Testing Rank Similarity

Brigham R. Frandsen* Lars J. Lefgren*

November 13, 2015

Abstract

We introduce a test of the rank invariance or rank similarity assumption common in treatment effects and instrumental variables models. The test probes the implication that the conditional distribution of ranks should be identical across treatment states using a regression-based test statistic. We apply the test to data from the Tennessee STAR class-size reduction experiment and JTPA job training program. We show that systematic slippages in rank can be important statistically and economically. We also illustrate the power of the rank similarity assumption for estimating and bounding structural parameters of interest in settings in which this assumption is plausible.
Keywords: instrumental variables, quantile regression, partial identification, treatment effects, nonseparable models

1 Introduction

Many recent developments in econometric theory concern the importance of treatment effect heterogeneity. Heterogeneity matters for the policy relevance and interpretation of treatment effects: a treatment which is beneficial on average may nevertheless inflict substantial harm, depending on how subjects differ in their response to the treatment. Treatment effect heterogeneity also matters for econometric identification of structural (causal) parameters. When individuals' relative ranks across treatment states are preserved, population—as opposed to local—average treatment effects (ATE) are identified and models restricting to scalar outcome disturbances are justified. Finally, characterizing treatment effect heterogeneity provides insights regarding the economics surrounding individuals' selection into treatment.

Generally researchers observe a subject in only one treatment state. Consequently, in order to characterize the heterogeneity of treatment effects, researchers must make some

*Department of Economics, Brigham Young University

assumption about a subject's rank in counterfactual distributions. A common benchmark is to assume rank invariance, which in the case of a binary treatment implies that the subject's rank is the same in both the treated and control distributions. Under this assumption, the distribution of treatment effects is identified and quantile treatment effects can be interpreted as effects on individuals at given points of the distribution. Furthermore, researchers can also identify an intervention's ATE even in the absence of perfect compliance to the instrument, where otherwise only a local average treatment effect (LATE) may be identified.

There is generally little reason to believe, however, that rank invariance holds. A generalization, rank similarity, requires only the conditional distribution of ranks—not the ranks themselves—to be identical in all treatment states, conditional on factors influencing treatment status (Chernozhukov and Hansen, 2005). This weaker assumption loses the ability to point identify the distribution of treatment effects, but it preserves the ability to identify ATE and population quantile treatment effects. Even the weaker rank similarity condition may be implausible in many settings, however, and deserves scrutiny. For example, it rules out the possibility that individuals use private information on comparative advantage when selecting into (or out of) treatment.

We propose a method for examining rank similarity and, by extension, its stronger version, rank invariance. We develop the test in a setting with a binary treatment that is either exogenously assigned, or partially determined by an instrument satisfying standard conditions, but note that it generalizes in a straightforward way to multivalued (including continuous) treatments. The test consists of comparing the estimated distribution of treated and control ranks, conditional on some observed variable S, using a regression-based statistic. S is taken to be a pre-determined variable in addition to any covariates X required for identification. Under rank invariance or similarity, the treated and control ranks will have identical distributions conditional on S and the test statistic has a well-known limiting distribution, while under violations the distributions will diverge, provided S is related to outcomes. For example, consider the case of a randomized job training program with perfect compliance. One might suspect that prior earnings—a candidate S variable—

is correlated with outcomes in the absence of treatment. Consequently, the distribution of ranks for control subjects with high prior earnings should have higher mass in the top of the distribution. If the treated and control distributions exhibit rank similarity with respect to prior earnings, the distribution of ranks for treated subjects with high prior earnings should be shifted in precisely the same way.

A failure to reject may justify rank similarity or invariance assumptions and the additional identification power they garner. For example, Chernozhukov and Hansen's (2005) estimator leverages rank similarity to identify ATE; imposing the special case of rank invariance further identifies the treatment effect on individuals at each quantile of distribution, as well as the distribution of treatment effects. Rank invariance also justifies classes of models with scalar outcome disturbances (Newey and Powell, 2003; Chesher, 2005; Chernozhukov, Imbens, and Newey, 2007).

Our test for rank similarity can also be used to motivate partially identifying restrictions for sharpening bounds on the distribution of treatment effects, something of a holy grail in program evaluation. For example, rank similarity's testable restrictions imply those of rank exchangeability and stochastic increasingness, assumptions which allow researchers to tighten bounds on the fraction of individuals harmed by treatment, and other features of the distribution of treatment effects (Frandsen and Lefgren, 2015).

