WHEN IS ATE ENOUGH? RISK AVERSION AND INEQUALITY AVERSION IN EVALUATING TRAINING PROGRAMS [*]

Rajeev Dehejia
Columbia University and NBER
rd247@columbia.edu

This paper explores the relationship between the theory and practice of program evaluation as it relates to training programs. In practice programs are evaluated by mean-variance comparisons of the empirical distributions of the outcome of interest for the treatment and control programs. Typically, earnings are compared through the average treatment effect (ATE) and its standard error. In theory, programs should be evaluated as decision problems using social welfare functions and posterior predictive distributions for outcomes of interest. This paper considers three issues. First, under what conditions do the two approaches coincide? I.e., when should a program be evaluated based purely on the average treatment effect and its standard error? Second, under more restrictive parametric and functional form assumptions, the paper develops intuitive mean-variance tests for program evaluation that are consistent with the underlying decision problem. Third, these concepts are applied to the GAIN and JTPA data sets.

First version: 4 April 2000
Current version: 27 March 2003

[*] The author gratefully acknowledges conversations and collaboration with Joshua Angrist that sparked and helped to refine this research and comments from Alberto Abadie, Gary Chamberlain, Richard Ericson, Andrew Gelman, Jinyong Hahn, Gur Huberman, Charles Jones, David Kranz, Adriana Lleras-Muney, and Dale Poirier. Thanks to seminar participants at Brown, Columbia, EUI Florence, UC Irvine, and University of Wisconsin, Madison.

When is ATE enough? Rules of Thumb vs. Decision Analysis in Evaluating Training Programs

1. Introduction

Program evaluation is typically carried out by considering the average treatment effect (ATE) of a new program under consideration (called the treatment) relative to a status quo program (called the control). This is true both in experimental settings where the ATE can be estimated by a simple difference in means for outcomes of interest between the treatment and control groups, and also in non-experimental settings where the ATE is often a parameter in a much more complicated model. Uncertainty regarding the treatment impact is summarized by the standard error of the ATE, often through the statistical significance of the point estimate.

In contrast, decision theory offers a more comprehensive method for evaluating programs. A decision-theoretic analysis leads to a choice that maximizes (minimizes) expected utility (loss) given a likelihood model of the data. This result was first established by Wald (1950) and has led many researchers to formalize and extend the decision-theoretic framework. Of course, one might argue that this formalized statement of objectives misses intangible features of the decision problem. It is clear, however, that the decision-theoretic framework, while adding a layer of complexity to simple Neyman-Fisher hypothesis-testing, provides a solid foundation for inference that explicitly links empirical estimates with the broader framework of rational, economic decision-making.

At a more practical level, a decision theory approach is more general along two dimensions than simply looking at ATE. First, it accounts for uncertainty regarding the treatment impact in a systematic way, through the decision-maker’s risk attitude in an expected utility setting. To the extent that the Neumann-Morgenstern (1944) / Wald (1950) approach is widely accepted in economics, the decision theoretic approach reflects how we should account for uncertainty. Second, the decision approach allows for the decision-maker to exhibit inequality aversion, which would lead him or her to consider the treatment impact on features of the distribution other than the average.

The aim of this paper is to develop rules of thumb for evaluating programs which are consistent with the decision framework. Interpreted literally – as adopting a program when its treatment effect is positive and statistically significant, we show that the traditional approach to evaluating programs is valid only under very strong assumptions. We then relax these assumptions to develop simple techniques – rules of thumb – for evaluating programs that are valid under more general conditions.

The paper proceeds as follows. In Section 2 we set up the general framework for evaluation. In Section 3 we establish conditions for equivalence between the traditional approach and the decision approach. In Section 4 we develop several rules of thumb for evaluating programs. In Section 5 we apply these rules in evaluating the GAIN and JTPA data sets. Section 6 concludes.

2. A Framework for Evaluation

2.1 Decision Theory and Evaluation

In a world without uncertainty, the policy-maker adopts a social welfare function, $S(u_1(y_1), u_2(y_2), \ldots, u_N(y_N))$ which, for the population of interest, n=1,…, N, aggregates individual utility over the outcome of interest, y_n. In general we think of y_n as being the net outcome of interest. For example, in an evaluation context, y_n could be post-program earnings or post-program earnings net of costs. Note that this is a "welfarist" social welfare function (see Sen [1973], and Heckman and Smith [1998]), because it is defined over individual utility, $u_n(\cdot)$, from the outcome of interest.

In the context of program evaluation, there are two (or more) outcomes of interest for each individual, $y_i^{(t_i)}$, with t_i=1 representing the treatment program and t_i=0 representing the control program. The utility-relevant outcomes are summarized by $W = \{y_i^{(0)}, y_i^{(1)}\}_{i=1}^n$. The observed data are given by $Z = \{y_i, t_i\}_{i=1}^n$, where $y_i = t_i y_i^1 + (1-t_i) y_i^0$. There is a distribution for Z and W:

$$(Z,W) \sim Q$$

where

$$Q = \int_{q\in\Theta} P_q \, dp(q) \, .$$

The outcome that will be realized depends on a decision by the policy-maker, a=0,1:

$$h(W,a) = \{y_i^{(a)}\}_{i=1}^n \, .$$

Ex post expected social welfare then is given by:

$$SWF(a) = E_Q\left\{U\left[S\left(\left\{y_i^{(a)}\right\}_{i=1}^n\right)\right]\right\},$$

where a=0,1 and $U(\cdot)$ is a Neumann-Morgenstern utility function and S is a social welfare function. A simplification which we will often adopt is to assume that $u_n(y_n)=y_n$, so that the policy-maker is concerned only with outcomes at the individual level. The expectation is with respect to the predictive distribution, $\left\{y_i^{(a)}\right\}_{i=1}^{n}\Big| Z$.

An alternative formulation considers ex ante uncertainty. Uncertainty is accounted for at the individual level. Each individual faces uncertainty regarding the outcome of interest, represented by $Y_n^{(a)}$, a distribution over $y_n^{(a)}$, and realizes some expected utility $u_n(Y_n^{(a)})$. The policy-maker's decision is based on a social welfare function defined over these individual expected utilities, $S(u_1(Y_1^{(a)}), u_2(Y_2^{(a)}),\ldots, u_n(Y_n)^{(a)})$ or their certainty equivalents:

$$S(y_{1*}^{(a)}, y_{2*}^{(a)}, \ldots, y_{n*}^{(a)})\,, \tag{2}$$

where $y_{1*}^{(a)}$ is the certainty equivalent of individual i's earnings (see Dehejia [1999]).

Both formulations are useful, depending on the context. We adopt the first formulation when focusing on the role that risk attitude plays in the evaluation, since it allows us to assume inequality neutrality while remaining agnostic about risk attitude. Instead, we adopt the latter approach when considering the role of inequality aversion, since this is more transparent in the latter approach.

