analysis). The strength of the prior required to obtain credible estimates from the available experimental data is a measure of the lack of information in the data.

In considering outcomes like employment and earnings, plausible models of program 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. As discussed in Section (2), positive regression dependence for  $Y_1$  given  $Y_0$  and  $d = 1$  is a consequence of many models of rational choice. This presumption is strengthened by the widely-held intuition that more motivated or more able persons apply to programs and are better at what they do regardless of what it is than less able or motivated persons. Willis and Rosen (1979) feature this case in their influential paper on college choice calling it the "one-factor" model. Less than perfect positive dependence is the natural point of departure from the assumption of perfect positive dependence that characterizes the classical econometric evaluation literature.

In order to make this notion operational it is necessary to specify more precisely what we mean by dependence. Notions of dependence in  $2 \times 2$  tables are presented in Goodman and Kruskal (1979) and in Bishop, Feinberg and Holland (1975). In terms of the table in Figure 2,

the most commonly used measure of association between the two outcomes is the cross product ratio:

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

When  $\alpha = 1$ , the treatment and non-treatment outcomes are independent. This measure has several advantages:

- (i) It is invariant under the interchange of rows and columns.
- (ii) It is invariant to the proportion of persons participating in the program.
- (iii) It is interpretable and is the ratio of the odds of being employed in the no-program state conditional on being employed in the program state ( $P_{EE}/P_{EN}$ ) and the odds of employment in the no-program state conditional on not being employed in the program state ( $P_{NE}/P_{NN}$ ).

By property (iii), the higher is  $\alpha$ , the more likely it is for a person employed in the program state to be employed in the no-program state. As the conditional (on employment in the program state) odds ratio of employment in the no-program state ( $P_{EE}/P_{EN}$ ) becomes large and the conditional (on no employment in the program state) odds ratio of employment in the no-program state becomes small,  $\alpha$  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,  $\alpha \rightarrow 0$ . Thus  $\alpha$  is an attractive measure of the association of outcomes for the  $2 \times 2$  table.

It is well known that in the  $2 \times 2$  table many diverse notions of positive dependence are equivalent. Positive covariance, association, positive regression dependence, right tail increasing

dependence and quadrant dependence all describe the same positive "association" of E and N. See, e.g., Esary, Proschan and Walkup (1967). Thus there is no loss in generality in using  $\alpha$  or its transform Q defined below.

Given  $\alpha$ , the row and column marginals  $P_{E\cdot}$  and  $P_{\cdot N}$ , and the requirement that the probabilities sum to one, we can uniquely determine the elements of the  $2 \times 2$  table. For adult women from the JTPA experiment, Figure 3 presents this relationship in terms of Yule's measure of association Q, which lies in the interval [-1, 1]:

$$Q = \frac{\alpha - 1}{\alpha + 1}$$

where  $Q = 0$  when  $\alpha = 1$  (and the rows and columns are independent). Higher values of Q are associated with greater dependence in outcomes in the two states (i.e. higher values of  $\alpha$ ). As we specify higher values of Q we reduce both  $P_{EN}$  and  $P_{NE}$  in absolute value. The difference  $P_{EN} - P_{NE}$  is the mean treatment effect T and is constant over all Q.

Intuitions that outcomes are strongly positively related across the two states translate into statements that Q is positive and close to one. Rather than picking a specific value of Q, it seems more plausible to present a weighted average value of  $P_{EN}$  and  $P_{NE}$ , placing more weight on positive values of Q. It seems unreasonable to place too much emphasis on  $Q = 1$ ; those employed in one state are only deemed more likely to be employed in the other state. Accordingly, it seems appropriate to use a prior or weighting function on Q that peaks in the interval (.5, 1) and that allows for some slippage in status between the two states. We use a spline prior of the general form presented in Figure 4 to produce posterior point estimates of the elements in the table. This prior places most of its weight on values of Q that are positive and

bigger than 0.5 but does not place too much weight on 1. Using this prior, and varying its peak and slopes, we produce a range of estimates for  $P_{EN}$ ,  $P_{EE}$ ,  $P_{NE}$  and  $P_{NN}$  all of which have the same mean treatment effect:

$$T = P_{EN} - P_{NE} .$$

For the four priors displayed in Table 5, posterior estimates are shown in Table 6. Our procedure places no restrictions on any combination of parameters that can be identified using the experimental data. Instead we supplement the experimental data by postulating priors on unidentified parameters. We thereby improve on the Hoeffding-Frechet bounds and reduce the range of uncertainty about other unidentified parameters of interest.