Rejecting rank similarity is also potentially interesting. First, the testing procedure itself shows which groups have comparative advantage under the treatment. For example, finding that conditional on low levels of education treated ranks are shifted up relative to untreated ranks implies that the intervention benefits less-educated individuals relative to the more-educated. Second, a rejection implies that at a minimum the testing variable S should be included as a control if researchers nevertheless want to impose rank similarity.

Our paper complements the literature on treatment effects, nonseparable models, and quantile instrumental variables models. The notion of rank invariance in treatment effects was introduced by Doksum (1974), who used the idea of an underlying "proneness" to interpret effects on different quantiles of the outcome distribution (see also Lehmann, 1975;

Imbens and Wooldridge, 2009). Heckman et al. (1997) consider rank invariance (perfect positive dependence) as an extreme case to obtain bounds on the joint distribution of potential outcomes. In the nonparametric IV literature, the assumption that the outcome is monotonic in a scalar disturbance—separable or nonseparable—imposes rank invariance. For example, Newey and Powell's (2003) nonparametric instrumental variables model relies on a rank invariance assumption in the form of a scalar separable outcome disturbance for identification, as do Chesher (2005), Athey and Imbens (2006) and Chernozhukov, Imbens, and Newey (2007), but with a nonseparable outcome disturbance. The IV quantile model of Chernozhukov and Hansen (2005, 2008) imposes rank similarity for identification of average treatment effects and rank invariance for interpretation of quantile treatment effects. Finally, Imbens and Newey's (2009) nonseparable triangular model achieves identification for a general non-scalar outcome disturbance, but introduces rank invariance in the form of a scalar outcome disturbance for interpretation of the estimates. Our paper builds on this literature by introducing a test of the assumptions needed for identification or interpretation in these models.

The intuition for our testing procedure is similar to Bitler et al.'s (2005) idea for testing for rank reversals using pre-determined covariates in the context of an exogenous treatment. Their method examines whether individuals at a given quartile in the control distribution exhibit similar pre-treatment characteristics to individuals at the corresponding quartile of the treatment distribution. Our method builds on their work by generalizing to endogenous or multi-valued treatments. In concurrent work, Dong and Shen (2015) develop a test for rank similarity based on a similar framework. Our paper complements theirs in three primary ways. First, their test relies on the assumption of monotonicity in the relationship between the instrument and treatment to calculate ranks in the treated and untreated distributions of complier ranks. In contrast, we set up our test in a quite general estimation framework which can include monotonicity as a special case when that assumption is appropriate, but need not impose it when it is not. Second, our instrumental variables estimation-based test is directly informative about the magnitude and nature of any deviations from rank

similarity, while the reduced-form test of Dong and Shen is not. Third, we show how the test motivates partially identifying restrictions on the distribution of treatment effects.

In the remainder of this paper we describe our econometric framework and define rank similarity and specify our test in a setting with an exogenous treatment. We then generalize our test to consider cases where treatment is endogenous but where an instrument is available. We illustrate the finite-sample size and power of the test using Monte Carlo simulations. We conclude with two empirical applications in which we examine the effect of class size on student achievement and job training on wages.

2 Econometric Framework and Test Procedures

Consider the standard treatment effects framework with a binary treatment. The test we develop can be generalized to multi-valued (including continuous) treatments. A binary treatment, D, potentially affects a continuously distributed outcome Y. Let $Y(1)$ and $Y(0)$ be potential outcomes with and without treatment, with cdfs F_1 and F_0. The observed outcome is $Y = Y(D)$. In addition to outcomes and treatment, we observe a q-vector of pre-treatment variables S and, in the case of an endogenous treatment, an instrumental variable Z. The pre-treatment variables S are in addition to any covariates X required for identification, which we suppress for clarity's sake. All results below can be taken to hold conditional on X.