2.2 Decision Theory in Practice

Three ingredients must be specified in order to implement this approach. First, given the observed data, an appropriate model must be specified for the outcome. For example, in an evaluation context, a model for earnings would be needed. If the data are drawn from a randomized trial, then the model would be straightforward; if there are issues of sample selection, then a more complicated model would be needed. From the model, we either approximate (using maximum likelihood) or simulate (using Markov Chain Monte Carlo) the posterior distribution of the parameters, $p(q|Z)$. Based on the posterior distribution of the parameters, we are interested in the predictive distribution of individual earnings, $\left\{y_i^{(a)}\right\}_{i=1}^{n} \Big| Z$. This is the distribution in the space of the outcome that accounts for all parameter and intrinsic uncertainty, conditional on model choice, and is readily simulated. Second, we must specify the policy-maker's risk attitude, and third, must specify the policy-maker's inequality attitude.

Two simplifications that are often adopted are risk neutrality and inequality neutrality. Inequality neutrality implies that the policy-maker is concerned only with the average outcome over the population of interest. This is stated in Observation 1.

> **Observation 1**: When $S(\cdot)$ is utilitarian (linear) and $u_n(y_n)=y_n$, the policy decision turns on comparing $U(\bar{y}_0)$ with $U(\bar{y}_1)$.

This implies that the policy-maker would examine the predictive distribution of $\bar{y}$ for both programs under consideration. The assumption of risk neutrality implies that the policy-maker is concerned only with the mean predicted value of social welfare. The two assumptions together have the strong implication that the policy-maker will compare the mean of the predictive distribution of average earnings across the two programs. Under the additional assumption of normality of the outcome, the mean of the predictive distribution of earnings will simply correspond to the sample mean. This is stated in Observation 2.

Observation 2: Assume $S(\cdot)$ is inequality neutral, $U(\cdot)$ is risk neutral, $u_i(y_i)=y_i$, y_{i1}~$\mathrm{N}(m_1, s_1)$, and y_{i0}~$\mathrm{N}(m_0, s_0)$. Then the treatment program will be adopted if $\bar{y}_1 - \bar{y}_0 > 0$.

2.3 The Traditional Approach

Rather than focusing on the predictive distribution, the traditional approach to program evaluation instead examines the average treatment effect (ATE), and its standard error. One version of this is stated below as Rule of Thumb 0.

Rule of Thumb 0 (RT0): Adopt program 1 (the treatment) if

$$t = \frac{\bar{y}_1 - \bar{y}_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} > t_{a/2}\,, \qquad \text{(RT 0)}$$

i.e., if the treatment effect is positive and statistically significant, where $s_p^2 = \dfrac{(n_1 - 1)s_1 + (n_0 - 1)s_0}{n_1 + n_0 - 2}$ and n_i, $\bar{y}_i$, and s_i^2 are the sample size, mean, and standard deviation for group i=1,0.

Essentially the salience of the treatment effect is evaluated through its t-statistic. Though this approach might seem overly simplistic, it is the most widely used rule of thumb in evaluating programs (see for example Friedlander, Greenberg, and Robins [1997], Table 2).

3. Looking for Equivalence

In this section we discuss the assumptions within a decision framework that justify the use of the average treatment effect in evaluating programs.

3.1 There Is No General Equivalence

Proposition 1, below, establishes that there are no assumptions that in general justify RT0 as the outcome of an expected utility decision process.

> **Proposition 1**: Assume $S(\cdot)$ is inequality neutral, y_{i1}~N(m_1, s^2), and y_{i0}~N(m_0, s^2). RT0 is not valid under a decision framework for all n_1 and n_0.
> **Sketch Proof**: For simplicity, assume that the difference in program cost between the treatment and control programs is constant, c, so that y_{i1} and y_{i0} are the outcomes. Consider three cases.
> (A) Positive program cost, c, and symmetric priors for programs 1 and 0: Consider $n_1=n_0=n\rightarrow\infty$. As n grows large, numerator of RT0 approaches $m_1 - m_0$, and the denominator of RT0 shrinks . For a sufficiently large n, any positive mean difference will be statistically significant. But then program 1 will be chosen if there is a small mean difference with program 0, even though the difference is less than c. This is inconsistent with $U(\cdot)$ increasing in earnings net of program costs.
> (B) Zero cost and symmetric priors: Consider the case with $s_1 = s_0$ and $n_1=n_0$. If utility is increasing in y, $\bar{y}_1 > \bar{y}_0$ is sufficient to choose program 1, which is not satisfied by RT0.
> (C) Zero cost and asymmetric prior: This case includes situations where there is a prior preference for one of the programs or extra information about one of the programs. However, even in this case, for n_1 and n_0 sufficiently large, the data will dominate the prior and we return to case (B).

3.2 A Special Case

Thus, the justification, if any, for the standard decision procedure will rely on fixed sample size arguments. One argument justifying RT0 for a fixed sample size is offered in Proposition 2.

> **Proposition 2**: RT0 can be justified by state-dependent preferences which are defined on the decision space rather than the outcome space.

Proof: Consider the following table:

		Choice made (c)	
		Treatment (1)	Control (0)
State of the world (s)	$m_1>m_0$ (1)	a_{11}	a_{10}
	$m_1<m_0$ (0)	a_{01}	a_{00}

The elements of the table would normally correspond to the utility from the choice made, given that state of the world. The decision rule is given by:

$$\text{EU}(c=1\,|\text{data}) = a_{11}\cdot\Pr(m_1>m_0\,|\text{data}) + a_{01}\cdot\Pr(m_1<m_0\,|\text{data}),$$

$$\text{EU}(\text{c}=0\,|\text{data}) = a_{10}\cdot\Pr(m_1>m_0\,|\text{data}) + a_{00}\cdot\Pr(m_1<m_0\,|\text{data}),$$

and

$$\text{EU}(c=1\,|\text{data}) - \text{EU}(c=0\,|\text{data}) = a_{01} - a_{00} + (a_{11} - a_{10} - (a_{01}-a_{00}))\cdot\Pr(m_1>m_0\,|\text{data}).$$

Thus c=1 if

$$\Pr(m_1 > m_0|\text{data}) > \frac{a_{00} - a_{01}}{a_{11} - a_{01} - (a_{10} - a_{00})} \equiv k\,. \tag{3}$$

For fixed n_1 and n_0 and appropriate values of a_{01}, a_{11}, a_{10}, and a_{00}, $k=t_{a/2}$.