#### *(8) Extensions To The Continuous Case*

Plackett (1965) presents a bivariate distribution system with given marginal distributions  $F_1$  and  $F_0$  that provides a fruitful point of departure for investigating the continuous case. Plackett's distribution system is generated by one additional parameter  $\psi$  that plays a role comparable to  $\alpha$  in the  $2 \times 2$  table.

Any bivariate distribution  $F(y_1, y_0 | d = 1)$  can be dichotomized at some arbitrary point  $(\tilde{y}_1, \tilde{y}_0)$ . Setting

$$p_{11} = F(\tilde{y}_1, \tilde{y}_0)$$

$$p_{12} = F_1(\tilde{y}_1 | d = 1) - F(\tilde{y}_1, \tilde{y}_0 | d = 1)$$

$$p_{21} = F_0(\tilde{y}_0 | d = 1) - F(\tilde{y}_1, \tilde{y}_0 | d = 1)$$

$$p_{22} = 1 - F_1(\tilde{y}_1 | d = 1) - F_0(\tilde{y}_0 | d = 1) + F(\tilde{y}_1, \tilde{y}_0 | d = 1)$$

we can write the cross product ratio as

$$\psi = \frac{p_{11} p_{22}}{p_{12} p_{21}} = \frac{F(\bar{y}_1, \bar{y}_0 | d=1)[1 - F_1(\bar{y}_1 | d=1) - F_0(\bar{y}_0 | d=1)]}{[F_1(\bar{y}_1 | d=1) - F(\bar{y}_1, \bar{y}_0 | d=1)][F_0(\bar{y}_0 | d=1) - F(\bar{y}_1, \bar{y}_0 | d=1)]}$$

which is clearly analogous to the parameter  $\alpha$ , introduced in the analysis of  $2 \times 2$  contingency tables. For each  $\psi$  we can generate a joint distribution  $F(\bar{y}_1, \bar{y}_0 | d = 1)$  from  $\psi$ ,  $F_1$  and  $F_0$ . Mardia (1970) establishes that:

- (i)  $F(\bar{y}_1, \bar{y}_0 | d = 1)$  is a proper probability distribution
- (ii)  $F(\bar{y}_1, \bar{y}_0 | d = 1)$  attains the Frechet-Hoeffding upper bound as  $\psi \rightarrow \infty$  and it attains the lower bound as  $\psi \rightarrow 0$ .
- (iii)  $\psi = 1$  corresponds to independence,  $\psi > 1$  corresponds to positive quadrant dependence, and  $\psi < 1$  corresponds to negative quadrant dependence where quadrant dependence is defined below.

In this parameterization,  $\psi$  is the same constant for all values of  $(y_1, y_0)$  and is a measure of quadrant dependence. Recall that  $F(y_1, y_0)$  is positive (negative) quadrant dependent if  $F(y_1, y_0 | d = 1) \geq (\leq) F_1(y_1 | d = 1)F_0(y_0 | d = 1)$ . An immediate consequence is that  $Y_1$  and  $Y_0$  are positively correlated (given  $d = 1$ ).

Kendall's  $\tau$ , Spearman's  $\rho$  and the product-moment correlation  $r$  all increase as a distribution becomes more positively quadrant dependent. Thus as  $\psi$  increases, those conventional measures of dependence between  $Y_1$  and  $Y_0$  increase. As shown in Section (2), under conditions of partial information about  $Y_1$  and  $Y_0$ , positive quadrant dependence (given  $d = 1$ ) is a consequence of rational choice by income maximizing agents who make unbiased guesses about their post-program outcomes although actually a stronger form of dependence is implied by rational choice behavior.

Transforming  $\psi$  into the interval  $[-1, 1]$  using Blomquist's  $Q$

$$Q = \frac{\psi - 1}{\psi + 1}$$

produces a bivariate system that "fills in" the missing data in a manner that depends on the strength of the quadrant dependence between  $Y_0$  and  $Y_1$ . When  $Q = 1$ , we obtain the Fréchet-Hoeffding upper bound, where  $Y_0$  and  $Y_1$  are functionally related.

Table 7A presents quantiles of the impact distribution for each several values of  $Q$  in the range  $[-1, 1]$ . Table 7B presents other parameters including the percentage of those treated with a positive impact, the impact standard deviation, and several measures of dependence for the same set of  $Q$  values. Details on the construction of these estimates and their standard errors appear in Appendix B.

Eliminating the negative quadrant dependent distributions eliminates the extreme variability in the quantiles of the gain distribution. However, Table 7A reveals that it is necessary to assume a high level of dependence between the treatment and non-treatment outcomes to produce plausible variation in the 95-5 or 75-25 range in the impact distribution.