Define an individual's rank in the untreated and treated distributions to be, respectively:

$$\begin{aligned} U(0) &:= F_0(Y(0)) \\ U(1) &:= F_1(Y(1)). \end{aligned}$$

The marginal distributions of $U(0)$ and $U(1)$ are, as a consequence of the Skorokhod

representation of $Y(0)$ and $Y(1)$, uniform:

$$\begin{aligned} U(0) &\sim U(0,1) \\ U(1) &\sim U(0,1). \end{aligned}$$

Rank invariance and rank similarity restrict the relationship between $U(0)$ and $U(1)$. The stricter notion, rank invariance, is defined as follows:

Definition 1 (Rank Invariance) *A treatment exhibits rank invariance if and only if* $U(0) = U(1)$ *almost surely.*

Rank invariance means an individual of a given rank in the control distribution would have the same rank in the treated distribution. It implies that individuals with the same control outcome would have responded identically to treatment, as noted by Doksum (1974). While this is restrictive, it is imposed by models with scalar outcome disturbances (Newey and Powell, 2003; Chesher, 2005; Athey and Imbens, 2006; Chernozhukov, Imbens, and Newey, 2007).

A generalization of rank invariance that captures a similar notion is rank similarity, introduced by Chernozhukov and Hansen (2005):

Definition 2 (Rank Similarity) *A treatment exhibits rank similarity with respect to* S *if and only if conditional on* $S = s$, $U(0)$ *and* $U(1)$ *are identically distributed for each* s *in the support of* S.

Rank similarity means a subpopulation (defined by $S = s$) will have the same distribution of ranks across treatment states. Rank similarity only has meaning with respect to some "rank-shifting" variable S, since unconditionally $U(0)$ and $U(1)$ have identical distributions by definition. Figure 1 illustrates the concept of rank similarity graphically. The left panel shows that potential ranks are uniformly distributed marginally. The middle and right panels illustrate that the conditional distribution of potential ranks may not be

uniform if S predicts ranks. The middle column shows that under rank similarity $U(0)$ and $U(1)$ are identically distributed conditional on S, while the right panel shows that when rank similarity is violated, the conditional distributions of $U(0)$ and $U(1)$ will differ. Note that rank invariance implies rank similarity, and can be seen as the special case of rank similarity where S is itself $U(0)$ or $Y(0)$. The rank shifting variable S plays a crucial role in the testability of this restriction; it obviates the need for knowledge of or assumptions about the joint distribution of potential outcomes $(Y(0), Y(1))$ required in prior work on treatment effect heterogeneity (Heckman et al., 1997).

Rank similarity, and by extension rank invariance, imposes testable restrictions on observed data, provided the distributions of potential outcomes are identified. Intuitively, it means that conditional on S, treatment status has no effect on the distribution of ranks. The proposed test directly examines this implication via the following procedure:

1. Estimate potential outcome cdfs $\hat{F}_0$ and $\hat{F}_1$;

2. Construct sample ranks

$$\hat{U}_i = (1 - D_i)\,\hat{F}_0(Y_i) + D_i \hat{F}_1(Y_i); \tag{1}$$

3. Estimate the following specification:

$$\hat{U}_i = \alpha_0 + \alpha_1 D_i + S_i'\alpha_2 + D_i S_i'\delta + \varepsilon_i; \tag{2}$$

4. Test the hypothesis

$$H_0 : \delta = \begin{pmatrix} 0 & \cdots & 0 \end{pmatrix}'$$

using the test statistic

$$\hat{\Delta} = n\hat{\theta}' h \left(h'\hat{V}h\right)^{-1} h'\hat{\theta}, \tag{3}$$

where n is the sample size, $\hat{\theta}$ is an appropriate estimator (OLS, quantile regression,

or instrumental variables) for specification (2), h is the selection matrix

$$h = \begin{pmatrix} \mathbf{0}_{q\times q+2} & \mathbf{I}_q \end{pmatrix}',$$

and $\hat{V}$ is the estimated asymptotic variance-covariance matrix of $\hat{\theta}$.

As shown below, under the null hypothesis of rank similarity the test statistic converges asymptotically to a χ^2 random variable with q degrees of freedom. Since the $\hat{U}_i$s are constructed conditional on D_i, they are normalized within each treatment category. Therefore, α_1 is not a free parameter in regression (2), and so is not included in the test statistic. Note that in the special case of a scalar rank shifter ($q = 1$), this test is asymptotically equivalent to a standard t-test on the interaction of D_i and S_i.

The regression-based procedure allows testing violations of similarity for any feature of the rank distribution. To test for differences in the conditional expectation, equation (2) can estimate using mean regression or instrumental variables estimators for average treatment effects. To test for differences in other parts of the distribution, equation (2) can be estimated using quantile regression or instrumental quantile methods for a range of quantile indices τ. In this way the proposed test maintains power against any departure from rank similarity.