There are two limitations to this result. First, as mentioned above, it relies on a fixed sample size. Second, the preferences adopted in Proposition 2 are state dependent and defined over choices rather than outcomes. This is a problematic assumption, because the policy-maker’s preferences depend on a state ($m_1>m_0$ versus $m_1<m_0$) which is unobserved both before and after the decision is made.[1] . Normally state-dependent preferences rely

[1] For example, in choosing between a picnic and an afternoon at the museum, preferences may be influenced by the weather, which is unknown when choosing, but during the afternoon the weather becomes known and of course affects the enjoyment of the activity. In an evaluation context the state of the world, $m_1>m_0$ or $m_1<m_0$, is unknown both when making the decision and also after the decision has been taken. The policy-maker’s utility from the realization of each program cannot reasonably be state dependent, because the underlying states are never observed by the policy-maker. In the previous example, it would be like having utility from a picnic and an afternoon at the museum depend on the weather on Mars.

on a state of the world that is unobserved when a decision is being made, but which is observed when the outcome is experienced.[2]

In conclusion, the rule of thumb that is widely adopted to evaluate programs is not in general valid. In next section we examine rules of thumb that are derived directly from a Bayesian decision procedure.

4. Relaxing Risk and Inequality Neutrality

In this section we develop rules of thumb that unlike RT0 are valid within the standard decision framework. In Section 4.1, we develop these rules by making specific assumptions about the likelihood model for earnings (normal, log normal, a mixture distribution) while maintaining the assumption of inequality neutrality. In Section 4.2, while still assuming inequality neutrality, we relax the parametric assumptions by using a multinomial likelihood. In Section 4.3 we discuss the inclusion of covariates, using a maximum likelihood framework. Finally in Section 4.4 we allow for inequality aversion.

4.1 Relaxing Risk Neutrality

A policy-maker who is inequality neutral, as we have seen in Observation 1, will consider $\bar{y}$, the average earnings in each program, but will also have to account appropriately for uncertainty. Table 0, rows 1 to 7, summarize rules of thumb that allow for risk aversion, but maintain the assumption of inequality neutrality, under a range of parametric assumptions on the earnings process.

[2] To adopt state-dependent preferences we would have to assume that the policy-maker eventually knows the true state of the world. This could be justified if the program will be evaluated multiple times, so over many trials the true parameter values will become known. However, in reality after a few evaluations, either the treatment or control program is chosen, so either m_1 or m_0 will become known, but not both.

Table 0: Rules of Thumb for Evaluating Programs

	Assumption on $U(\cdot)$	Assumption on $S(\cdot)$	Distribution	Solution Concept*	Adopt treatment if:
1	--	linear	normal	SD1	$m_1 > m_0$ and $s_1^2 < s_0^2$ [a]
2	--	linear	log normal	SD2	$s_1^2 < s_0^2$ and $m_1 + s_1^2/2 > m_0^2 + s_0^2/2$ [b]
3	--	linear	mixture**	SD1	$p_0 > p_1$, $m_0 < m_1$, and $s_0 > s_1$ [c]
4	--	linear	mixture	SD1	$p_0 - p_1 > \Pr(y_1 \le c \mid m_1, s_1)$ $- \Pr(y_0 \le c \mid m_2, s_2)$, $m_0 < m_1$, and $s_0 < s_1$ [d]
5	--	quadratic	--	EU	$b\mathrm{m}_1 + c\mathrm{v}_1 + c\mathrm{m}_1^2 > b\mathrm{m}_0 + c\mathrm{v}_0 + c\mathrm{m}_0^2$ [e]

Notes:
* SDn = n-th order stochastic dominance, EU=expected utility.
** Mixture=Pr(y_{it}=0)=p_t, Pr(y_{it}>0)=1–p_t, t=1,0. For y_i~log N | y_i>0.

[a] $m_i = \bar{y}_i, s_i^2 = \frac{1}{n_i}(\frac{1}{n_i}+1)s_i^2$, i=1 for treatment group, =0 for control group.

[b] $m_i = \bar{y}_i, s_i^2 = \frac{1}{n_i}(\frac{1}{n_i}+1)s_i^2$, sample statistics from log earnings data; i=1 for treatment group, =0 for control group.

[c] p_i=proportion of zeros in earnings data, $m_i = \bar{y}_i, s_i^2 = \frac{1}{n_i}(\frac{1}{n_i}+1)s_i^2)$, sample statistics from log of positive earnings data; i=1 for treatment group, =0 for control group.

[d] sample statistics as in (c) with

$$c = \frac{\dfrac{m_0}{s_0^2} - \dfrac{m_1}{s_1^2} - \left(\left(\dfrac{m_0}{s_0^2} - \dfrac{m_1}{s_1^2}\right)^2 - 4\left(\dfrac{1}{2s_0^2} - \dfrac{1}{2s_1^2}\right)\left(\dfrac{m_0^2}{s_0^2} - \dfrac{m_1^2}{s_1^2} + \log s_0 - \log s_1\right)\right)^{1/2}}{\dfrac{1}{2s_0^2} - \dfrac{1}{2s_1^2}}.$$

[e] $S(y)=a + by + cy^2$, c_n=certainty equivalent of individual predictive distributions of earnings,

m_i=mean(c_n), v_i=variance(c_n), i=1 for treatment group, =0 for control group.

The simplest assumption is that y_{it}~N(m_t, s_t), for t=1 (treatment) and 0 (control), in which case the predictive distribution of average earnings under treatment (control) is given by,

$$\bar{y}_t \middle| Data \sim N(\bar{y}_{T=t}, \tfrac{1}{n}(\tfrac{1}{n_t}+1)s_t^2) \equiv N(m_t, s\,_t^{\,2}) , \tag{4}$$

where $\bar{y}_{T=t}$ and s_t^2 represent the sample mean and variance of the empirical distribution and n_t is sample size, for t=1,0. Note that $\bar{y}_t$ is the average *predicted* earnings for the sample of interest, whereas $\bar{y}_{T=t}$ is the sample mean of earnings for the treatment (control) group.[3,4]

We derive a decision rule by using the concept of stochastic dominance. This is a useful concept because it allows us to compare distributions without making specific assumptions on the utility function. The limitation of the concept is that not all distributions can be ranked according to stochastic dominance. Row 1, Table 0, states the conditions under which the treatment first-order stochastically dominates the control. The conclusion is that -- if all we know about the decision-maker is that his utility function is increasing in the average earnings from the program -- the programs can be ranked unambiguously only if one of the programs has both a higher mean and a lower variance.

A limitation of the rule described in row 1 is the assumption that the outcome is normally distributed, which is unrealistic for earnings. A more common assumption is that earnings are log normally distributed. Equation (1) can be reinterpreted as describing the predictive distribution of average log earnings (where the sample moments are also for the log distribution). The policy-maker compares the exponential of this distribution

[3] In principle, (4) can also be interpreted as the predictive distribution from a classical regression model. This would allow us to relax the assumption of no sample selection bias.