One way to explore the uncertainty in the data is to impose prior beliefs about  $Q$ . For various specifications of the spline prior depicted in Figure 4, we produce estimates of parameters of the gain distribution for different priors over  $Q$ . As before, these priors place all (or almost all) of their weight on positive values of  $Q$ . Table 8 presents the posteriors resulting from applying the priors displayed in Table 5. Figures 5,6,7 show how the median, the 75<sup>th</sup> percentile and the fraction benefitting from the program vary with the posterior mean of  $Q$  over a larger set of 19 prior distributions listed in Appendix C. The variability in the experimental

data is greatly reduced when spline priors emphasizing strong positive dependence are imposed.

An important limitation of this approach is that the dependence relationship between  $Y_0$  and  $Y_1$  given  $d = 1$  is very tightly specified. Although Blomquist's  $Q$  is analogous to the dependence relationship specified by Yule's  $Q$  for the  $2 \times 2$  table, as we trace out Yule's  $Q$  over  $[-1, 1]$ , we recover all  $2 \times 2$  Tables. We do not recover all bivariate distributions using Plackett's distribution when we vary Blomquist's  $Q$  over the same interval. Put more formally, the Plackett family is not dense in the space of all bivariate distributions. There are many bivariate distributions for continuous data that do not have the same value of  $Q$  for all  $y_0, y_1$  values, which is the defining property of the Plackett class. A more general approach specifies priors over the permutations of the data distributions introduced in Section (3). We turn to this approach next.

#### *(9) Putting Priors Over Permutations of The Data Distributions*

The evidence presented in Section (3) suggests that the range of impacts estimated from all permutations of the data distributions is rather wide. As in the preceding section, it is useful to consider mild perturbations away from the case of perfect positive ranking of  $Y_0$  and  $Y_1$ . In this section, we consider a measure of disarray from perfect ranking that characterizes all possible bivariate data distributions and that is based on a somewhat more intuitively satisfying measure of dependence than Blomquist's  $Q$ . We assume in this section that the data are from absolutely continuous distributions and we assume no ties in the sample distribution.

Consider any permutation of the  $Y_0$  associated with the  $Y_1$  via permutation rule  $\Pi$ .  $Y_1$  and  $Y_0$  are perfectly arrayed if  $\Pi = I$ . For other permutations, there is some level of disarray.

An inversion (relative to the perfectly increasing rank order assumed for  $Y_1$ ) is said to occur each time, in binary comparisons, an element in the  $Y_0$  array is bigger than a succeeding element, going down the full  $Y_0$  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). More generally, for any permutation of the  $Y_0$  associated with the  $Y_1$ , we can define the total number of inversions in the array as

$$V = \sum_j \sum_{i < j} h_{ij} \quad h_{ij} = \begin{cases} 1 & \text{if } Y_0^{(i)} > Y_0^{(j)} \\ 0 & \text{otherwise} \end{cases} .$$

$V$  may range from 0 to  $\frac{1}{2} N(N-1)$ , with the former value arising in the case of perfect positive dependence in the ranks and the latter value arising when there is perfect inverse ranking.

Kendall's rank correlation measure  $\tau$  may be written as

$$\tau = 1 - \frac{4V}{N(N-1)} = 1 - 4 \frac{\sum_j \sum_{i < j} h_{ij}}{N(N-1)} .$$

It normalizes  $\tau$  to lie in the interval [-1, 1]. (See, e.g. Kendall, (1970) and Daniels (1944, 1948)). All bivariate data distributions with the given marginals are produced by letting  $\tau$  vary over the entire interval. This is a major advantage by comparison with distributions produced from the Plackett class. While other measures of distance between permutations have been proposed, (Critchlow (1985)),  $\tau$  has many convenient properties and is a natural point of departure for our analysis.

Since  $h_{ij}$  is a superadditive function (fixing the  $Y_1$  ranking), and since sums of superadditive functions are superadditive, we know from the theorem stated in Section 5 that  $\tau$

attains its maximum value at the Frechet-Hoeffding upper bound and its minimum value at the Frechet-Hoeffding lower bound. Thus we can characterize the bounding distributions as producing minimal and maximal disarray between  $Y_1$  and  $Y_0$ . By specifying  $\tau$  we pick a level of dependence between the two outcomes and hence a level of permutational disarray.  $\tau$  is a measure of slippage in the ranks with  $\tau = 1$  corresponding to perfect rank correlation. Varying  $\tau$  between -1 and 1 traces out all possible permutational distributions.