Note that the testing procedure does not require equation (2) to correspond to a correctly specified conditional mean or conditional quantile function. Least-squares and quantile regression estimators are well known to converge to a population quantity that linearly approximates the underlying conditional mean or quantile functions (White, 1980; Angrist et al., 2006). Under the null hypothesis of rank similarity, those population quantities will be identical for $U(0)$ and $U(1)$. The δ parameter in equation (2), which corresponds to the treatment-control difference in those population quantities will therefore be zero even under misspecification.

The choice of estimators for $\hat{F}_0$, $\hat{F}_1$, and specification (2) depends on the specific empirical setting. For exogenously assigned treatments, empirical cdfs and ordinary least

squares estimators suffice; settings with endogenous treatments require instrumental variables methods. The following subsections specify the details of the test for these two cases in turn.

2.1 Unconfounded treatment

First, consider the case where treatment is as good as randomly assigned:

Condition 3 (Treatment Unconfoundedness) $(Y(0), Y(1))$ *are jointly independent of* D.

Under this condition, the marginal distributions of potential outcomes are identified by the conditional empirical distribution functions:

$$\begin{aligned} \hat{F}_0(y) &= \frac{\sum_{i=1}^n 1(Y_i \leq y)(1-D_i)}{\sum_{i=1}^n (1-D_i)}, \\ \hat{F}_1(y) &= \frac{\sum_{i=1}^n 1(Y_i \leq y) D_i}{\sum_{i=1}^n D_i}. \end{aligned}$$

Sample ranks $\hat{U}_i$ are constructed from these estimators via (1), and specification (2) can be estimated via ordinary least squares:

$$\hat{\theta} = \left(W'W\right)^{-1} W'\hat{U},$$

or quantile regression:

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} \rho_\tau \left(\hat{U}_i - W_i'\theta\right)$$

where W is a matrix of observations on $W_i' := \begin{pmatrix} 1 & D_i & S_i' & D_i S_i' \end{pmatrix}$ and ρ_τ is the asymmetric loss "check" function.

2.2 Endogenous treatment

Now suppose treatment status D_i is possibly confounded, but there exists an exogenous instrument Z_i; for example, randomized treatment *assignment*. Let potential treatment status

as a function of the instrument be $D(0)$ and $D(1)$. We assume the standard conditions for a valid instrument:

Condition 4 (Instrument Validity) $(Y(0), Y(1), D(0), D(1))$ *are jointly independent of* Z; *and* $E[D|Z=1] > E[D|Z=0]$.

Condition 4 and the assumptions implicit in the potential outcomes notation correspond to the usual exclusion and relevance conditions. Under this condition and rank similarity the marginal distributions of potential outcomes are identified (Chernozhukov and Hansen, 2005). If rank similarity is replaced by a monotonicity condition on the response of treatment status to the instrument, the marginal distributions of potential outcomes among the subpopulation that responds to the instrument—compliers—are identified (Imbens and Angrist, 1994).

Implementing the test in a setting with an endogenous treatment requires instrumental variables estimation of $\hat{F}_0$, $\hat{F}_1$, and regression specification (2). This can be done in at least two ways, depending on which assumptions are appropriate for the specific empirical setting. The first is Chernozhukov and Hansen's (2006) quantile instrumental variables estimator. Given Condition 4 this approach is valid under the null hypothesis of rank similarity, and leads to a test of asymptotically correct size. The cdfs F_1 and F_0 can be estimated by inverting estimates of the quantiles of $Y_i(1)$ and $Y_i(0)$ obtained via Chernozhukov and Hansen's (2006) instrumental variables quantile regression procedure. The resulting estimates $\hat{F}_1$ and $\hat{F}_0$ then can be used to construct the sample ranks $\hat{U}_i$ via (1). The test consists of estimating the effect of the vector $W_i' := \begin{pmatrix} 1 & D_i & S_i' & D_i S_i' \end{pmatrix}$ on $\hat{U}_i$, which can be done for a particular quantile (say, the median), or integrated over all quantiles for the mean effect.