[4] The assumption of normality is common, indeed one of the foundations, of mean-variance comparisons in the portfolio choice literature (see Ingersol 1987 and Samuelson 1970). The normality assumption can be justified asymptotically, depending on the model. If the variable of interest is defined as a parameter in a likelihood model (e.g., the mean of the earnings distribution), then we know that it will be asymptotically normal (see inter alia De Groot 1970 and Le Cam 1953).

for each program.[5] From row 2, we see that if a distribution is to be preferred by a decision-maker under all risk-averse preferences to another, then it must have a lower variance. However, its mean can be lower, because of the second condition. Thus, a mean-variance tradeoff exists to the extent that if the first distribution has a lower variance, its mean must be higher to a sufficient degree to obtain second-order stochastic dominance. Thus this rule is more demanding than the previous one.

Log normality can also be a problematic assumption under some circumstances. In particular, in many labor training programs a large proportion of observations is concentrated at zero (representing zero earnings, which is common among welfare recipients). Rows 3 and 4 generalize the likelihood to a mixture distribution, in which there is a probability *p* of zero (and a probability 1–*p* of positive) earnings. Conditional on earnings being positive, they are described by a log normal distribution. Row 3 presents the obvious generalization of the rule in row 2: it requires the mass point in the control distribution to be larger, along with the conditions described in row 2. Row 4 considers an important extension. Even if the conditions from row 2 are not satisfied, if the mass point in the control distribution is sufficiently larger than the mass point in the treatment distribution, the treatment can still first-order stochastically dominate the control.

[5] Note that this is not precisely correct, since the exponential is not a linear operator. The policy-maker should compare the average of the exponential of the individual predictive distributions of earnings. Unfortunately this distribution will not have a convenient parametric form, which is why we consider this alternative.

4.2 Non-Normal Data: A Nonparametric Approach

In this section we consider a non-parametric (or flexibly parametric) approach to modeling the data and ranking programs. In order to achieve this generalization, we assume that the data are discrete and use a multinomial distribution.

Divide the support of the earnings distribution into M bins, $b_1,b_2,\ldots, b_M$. For each bin, we observe $n_{11},n_{21},\ldots, n_{M1}$ individuals in the treatment group and $n_{10}, n_{20},\ldots, n_{M0}$ units in the control group. In this setting, a multinomial likelihood with probability of a draw from each bin $(q_{1t}, q_{2t}, \ldots,q_{Mt})$, t=1,0, is completely unrestrictive. The advantage of this setup is that it allows us to consider risk aversion under more general parametric assumptions. Since the model does not incorporate covariates, we still cannot consider the role of inequality aversion.

A conjugate prior for this model is a Dirichlet prior of the form $D(z_{1t}, z_{2t}, \ldots, z_{Mt})$, which leads to a posterior of the form $(q_{1t}, q_{2t}, \ldots,q_{Mt})| \, data \sim D(z_{1t}+n_{1t}, z_{2t}+n_{2t}, \ldots, z_{Mt}+n_{\mathrm{M}t})$, for t=1,0. Chamberlain and Imbens (1996) suggest that an improper prior where $n_{mt}\to 0$ is preferable. The predictive distribution of earnings is given by:

$$\begin{aligned}
&\Pr(Y_{it} = b_m | data) \\
&= \int \Pr(Y_{it} = b_m | q_{mt})\, p(q_{mt} | data) dq_{mt} \\
&= \int q_{mt}\, p(q_{mt} | data) dq_{mt} \\
&= \frac{n_{mt}}{\Sigma n_{mt}},
\end{aligned}$$

where the last step is from the properties of the Dirichlet distribution. Thus the predictive distribution of y_{it} is essentially the empirical distribution of the data.

Since the model does not allow for covariates, each individual has the same expected utility in each program,

$$EU_t = \sum_m p_{mt} U(b_m),$$

where N_t=Σn_{mt}, and $p_{mt} = n_{mt}/N_t$, for t=1,0, and the ranking of each program will be determined by comparing EU_1 with EU_0. The stochastic dominance concepts discussed in the previous section could also be applied in this context, but there are no additional restrictions, beyond the definition of stochastic dominance, under which a positive mean difference becomes a sufficient condition for stochastic dominance.[6]

4.3 Non-Normal Data: A Maximum Likelihood Approach

An alternative to the multinomial model discussed in the previous section is to use maximum likelihood methods. This would allow for the inclusion of covariates, and more generally allows for the estimation of relatively complicated models with ease. Once a likelihood model has been specified for the outcomes of interest, then the posterior distribution of the parameters can be approximated from the maximum likelihood estimate using the usual formula:

$$q|z \overset{approx}{\sim} N\left(\hat{q}_{ML}, \left[\frac{\partial^2 \ell(q_{ML})}{\partial q \partial q'}\right]^{-1}\right).$$

Then the predictive distribution of the outcome $p(y|z)=p(y|q)\mathrm{p}(q|z)$ is readily simulated by first drawing from the posterior distribution of the parameters, and then drawing from the outcome distribution specified in the likelihood. The predictive distributions of the

[6] A positive difference in means is a necessary, but not sufficient, condition for the treatment program to first- and second-order stochastically dominate the control. But in the case of discrete distributions, it is very easy to check for dominance. The treatment first-order stochastically dominates the control if and only if:

$$P_{m1} \leq P_{m0} \quad \forall m, \text{strict for some } m,$$

where $P_{mt} = \sum_{i=1}^{m} p_{it}$, t=1,0. The treatment second-order stochastically dominates the control program if:

outcomes for the individuals of interest can then be fed into the decision problem outlined in Section 2.

4.4 Relaxing Inequality Neutrality

The final extension we consider is to allow for inequality neutrality. We use the framework described in equation (2), Section 3, to allow for the possibility of inequality aversion. The policy-maker's social welfare for a program is a function of each individual's certainty equivalent of earnings under that program. The policy-maker chooses the program with the higher social welfare. At this level we can be agnostic about not only individual preferences, but also the likelihood model used to produce the predictive distributions of earnings.[7] All that is required is a certainty equivalent of the earnings distribution that each individual faces.

In order to obtain rules of thumb for ranking programs, we must make functional form assumptions about $S(\cdot)$. If $S(\cdot)$ is quadratic, then the programs can be ranked by the means and variances of the certainty equivalents of earnings for each individual in each program. This is shown in row 5 of Table 0.

5. Two Examples

5.1 The GAIN Demonstration, Riverside County

In this sub-section we re-analyze data from the Riverside county portion of the Greater Avenues for Independence (GAIN) experiment. Riverside was widely viewed as the most

$$\sum\nolimits_m (P_{mt} - P_{mt}) \leq 0 \quad \forall m, \text{strict for some } m.$$

[7] The likelihood must allow for heterogeneity in individual earnings. Without heterogeneity, the distribution of the certainty equivalents of earnings under treatment (control) across individuals would be

successful portion of the GAIN experiment, and more generally as one of the success stories of welfare reform. We demonstrate that the traditional approach of analyzing t-statistics provides a narrow and misleading view of this program.