Given two joint distribution  $F^{(a)}(y_1, y_0 | d = 1)$  and  $F^{(b)}(y_1, y_0 | d = 1)$ ,  $F^{(a)}$  is more concordant than  $F^{(b)}$  if they both share common marginals and  $F^{(a)} > F^{(b)}$  for all  $(y_1, y_0)$  i.e.  $F^{(a)}$  has more mass near the diagonal  $y_0 = y_1$  than  $F^{(b)}$ . Kendall's tau is higher for the more concordant distribution, as is intuitively satisfactory.<sup>5</sup> (See Schriever, 1987).

Using the sample distributions we can pair each  $Y_1$  with each possible  $Y_0$  and, subject to the limitations already discussed, generate all possible permutational contrasts. The generated distributions can be used to produce sample gain distributions for different assumed levels of disarray.

Table 9A and 9B present estimates of quantiles of the gain distribution and other parameters of interest conditional on various values of  $\tau$ . Table 10 presents measures of association and selected percentiles of the posterior gain distribution for the four priors listed in Table 5, now defined over  $\tau$  rather than  $Q$ . Priors 1 and 2, which place most of their weight on the strongly positively dependent permutations, reduce the variability in the percentile gain distributions the most. The percentile gains seem most credible for prior 1. Note, however, that

---

<sup>5</sup>Blomquist's  $Q$  shares the same property. However, the conventional product-moment correlation coefficient does not share this property.

by any measure of dependence, this prior requires a strong positive relationship between  $Y_1$  and  $Y_0$ . A central conclusion of this analysis and the analysis of the Plackett family is that plausible posterior gain distributions require high measures of positive dependence. The evidence from both the Plackett class and the class of priors placed on permutations indicates that a majority of the adult female participants benefitted from the program.

Our use of the quantiles of the implied gain distribution to calibrate the plausibility of a prior is indirect. It would be more direct to place priors on the quantiles of the gain distribution rather than to operate indirectly through alternative measures of dependence. The main problem with this approach is computational.

To understand the problem, observe that for absolutely continuous  $Y_1, Y_0$  the density of gains is obtained from the joint density of  $(Y_0, Y_1)$ :

$$f(\Delta) = \int f(y_0 + \Delta, y_0 | d = 1) dy_0.$$

The  $p^{\text{th}}$  quantile of the gain distribution is

$$\Delta_p = \inf_z F_{\Delta}(z) > p.$$

Any prior on  $\Delta_p$ , say  $g(\Delta_p)$ , imposes implicit restrictions on the dependence between  $Y_1$  and  $Y_0$  given the marginal distributions. Dependence between  $Y_1$  and  $Y_0$  is implicitly specified by selecting  $g$ , and it must be specified consistently.

Let  $K$  denote a function that combines the marginal densities using some measure of dependence  $\theta$  (which may be vector valued) to create a joint density.

$$K(f_1(y_0 + \Delta | d = 1), f_0(y_0 | d = 1); \theta).$$

In the Plackett family,  $\theta = \psi$  and  $K$  is specified as in Mardia. For the permutational families,

$\theta = \tau$ , and  $K$  is given by enumerating all possible permutational arrays. A major difficulty is inferring the prior for  $\theta$ , say  $\phi(\theta)$ , given the prior  $g(\Delta_p)$  for the quantiles. Even if this difficulty is overcome, we are left with computation of the posterior:

$$\int \int K(f_1(y_0 + \Delta | d = 1), f_0(y_0 | d = 1); \theta) dy_0 \phi(\theta) d\theta.$$

Deducing the prior on  $\theta$  and computing the posterior of  $\Delta$  are major computational problems. Our procedure of imposing priors on the dependence between  $Y_1$  and  $Y_0$  directly circumvents the first problem and simplifies the second. We use our prior information about the plausibility - or implausibility - of gains at certain quantiles to place more or less weight on the values of the dependence parameter  $\theta$  indicating a high level of dependence.

*(10) Allowing For Mass Points at Zero in The Population*

In many cases - and in particular for the JTPA earnings data - it is plausible that there are mass points at zero for  $Y_1$  and  $Y_0$ . (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 paper). The analysis of this case combines the analysis of Section (6) with the analysis of Section (9). 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.

Define the following notation: Let  $\Pr(Y_1 = 0, d = 1) = P_{0.1} > 0$  and  $\Pr(Y_0 = 0, d = 1) = P_{.0} > 0$ . The density of  $Y_1$  for  $Y_1 > 0$  is

$$f(y_1 | y_1 > 0, d = 1)$$

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

$$f(y_0 | y_0 > 0, d = 1).$$

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 both be discrete or continuous.

We propose the following three step generalization of the procedures used in Sections (6) and (9).