The second method is Abadie's (2003) κ-weighted least squares estimator. Given Condition 4, this approach is valid when treatment status can be assumed to respond monotonically to the instrument almost surely. A procedure based on this approach then tests rank similarity among the subpopulation of compliers. Rank invariance (and rank similarity as

imposed in Chernozhukov and Hansen (2005)) implies rank similarity among compliers, so a test based on this approach has asymptotically correct size. Using this approach, the cdfs F_1 and F_0 can be estimated directly by weighted least squares estimation of the effect of D_i on $1(Y_i \leq y) D_i$ (for $F_1(y)$) and the effect of $(1 - D_i)$ on $1(Y_i \leq y)(1 - D_i)$ (for $F_0(y)$), using Abadie's κ as weights:

$$\kappa_i = 1 - \frac{D_i(1 - Z_i)}{\Pr(Z_i = 0)} - \frac{(1 - D_i) Z_i}{\Pr(Z_i = 1)}. \tag{4}$$

The resulting estimates $\hat{F}_1$ and $\hat{F}_0$ then can be used to construct the sample ranks $\hat{U}_i$ via (1). The test is then carried out by weighted regression of $\hat{U}_i$ on the vector W_i.

3 Asymptotic Theory

The test statistics are based on consistent, asymptotically normal estimators, and thus have straightforward limiting distributions under standard regularity conditions. Critical values and p-values can then be calculated by consistent estimation of the limiting distribution, or by simulation methods. The test may be based on a variety initial estimators for the potential outcome cdfs F_0 and F_1 and the rank regression (2), depending on whether D_i is endogenous and (under endogeneity) which instrumental variables assumptions are most appropriate in the particular setting (e.g., monotonicity). The result below establishes the limiting distribution given general uniformly consistent asymptotically normal estimators for the potential outcome cdfs and the rank regressions. Leading examples include ordinary least squares, the Abadie (2003) IV estimator, and Chernozhukov and Hansen (2005) discussed in the previous section.

The limiting distribution of $\hat{\theta}$, and hence of the test statistic $\hat{\Delta}$, depends in general on the first-step estimators $\left(\hat{F}_0, \hat{F}_1\right)$, which can be viewed as infinite-dimensional nuisance parameters. The result here adopts the framework of Newey (1994) and Newey and McFadden (1994) for taking into account the first-step estimation in the final limiting distribution. The correction for the first step consists of an adjustment term added to the limiting variance-

covariance matrix, but the final estimator $\hat{\theta}$ continues to be consistent and asymptotically normal under regularity conditions. The test statistic (3), therefore, as a quadratic form in an asymptotically normal estimator, converges to a χ^2 random variable, as the following theorem establishes.

Theorem 5 *Let the vector of observed data be* $T_i' = (Y_i, W_i', Z_i)$ *with joint distribution* P_0, *where* $W_i' := \begin{pmatrix} 1 & D_i & S_i' & D_i S_i' \end{pmatrix}$. *Suppose (i)* $\left(\hat{F}_0, \hat{F}_1\right)$ *are uniformly consistent asymptotically normal estimators for* (F_0, F_1) *such that* $\sqrt{n}\left(\hat{F}_d(y) - F_d(y)\right) \rightsquigarrow \mathbb{G}_d(0, \sigma_d(y, y'))$ *for* $d \in \{0, 1\}$, *where* $\mathbb{G}_0$ *and* $\mathbb{G}_1$ *are independent Gaussian processes, and where* $\rightsquigarrow$ *denotes uniform convergence as a process indexed by* y; *(ii)* $\hat{\theta}$ *satisfies* $\sum_{i=1}^n m\left(T_i, \hat{\theta}, \hat{F}\right)/n = 0$ *with* $E\left[\|m(T_i, \theta_0, F)\|\right] < \infty$ *and* $M := \frac{\partial}{\partial \theta} E\left[m(T_i, \theta, F)\right]|_{\theta=\theta_0}$ *nonsingular; and (iii) assumptions A1-A6 in the Appendix are satisfied. Then the test statistic* $\hat{\Delta}$ *has the following limiting distribution:*

$$\begin{aligned} \hat{\Delta} &= n\hat{\theta}' h \left(h' \hat{V} h\right)^{-1} h' \hat{\theta} \\ &\underset{d}{\to} \chi^2(q), \end{aligned}$$

where $\hat{V}$ *is a consistent estimator of* $V = M^{-1}\Omega M^{-1\prime}$, *and* Ω *is the limiting variance-covariance matrix of* $\sum_{i=1}^n \left(m(T_i, \theta_0, F) + \alpha(T_i)\right)/\sqrt{n}$, *and* $\alpha(T_i)$ *is defined in the appendix.*

The result guarantees that the test will have asymptotically correct size. The test's power derives from differences in the conditional distribution of ranks across treatment states under alternatives to rank similarity. If these differences lead to shifts in the mean or other features of the distribution captured in specification (2), then $h'\theta$ will have nonzero elements, and the test's power will depend on the corresponding noncentral χ^2 distribution.