GAIN was a welfare-to-work demonstration in which the randomly assigned treatment consisted of a set of labor training and job-search initiatives imposed on recipients of Aid to Families with Dependent Children. Table 1 summarizes outcome information for Riverside County.

Table 1 reveals that an analysis of Riverside based on the statistical significance of the treatment impact leads to an ambiguous ranking. From panel (a), we note that the overall treatment effect is positive and significant, but if we break out positive earnings and consider them separately, panels (b) and (c) reveal that the treatment impact, though positive, is no longer statistically significant. Note that there is a substantial difference in the proportion of zeros in the treatment and control groups.

degenerate. The models considered in Sections 4.1 and 4.2, for example, do not allow for treatment effect heterogeneity.

Table 1: GAIN, Riverside County

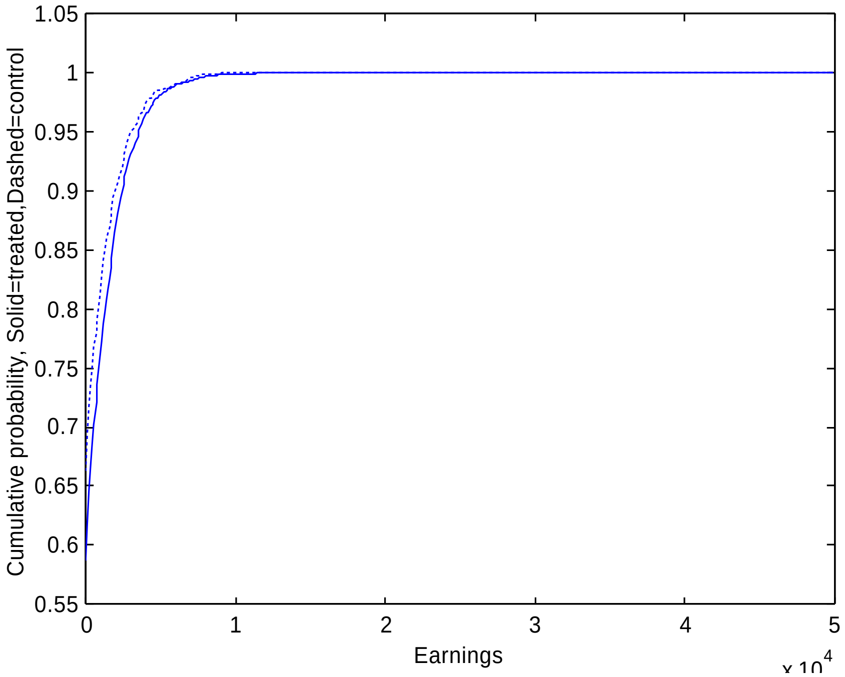
	Treatment		Control
(a) Average Quarterly Earnings			
Sample Mean	712		550
Sample Standard Deviation	1,351		1,199
Proportion of Zeros	0.5485		0.6249
Average Treatment Effect		163 (36)	
(b) Average Quarterly Positive Earnings			
Sample Mean	1,578		1,465
Sample Standard Deviation	1,637		1,579
Average Treatment Effect		112 (72)	
(c) Average Log Positive Earnings			
Sample Mean	6.6499		6.5209
Sample Standard Deviation	1.4870		1.5257
Average Treatment Effect		0.1290 (0.0660)	

The rules of thumb from Table 0 give a more consistent ranking of the program. Based on positive earnings, we note that treatment earnings have both a higher mean and higher variance than control earnings; thus, rules 1 to 3 do not apply. But when we consider the zeros, since the mass point at zero is greater for the control distribution by a sufficient degree than the mass point for the treated distribution, rule 4 implies first-order stochastic dominance. Thus, the ranking of programs is unambiguous.

Figure 1 confirms this conclusion in a non-parametric setting. Figure 1 plots the predictive distribution of earnings under treatment and control from the multinomial model. We see that the CDF of treatment earnings is everywhere to the left of the CDF of control earnings, implying first-order stochastic dominance.

The case of Riverside offers one illustration of the fact that t-statistics can be a very misleading means of accounting for the uncertainty in an evaluation.

Figure 1: Cumulative Distribution Function, Average Earnings, Riverside



Another dimension along which we can extend the analysis beyond the traditional approach is to consider the role played by inequality aversion in evaluating the Riverside experiment. Recall that a t-statistic implicitly assumes inequality neutrality. To allow for inequality aversion, we assume that individual preferences are constant relative risk aversion, and apply rule 5 from Table 0. Table 2 presents the means and variances of the certainty equivalents. The means for the treatment program are higher than for the control program, but the variance is also higher. The final column gives the restrictions on quadratic preferences for which the treatment is preferred. Since $c<0$, $b>>0$ is required.

Table 2: Certainty Equivalents, Riverside

Coefficient of relative risk aversion	Treatment		Control		Restrictions to prefer treatment
q	mean	variance	mean	variance	$b/\lvert c\rvert >$
2	68	11740	45	1865	0.0018
3	34	941	24	295	0.0082
4	23	283	16	107	0.0150
5	17	131	12	54	0.0217

Calibrating quadratic preferences is not an intuitive exercise. Table 3 instead uses CRRA preferences to illustrate how risk aversion and inequality aversion are combined into a ranking of the programs. The table depicts the ranking of the programs using CRRA preferences both for individuals and the policy-maker, with a range of coefficients of risk and inequality aversion. Note that for a range of values of risk and inequality aversion the treatment is not preferred to the control program. Interestingly this includes the case of $e=0$, inequality neutrality, which appears to contradict Figure 1 which indicates first-order stochastic dominance when $e=0$. This is accounted for by the fact that Table 3 is based on a model which allows for treatment effect heterogeneity, whereas Figure 1 assumes a constant treatment effect. This reinforces a point that has been made in the evaluation literature, namely the importance of allowing for treatment effect heterogeneity (see Dehejia 1999 and Heckman and Smith 1998). When we increase the degree of inequality aversion, the control program is preferred for a broad range of coefficients of risk aversion.

Table 3: Welfare Rankings, Allowing for Risk- and Inequality-Aversion, Riverside

		Coefficient of Inequality Aversion (*e*)									
		-4	-3	-2	-1	0	1	2	3	4	5
Coefficient of Relative Risk Aversion (*q*)	-4	+	+	–	–	–		–	–	–	–
	-3	+	+	–	–	–		–	–	–	–
	-2	+	+	–	–	–		–	–	–	–
	-1	+	+	+	–	–		–	–	–	–
	0	+	+	+	+	+		+	–	–	–
						(*)					
	2	+	+	+	+	+		+	+	–	–
	3	+	+	+	+	+		+	+	–	–
	4	+	+	+	+	+		+	+	–	–
	5	+	+	+	+	+		+	+	–	–

Notes: +=treatment preferred to control, –=control preferred to treatment.
(*) *q*=0.4, *e*=0 corresponds to RT0, which accepts the treatment.