**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.

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

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

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

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

We know the left hand side of each of the following equations but the available population information does not afford a further resolution into the components on the right hand side.

$$P_{1\cdot} = P_{10} + P_{11} = \Pr(Y_1 > 0 | d = 1)$$

$$P_{0\cdot} = 1 - P_{1\cdot}$$

and

$$P_{\cdot 1} = P_{01} + P_{11} = \Pr(Y_0 > 0 | d = 1)$$

$$P_{\cdot 0} = 1 - P_{\cdot 1}.$$

Following the procedure outlined in Section (3), we can represent all of the possible  $2 \times 2$

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

$$f(y_1 | y_1 > 0, y_0 = 0, d = 1) \text{ and } f(y_0 | y_1 = 0, y_0 > 0, d = 1).$$

We know the left hand side of the following equations:

$$(4a) \quad f(y_1 | y_1 > 0) = f(y_1 | y_1 > 0, y_0 = 0, d = 1) \frac{P_{10}}{P_{11} + P_{10}} \\ + f(y_1 | y_1 > 0, y_0 > 0, d = 1) \frac{P_{11}}{P_{11} + P_{10}}$$

and

$$(4b) \quad f(y_0 | y_0 > 0) = f(y_0 | y_0 > 0, y_1 = 0, d = 1) \frac{P_{01}}{P_{11} + P_{01}} \\ + f(y_0 | y_0 > 0, y_1 > 0, d = 1) \frac{P_{11}}{P_{11} + P_{01}}$$

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

We may construct  $f(y_1 | y_1 > 0, y_0 = 0, d = 1)$  by weighting  $f(y_1 | y_1 > 0, d = 1)$ :

$$f(y_1 | y_1 > 0, y_0 = 0, d = 1) = f(y_1 | y_1 > 0, d = 1) w_1(y_1 | y_1 > 0)$$

where we require that  $w_1(y_1 | y_1 > 0) \geq 0$  and

$$1 = \int_0^{\bar{y}_1} f(y_1 | y_1 > 0, y_0 = 0, d = 1) dy_1 = \int_0^{\bar{y}_1} f(y_1 | y_1 > 0, d = 1) w_1(y_1 | y_1 > 0) dy_1.$$

Similarly we may construct

$$f(y_0 | y_1 = 0, y_0 > 0, d = 1) = f(y_0 | y_0 > 0, d = 1) w_0(y_0 | y_0 > 0)$$

with

$$w_0(y_0 | y_0 > 0) \geq 0 \text{ and}$$

$$1 = \int_0^{\infty} f_{Y_0|Y_0 > 0, d=1} w_0(y_0 | y_0 > 0) dy_0.$$

For consistency with (4a) and (4b), we require:

$$(5a) \quad f_{Y_1|Y_1 > 0, Y_0 > 0} = \frac{f_{Y_1|Y_1 > 0} \left(1 - \frac{w_1 P_{10}}{P_{11} + P_{10}}\right)}{\left(\frac{P_{11}}{P_{11} + P_{10}}\right)}$$

$$(5b) \quad f_{Y_0|Y_1 > 0, Y_0 > 0} = \frac{f_{Y_0|Y_0 > 0} \left(1 - \frac{w_0 P_{01}}{P_{11} + P_{01}}\right)}{\left(\frac{P_{11}}{P_{11} + P_{01}}\right)}$$

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 5(a) that

$$\frac{P_{11}}{P_{10}} + 1 \geq w_1 \geq 0 \quad \text{for all } y_1 \text{ in the support of } Y_1$$

and for (5b) that

$$\frac{P_{11}}{P_{01}} + 1 \geq w_0 \geq 0 \quad \text{for all } y_0 \text{ in the support of } Y_0.$$

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

$$\left( w_1, 1 - \frac{w_1 P_{10}}{P_{11} + P_{10}} \right)$$

and

$$\left( w_0, 1 - \frac{w_0 P_{01}}{P_{11} + P_{01}} \right)$$

bear a reciprocal relationship within each pair. For example, weighting  $f(y_1 | y_1 > 0)$  by placing more mass at the low values of  $y_1$  to obtain  $f(y_1 | y_1 > 0, y_0 = 0)$  ("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  $f(y_1 | y_1 > 0, y_0 > 0, d = 1)$ . 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

$$y_1 \in (0, \varepsilon)$$

$w_1^*$  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

$$\int_0^\varepsilon f(y_1 | y_1 > 0, d = 1) w_1^*(y_1 | y_1 > 0) dy_1 \geq \int_0^\varepsilon f(y_1 | y_1 > 0, d = 1) w_1^{**}(y_1 | y_1 > 0) dy_1.$$

If this relationship is true for all  $\varepsilon \in \text{Supp}(Y_1)$ , 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 (5a) and (5b) as the marginals for the permutation procedure developed in Section (9). ■

Proceeding in this fashion, we can specify priors over  $Q$ ,  $(w_1, w_0)$ , and the inversion classes  $\tau$  for  $Y_1 > 0, Y_0 > 0$  to produce posteriors on the joint distribution of  $(Y_1, Y_0)$ . Presumably, the priors should be jointly specified. Priors specifying high values of  $Q$ ,  $\tau$  and  $(w_1, w_0)$  in the neighborhood of the origin would be postulated. However, this joint positive dependence is not strictly required provided the consistency conditions on the weights are satisfied.