There is no combination of the parameters *q* and *e* that directly correspond to RT0, the rule which would normally be followed. But a partial correspondence can be established as follows. RT0 embodies inequality neutrality, since it focuses on the average treatment effect, hence *e*=0. It does not also correspond to *q*=0 because if the policy-maker is risk-neutral as well then program 1 is chosen if $\bar{y}_1 > \bar{y}_0$. For RT0, $\bar{y}_1 - \bar{y}_0 > c$, where *c* is the denominator of RT0 multiplied by the right-hand side of the same expression. Given the distribution of $\bar{y}_0$, we can numerically compute *q* such that the treatment would be chosen if $\bar{y}_1 - \bar{y}_0 > c$. For the Riverside program, this corresponds to *q*=0.4 Thus RT0 corresponds to slightly risk averse preferences.

In summary, the traditional analysis of GAIN is misleading along two dimensions. First, the t-statistic on the ATE in positive earnings is not significant, though for total earnings the difference is significant. Instead, the rules of thumb deliver a more consistent ranking, which is confirmed by a non-parametric analysis. Second, the traditional approach ignores heterogeneity in the treatment impact.

5.2 The JPTA Data

In this section we provide an analysis of the JPTA data. We also illustrate how program costs can be incorporated into the analysis. The outcome which we study is the sum of 30-month earnings. Costs are measured according to the type of services each individual received (classroom training, on-the-job training, or other services). We present results for the adult male sample. Table 4 presents average earnings and training costs from the program.

Table 4: JTPA Data, Earnings (in dollars)

	Treatment	Control
(a) Average Quarterly Earnings		
Sample Mean	19,520	18,404
Sample Standard Deviation	19,912	18,760
Proportion of Zeros	0.1027	0.1039
Average Treatment Effect	1,117 (580)	
(b) Average Quarterly Positive Earnings		
Sample Mean	21,754	20,538
Sample Standard Deviation	1,983	1,868
Average Treatment Effect	1216 (610)	
(c) Average Log Positive Earnings		
Sample Mean	9.3084	9.2782
Sample Standard Deviation	1.5375	1.4645
Average Treatment Effect	0.0302 (0.0475)	

The treatment program has both a higher level and greater variation in treatment earnings compared to control earnings. The average cost per trainee is $515, whereas the average treatment effect is $1,117 (with a standard error of 580). The treatment effect is not statistically significant.

The rules of thumb from Table 0 are not able to rank the programs. Likewise, Figure 2, which depicts non-parametric estimates of the CDFs of treatment and control earnings, is inconclusive. Unlike the GAIN example, stochastic dominance does not help to rank the distributions, which cross twice. One important difference between the JPTA and GAIN data is that the former has a much smaller proportion of zeros (typically 0.1, compared to 0.6 for GAIN). The cost component of the program can be taken into account by netting the training costs out of earnings under treatment. The CDFs of net earnings are presented in Figure 3. Training costs do not alter the basic conclusion: stochastic dominance remains an inconclusive means of comparison. The inability of stochastic dominance arguments to rank the programs suggests that their ranking will depend on the degree of risk aversion of the policy-maker. We consider this issue below.

Figure 2: CDFs of Treatment and Control Earnings, JPTA

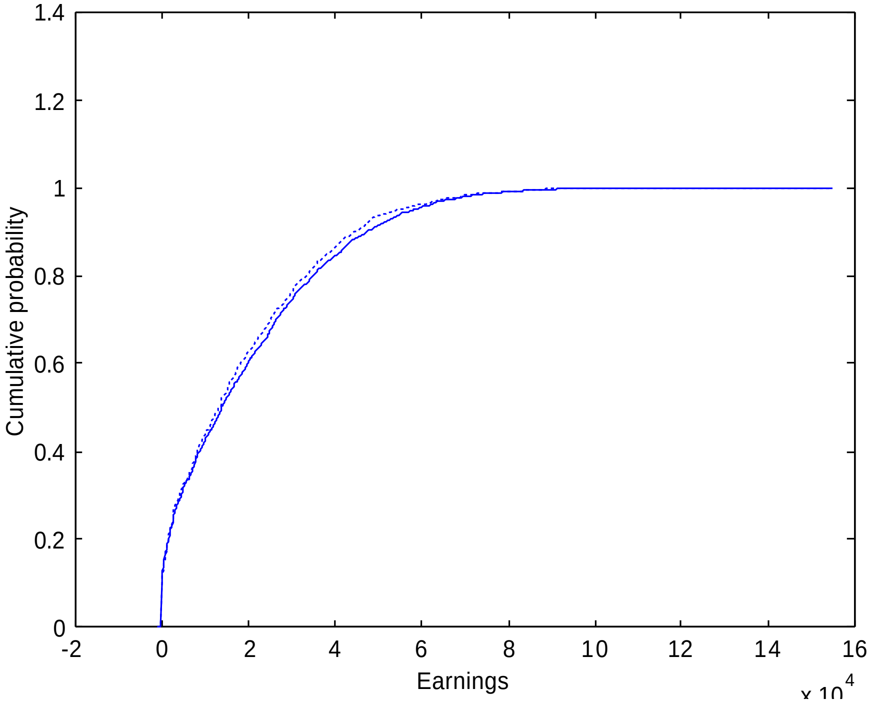
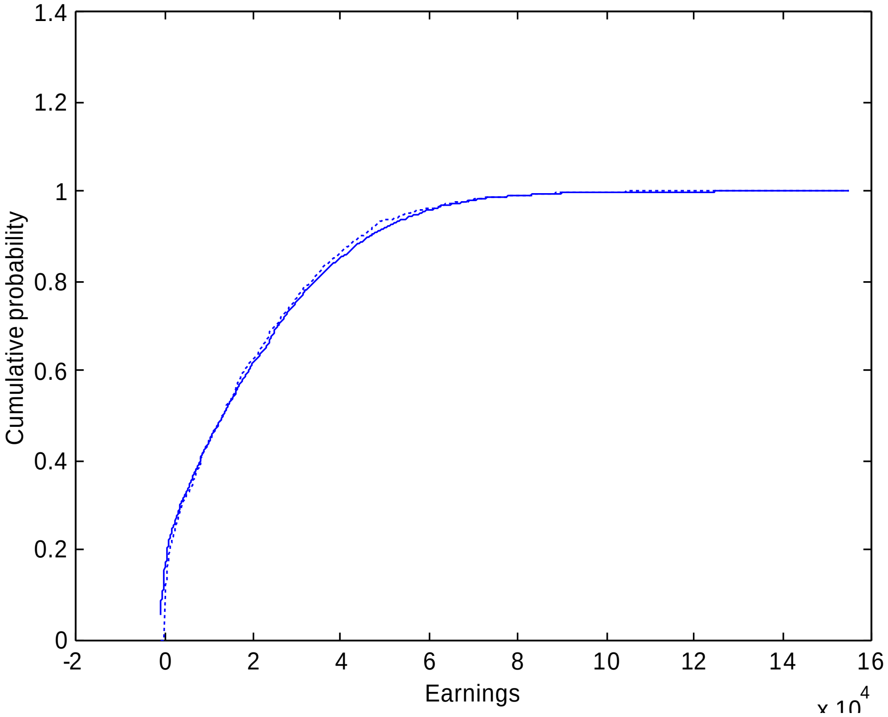


Figure 3: CDFs of Treatment and Control Earnings Net of Training Costs, JTPA



Before turning to specific functional forms for the social welfare functions, it is useful to consider the role that inequality aversion could play in ranking the programs. Figures 4 to 6 (below) present the percentiles of the earnings distribution of each program. The first figure considers earnings, ignoring training costs. We can see that in terms of percentiles of earnings the two programs start out together at $0 until about the 10th percentile. Then for the next range (roughly from the 10th to the 15th percentiles) the treatment distribution is higher. For the 15th to the 30 percentiles, control earnings are higher, and above the the 30th percentile treatment earnings are higher. In terms of inequality the treatment program raises earnings at both ends of the distribution.

When we add training costs into the analysis, the conclusions remain similar, except that the treatment distribution is pulled down by the training costs. This implies that the treatment distribution only overtakes the control distribution at the upper end of the distribution. This is depicted in Figures 5 and 6.

Figure 4: Percentiles of Earnings, JTPA

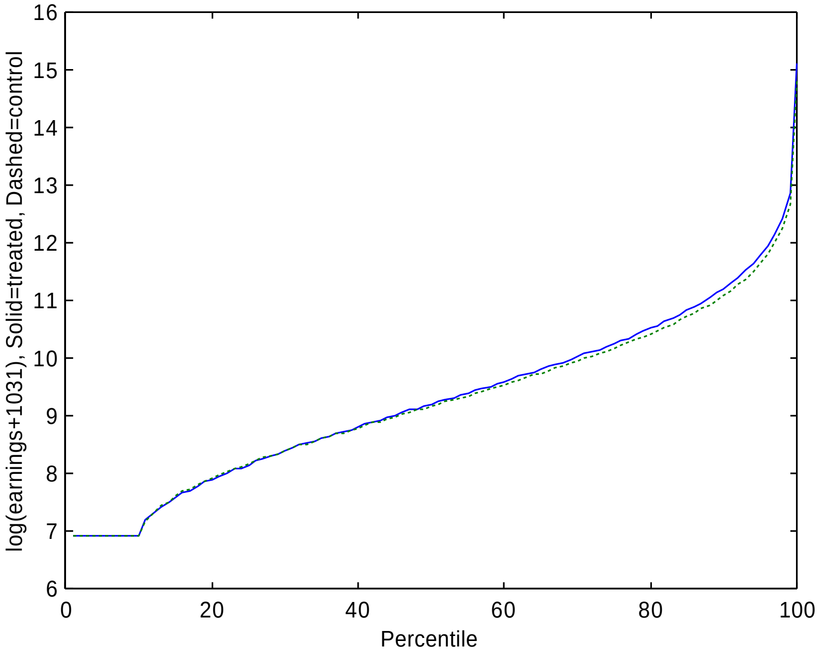


Figure 5: Percentiles of Earnings Net of Training Costs, JTPA

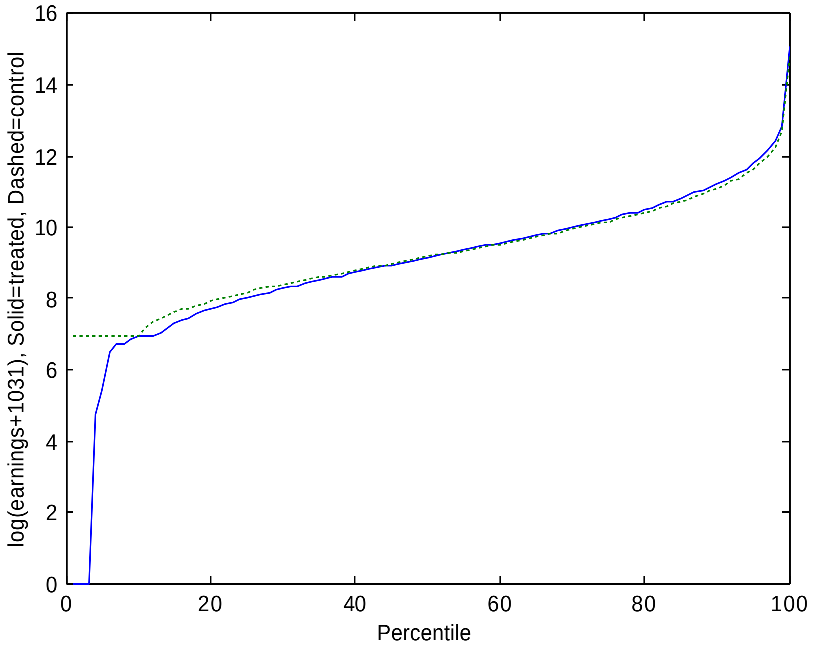
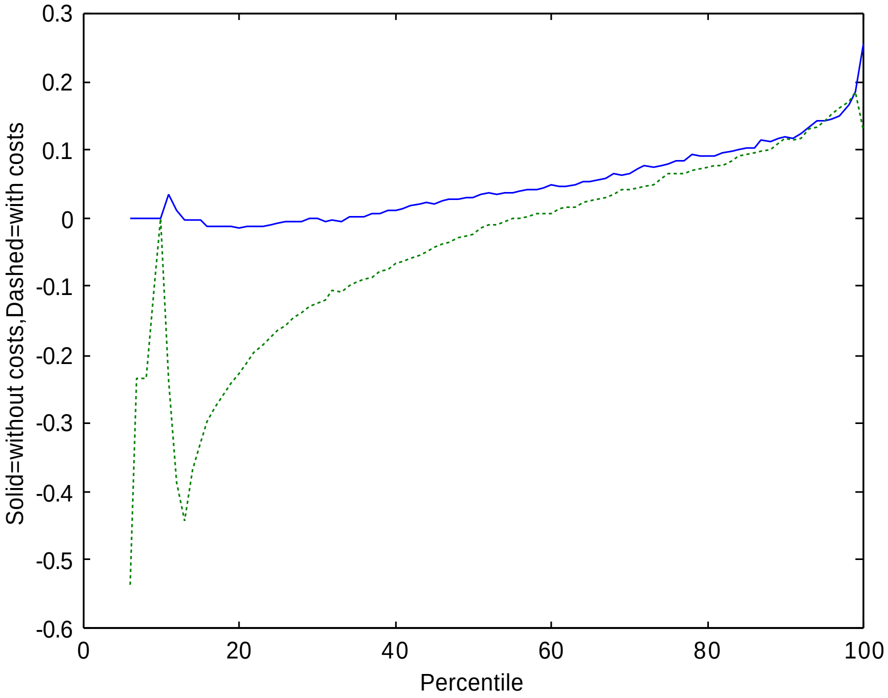