Table 11 presents the joint distribution of  $(E_1, E_0)$  for adult women for selected values of Yule's  $Q$ . This table gives results for the first step of our three step procedure. It differs from Table 6 because earnings over the entire 18 month period are considered rather than employment in months 16, 17 or 18 as presented in the previous table. As  $Q$  increases, the probability of a favorable outcome from the program increases. For all values of  $Q$ , the program produces net employment gains (i.e.  $P(Y_1 > 0, Y_0 = 0) > P(Y_1 = 0, Y_0 > 0)$ ).

Table 12 presents the empirical results from stages 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 value of  $Y_1$ ; for  $Y_1 = 0$ , place as much mass as possible at extreme upper values of  $Y_0$ ).

$w_0 = 1$  (independence; denoted  $w_0$  in the table)

$w \propto a + bq$       $a = 1, b = 3$ .

$w_+$  = point mass placed at same extremes (i.e. for  $Y_0 = 0$ , place as much mass as

feasible near  $Y_1 = 0$ ; for  $Y_1 = 0$  place as much mass as feasible near  $Y_0 = 0$ ).

These weighting functions are depicted in Figure 11.

Reading down the third and fourth columns of Table 12, the mean values of  $Y_0$  and  $Y_1$  allocated to the  $(0, Y_0)$  and  $(Y_1, 0)$  cells decline as they must. 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  $Y_0$  must rise as a consequence of 5(a) and 5(b). (See columns six and seven of these tables). As  $w_1(y_1 | y_1 > 0)$  decreases for higher values of  $y_1$ , the mass of  $f(y_1 | y_1 > 0, y_0 > 0)$  necessarily increases in the upper tail. Similar remarks apply to the behavior of  $f_0(y_0 | y_0 > 0)$  as high end values are downweighted.

There are several interesting features of the estimated gain distribution. First, for virtually the entire range of dependence parameters, the median impact is positive. The median impact is never greater than \$1100 for the full eighteen month period. Second, unless very high positive dependence is specified for the continuous outcome measures of dependence, (the  $\tau$  parameter), the interquartile range on the gain distribution is very large. Third, for most configurations of the dependence parameters, more persons benefit from participation than non-participation. This number is obtained by multiplying the number in the final column of the table (which presents the proportion of the persons with positive earnings in both states who benefit from participating in the program) by the proportion of persons in the cell  $Y_1 > 0, Y_0 > 0$  (obtained from Table 11) and adding the proportion of persons with earnings in the program state but no earnings in the no-program state ( $P(Y_1 > 0, Y_0 = 0)$ ). However, this is not universally true. Consider the case  $Q = 1$ ,  $w_+$  and  $\tau = .9$  given in Table 12. For this case 41.2% of the persons benefit from the program, 20% do not have their status changed, and

38.8% of the persons are harmed.

Summary statistics of the overall impact distribution are presented in Table 13. This distribution is formed by combining the types of three conditional distributions. The final two columns of the table reveal that for no combination of values of the dependence parameters are a majority of women harmed by participating in the program. Yet for some 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.

Confirming the impression of Table 12, the interquartile range is plausible only for high values of  $Q$  and  $\tau$  and for weighting functions  $w_p$  and  $w_+$ . The median gain ranges from -455 to 714. For virutally all configurations with positive dependence parameters, the median is positive.

*(11) Deconvolution When Gains Are Not Anticipated At The  
Time Program Participation Decisions Are Made*

Another type of dependence restriction postulates that the gain,  $\Delta$ , is independent of the base  $Y_0$  so

$$Y_0 \perp\!\!\!\perp \Delta \mid d = 1.$$

Then

$$Y_1 = Y_0 + R\Delta, \quad R\Delta \perp\!\!\!\perp Y_0,$$

where  $R = 1$  if a person is randomized into the experiment and  $R = 0$  otherwise, and where the conditioning on  $d = 1$  is left implicit.

This condition would be satisfied if the gain  $\Delta$  can not be forecast at the time decisions

are made about program participation. This case is extensively discussed in Heckman and Robb (1985, p.181) and is intermediate between the common-effect model and the variable - impact model when the impact is anticipated by agents.