Figure 6: Difference in Percentiles of Treatment and Control Earnings, JTPA, (detail: 6th to 100th percentile)



The two strands of the analysis – risk aversion and inequality aversion – are combined in the tables below. They present the welfare rankings of the treatment and control programs using CRRA preferences and allowing for a range of risk- and inequality-attitudes. Ignoring training costs, we see that the treatment program is preferred to the control as long as the policy-maker's risk attitude is risk-neutral or risk-loving, and if the coefficient of inequality aversion is less than 5. When the coefficient of inequaliy aversion is in an intermediate range, between 2 and 4, the treatment is also preferred when the policy-maker is risk averse. RT0 corresponds to *e*=0 and *q*=0.1, a very low level of risk aversion. From Table 4 we know that RT0 leads to the treatment being rejected.

Table 5: Welfare Rankings, Earnings, JTPA

		Coefficient of Inequality Aversion									
		-4	-3	-2	-1	0	1	2	3	4	5
Coefficient of Relative Risk Aversion	-4	+	+	+	+	+		+	+	+	–
	-3	+	+	+	+	+		+	+	+	–
	-2	+	+	+	+	+		+	+	+	–
	-1	+	+	+	+	+		+	+	+	–
	0	+	+	+	+	+		+	+	+	–
	1					(*)					
	2	–	–	–	–	–		+	+	+	–
	3	–	–	–	–	–		+	+	+	–
	4	–	–	–	–	–		+	+	+	+
	5	–	–	–	+	+		+	+	+	+

Notes: (*) *q*=0.1, *e*=0 corresponds to RT0, which rejects the treatment.

When we take the training costs into account, the range of risk and inequality preferences for which the treatment is preferred shrinks. In particular, now the treatment is strictly preferred if the policy-maker is risk neutral or risk loving and coefficient of inequality aversion is less than 5. Strong inequality aversion (*e*=5) can lead to a preference for the control, presumably since the treatment lifts earnings at the upper end of the distribution and reduces earnings for the lower range. But for less extreme degrees of inequality aversion, the preference over programs is decided purely by the degree of risk aversion. In this case, RT0 again corresponds to *e*=0 and *q*=0.1, and rejects the treatment.

Table 6: Welfare Rankings, Earnings Net of Training Costs, JTPA

		Coefficient of Inequality Aversion									
		-4	-3	-2	-1	0	1	2	3	4	5
Coefficient of Relative Risk Aversion	-4	+	+	+	+	+		+	+	+	–
	-3	+	+	+	+	+		+	+	+	–
	-2	+	+	+	+	+		+	+	+	–
	-1	+	+	+	+	+		+	+	+	–
	0	+	+	+	+	+		+	+	+	–
	1					(*)					
	2	–	–	–	–	–		–	–	–	–
	3	–	–	–	–	–		–	–	–	–
	4	–	–	–	–	–		–	–	–	–
	5	–	–	–	–	–		–	–	–	–

Notes (*) *q*=0.1, *e*=0 corresponds to RT0 and rejects the treatment.

6. Conclusion

This paper has develop a set of simple rules of thumb that can be used to evaluate programs. These rules of thumb account for risk aversion and for inequality aversion. By applying these rules to data from the GAIN and JTPA experiments, we demonstrated that these two factors are very important in ranking programs. For the GAIN data, which is widely cited as an example of a successful training program, the treatment is not always preferred to the control, when we allow for inequality aversion. For the JTPA, the simple rules of thumb were unable to rank the programs, but the full analysis which combines inequality aversion, risk aversion, and heterogeneous treatment impacts demonstrated that for risk or inequality averse preferences the control program is preferred.

One of the limitations of the methods discussed is that they rely on strong assumptions. The rules of thumb use assumptions such as normality and log normality, and the non-parametric analysis does not account for covariates. All of these can be readily corrected by applying more general models and techniques. But at that point the analysis would cease to be through simple rules of thumb and would become a full-blown decision analysis. Given the impracticality of such an analysis in many contexts, the methods that this paper has developed offer a useful overview of the evaluation problem.

References

Chamberlain, Gary, and Guido Imbens (1996). "Hierarchical Bayes Models with Many Instrumental Variables," Harvard Institute of Economic Research, Paper Number 1781.

DeGroot , Maurice (1970). *Optimal Statistical Decisions*. New York: Mc-Graw Hill.

Dehejia, Rajeev (1999). "Program Evaluation as a Decision Problem," NBER Working Paper No. 6954, forthcoming, *Journal of Econometrics*.

------ and Sadek Wahba (1999). "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, Vol. 94 (448), pp. 1053-1062.

Friedlander, Daniel, David Greenberg, and Philip Robins (1997). "Evaluating Government Training Programs for the Economically Disadvantaged," *Journal of Economic of Literature*, 35, 1809-1855.

Heckman, James and J. Hotz (1989), "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862-874.

------ and Jeffrey Smith (1998). "Evaluating the Welfare State," NBER Working Paper No. 6542.

Ingersol, Jonathan (1987). *Theory of Financial Decision Making*, pp. 37-42. New York: Rowman & Littlefield.

Lalonde, Robert (1986). "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604-620.

Le Cam (1953), "On some asymptotic properties of maximum likelihood estimates and related Bayes estimates," *University of California Publications in Statistics*, Vol. 1 (11), pp. 277-330.

Levy, Haim (1998). *Stochastic Dominance: Investment Decision Making Under Uncertainty*. Boston: Kluwer Academic Publishers.

Neumann, J. von, and O. Morgenstern (1944). *Theory of Games and Economic Behaviour*. Princeton: Princeton University Press.

Samuelson, Paul (1970). "The Fundamental Approximation Theorem of Portfolio Analysis in terms of Means, Variances and Higher Moments," *Review of Economic Studies*, Vol. 37 (4), pp. 537-542.

Sen, Amartya (1973). *On Economic Inequality*. Oxford: Clarendon Press.

Wald, Abraham (1950). *Statistical Decision Functions*. New York: John Wiley & Sons.