We may write the density of  $Y_1$  as a convolution of  $Y_0$  and  $\Delta$ :

$$f_1(y_1 | R = 1, d = 1) = f_{\Delta}(\Delta | R = 1, d = 1) * f_0(y_0 | R = 0, d = 1)$$

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  $\Delta$ , the characteristic function of  $Y_1$

$$E(e^{iy_1} | d=1) = \int_{-\infty}^{\infty} e^{iy_1} dF(y_1 | d=1)$$

may be written as

$$E(e^{iy_1} | d=1) = E(e^{i\Delta} | d=1)E(e^{iy_0} | d=1)$$

so

$$E(e^{i\Delta} | d=1) = E(e^{iy_1} | d=1) / E(e^{iy_0} | d=1) = \varphi(t) .$$

Then<sup>6</sup>

(See, e.g. Kendall and Stuart, Vol. I, 1977, p. 98). We can therefore recover the distribution of  $\Delta$  from the distributions of  $Y_1$  and  $Y_0$  produced by the experiment.

---

<sup>6</sup>The ratio of two characteristic functions need not be a characteristic function. By Bochner's Theorem (see, e.g. Gnedenko (1974)), for  $\varphi(t)$  to be a characteristic function, it must satisfy  $\varphi(0) = 1$  and  $\varphi(t)$  must be positive definite. This hypothesis could be tested using the methods presented in Heckman, Robb and Walker (1989). The test would consist of checking if the ratio of the two sample characteristic functions is "within sampling variation" of being positive definite.

$$(6) \quad F(\Delta | d=1) = \frac{1}{2} + \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{[e^{i\Delta} \varphi(-t) - e^{-i\Delta} \varphi(t)]}{it} dt.$$

Deconvolution is known to be a numerically treacherous operation (Jansson (1984)). Therefore before undertaking a deconvolution analysis, it is of some interest to examine a linear model approach to estimating the variance of  $\Delta$  using random coefficient models. If the variance is negligible, there is no point in undertaking more elaborate deconvolution operations. Our discussion of this approach helps to establish a link between our analysis and previous work by Heckman and Robb (1985).

*(a) A Random Coefficient Approach*

Setting  $Y_0 = X\beta + U$  as in the introduction to the paper, we obtain a conventional random coefficient model

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

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

$$Y = X\beta + R\bar{\Delta} + \varepsilon R + U \quad E(\varepsilon R + U | X, R) = 0.$$

The assumed independence between  $\Delta$  and  $Y_0$  translates into independence between  $\varepsilon$  and  $U$ . The increased variance in the residuals for participants can be used to estimate  $\text{VAR}(\varepsilon)$ . From participant residuals, we can identify

$$\text{VAR}(\varepsilon + U) = \text{VAR}(\varepsilon) + \text{VAR}(U).$$

From non-participant residuals, we can identify

$$\text{VAR}(U).$$

Thus we can test an implication of the assumption that  $\Delta \perp U$  by using the empirical analogs

of  $\text{VAR}(\varepsilon + U)$  and  $\text{VAR}(U)$  for participants and non-participants respectively.<sup>7</sup>

Table 14 presents estimates based on this approach. There is mild evidence in support of the hypothesis that  $\text{VAR}(\Delta) > 0$ , suggesting that a more elaborate deconvolution approach to estimating the distribution of  $\Delta$  is likely to be fruitful. If we maintain normality of  $Y_1$  and  $Y_0$  (given  $d = 1$  and  $X$ ), the distribution of  $\Delta$  is then normal with mean  $\hat{\Delta}$  and variance  $\hat{\Delta}$ . The line in Figure 12 labelled "Normal CDF" plots the graph of the cumulative distribution of the gain from treatment for participants obtained from the conventional variance components approach based on normality of the residuals.

*(b) Empirical Deconvolution*

A more general and robust approach exploits formula (4) and the empirical characteristic functions for  $Y_1$  and  $Y_0$  to estimate the distribution of  $\Delta$ . Details of the implementation of the deconvolution procedure are given in Appendix D. Figure 12 plots the estimated distribution function. It is clearly non-normal. Table 15 presents parameters calculated from this distribution. The evidence suggests that under this information assumption about 18% of adult women were

---

<sup>7</sup>We note parenthetically that the random coefficient model with gain unknown at the time program enrollment decisions are made is a halfway house between the ordinary variable coefficient model and the common effect model. In an ordinary non-experimental setting, we write

$$Y = dY_1 + (1 - d)Y_0$$

so

$$Y = X\beta + d\Delta + U$$

and

$$Y = X\beta + d\hat{\Delta} + [U + d\varepsilon].$$

By assumption,  $E(d | \varepsilon, X) = 0$  so the only form of selection rises if  $E(U | d, X) \neq 0$ . Thus standard instrumental variable estimators can be used to consistently estimate  $\hat{\Delta}$  just as in the common effect model. See Heckman and Robb (1985) for further details. Assuming that  $U$  is homoscedastic, the variance of  $\varepsilon$  can be estimated using the residuals for those with  $d = 1$  and for those with  $d = 0$  exploiting the fact that  $E(Ud\varepsilon) = 0$ . Thus  $\hat{\text{VAR}}(U + \varepsilon d | d = 1) - \hat{\text{VAR}}(U + \varepsilon d | d = 0)$  consistently estimates  $\text{VAR}(\varepsilon)$ .

harmed by participating in the program. There is substantial mass (43%) at the origin with  $\Delta = 0$ . The density conditional on  $\Delta > 0$  is presented in Figure 13. The estimated variance of the nonparametric gain distribution matches the variance for the gain distribution obtained from the random coefficient model within the range of sampling error produced from the two estimates.

### *(12) Discussion and Summary*

This paper considers the uncertainty about the joint distribution of treatment and non-treatment outcomes that is an inherent feature of data produced from randomized social experiments. In the special - but widely invoked - case in which the treatment and control outcome distributions differ only by a constant or a deterministic function, social experiments recover the joint distribution of outcomes. In the more general case they do not. Classical probability bounds widely used in the literature on clinical trials restrict features of the outcome distributions but still leave considerable variability in the earnings data for the adult females from the National JTPA Study that we analyze in this paper.

In the context of labor market outcomes, there is more information about the relationship between outcomes in the treatment and control populations than is used in the classical probability bounding literature with the conventional case here. With the conventional approach to program evaluation assumes that treatment and control outcomes differ by a constant or a deterministic function as our point of departure, we consider a somewhat more general case in which the outcomes for the treatment and control state are "closely related", and we define several precise meanings of the term "related" that apply to models with discrete, continuous and mixed discrete/continuous outcomes.

Our measure of dependence for continuous outcomes measures proximity in terms of the distance in the ranks of outcomes in the treatment and control distributions. Our measure for discrete outcomes uses the cross product ratio defined for  $2 \times 2$  tables. A combination of the continuous measure and the discrete measure coupled with a more conventional measure of stochastic ordering for univariate distributions is required to analyze models with both discrete and continuous outcome measures. We demonstrate that restricting the range of admissible dependence reduces the intrinsic uncertainty produced from randomized social experiments. We also present Bayesian estimators for features of models that cannot be exactly identified in data from social experiments. These estimators also reduce the variability in the estimates of interesting features of the joint outcome distribution. However, to produce credible estimates of the distribution of program gains at selected quantiles requires that the dependence between treatment and control outcomes be high and positive. Positive dependence is an outcome of certain optimizing models of program choice.

We also use deconvolution methods for estimating the distribution of program gains when agents do not know the gain from a program at the time program participation decisions are made. We compare estimates from these models with estimates obtained from more conventional random coefficient models.

Application of the methods developed in this paper to data on adult female earnings and employment from the National JTPA experiment produces a range of distributions of program impacts all of which indicate small but positive effects of the program. Median earnings gains are positive for most distributions. Most people are not harmed by the program and for all values of the dependence parameters we consider, employment gains are positive. The most

plausible earnings impact distributions are produced from models where the earnings in the treated and untreated states are strongly positively dependent.

We have not exploited information about the choice process that can be used to extrapolate experimental evidence to other environments where the ingredients of the decision rule are different to recover the joint distributions  $F(y_1, y_0)$  or  $F(y_1, y_0 | d = 0)$  rather than just  $F(y_1, y_0 | d = 1)$ . The full joint distribution of outcomes can be recovered for participants and non-participants ( $d = 0$ ) in the Roy Model (Heckman and Honoré (1990)) or in the dependent competing risks model (Heckman and Honoré (1989)). Both of these models rely on the assumption that agents participate in programs in order to maximize their incomes. It is possible to recover the full joint distribution for all persons ( $d = 0$  and  $d = 1$ ) using only the data on outcomes for controls and experimentals if there is sufficient variation in explanatory variables. It is also possible to estimate the impact on these distributions of changes in regressors resulting from policy changes that alter the package of conditioning variables confronting individuals. The existing published literature demonstrates how this can be done and we do not repeat its findings here. Viverberg (1993) presents bounds for the unidentified parameters of parametric versions of the two sector model of self-selection with general self-selection rules.

Without prior information, social experiments answer only a few of the questions of interest to program evaluators. With the methods presented in this paper and in the published literature, supplemented by a variety of forms of prior information, data from social experiments can be used to answer a much wider variety of interesting policy questions.