Note that the limiting distribution here is derived without assuming iid data, to accommodate applications with clustered sampling or other dependence structures.

4 Test Implications

Having outlined our method for testing for rank similarity, it is helpful to discuss the implications both of a rejection of the null of rank similarity as well as a failure to reject. If the test rejects rank similarity with respect to S, researchers may want to exercise caution in using models that impose rank similarity, or its stronger version, rank invariance, since a failure of rank similarity to hold with respect to observables may make it less plausible that it holds with respect to unobservables. The rejection would also affect the interpretation of the estimates from quantile regression and quantile IV estimates. In particular, the coefficients could not be interpreted as effects on individuals at particular points in the control distribution.

However, a failure of rank similarity with respect to S does not preclude rank similarity *conditional* on S with respect to unobserved factors, although as we mentioned above, conditional rank similarity may be less compelling if the test rejects. If researchers nevertheless found conditional rank similarity plausible, our test can be used to determine the set of covariates which must be conditioned upon in order to employ methods relying on rank similarity. For example, rank similarity may not hold with respect to gender even though it holds with respect to other variables of interest. In this case, instead of rejecting rank similarity altogether, researchers might simply wish to condition upon gender or perform their analysis separately by gender.

When our test fails to reject, researchers can leverage the assumption of rank similarity, and related assumptions, to employ a variety of very useful empirical methods. For example, one can employ the quantile IV methods developed by Chernozhukov and Hansen (2005). This allows researchers to employ IV methods in the absence of monotonicity. Researchers can also identify the effect of treatment on the quantiles of the distribution of outcomes not only for compliers but for the entire population. Consequently, researchers can identify the ATE in addition to the LATE.

A failure to reject rank similarity may also provide researchers with a justification to

impose a variety of related assumptions. In particular, rank invariance is a special case of rank similarity. If researchers are willing to impose this assumption, then quantile methods identify not only the effect of treatment on the distribution of outcomes but also individual-level treatment effects. For example, a median regression yields not only the effect of treatment on the median outcome but also yields the effect of treatment for individuals who are at the median of the control distribution. Researchers can also employ the models with scalar outcome disturbances (Newey and Powell, 2003; Chesher, 2005; Chernozhukov, Imbens, and Newey, 2007).

Even when rank similarity is justified by the test, however, one may not want to go as far as imposing rank invariance. A weaker assumption that is also justified by the rank similarity test is weak stochastic increasingness between ranks. This assumption, introduced in Frandsen and Lefgren (2015), means $U(0)$ and $U(1)$ are each weakly stochastically increasing in the other. It includes rank invariance and rank independence as special cases, and has the testable restriction that $U(0)$ and $U(1)$ are shifted in the same direction by observed variables S. This restriction is implied by rank similarity, and thus the rank similarity test we develop here can be interpreted as a particularly strict test of stochastic increasingness, since it requires that potential ranks be shifted in not only the same direction, but by the same magnitude. The test for stochastic increasingness in Frandsen and Lefgren (2015) does not require this restriction.

Rank similarity also motivates the related assumption of rank exchangeability. Rank exchangeability means the joint distribution of $U(0)$ and $U(1)$ is symmetric, and (when combined with weak stochastic increasingness) has the same testable restriction as rank similarity that $U(0)$ and $U(1)$ are shifted identically in distribution by observed variables S. As a result, the rank similarity test can serve also as a joint test of exchangeability and weak stochastic increasingness.

4.1 Bounds on the Treatment Effect Distribution

Frandsen and Lefgren (2015) demonstrate that the assumptions of stochastic increasing-

ness and exchangeability are useful for bounding features of the distribution of treatment effects. Stochastic increasingness implies bounds on the distribution of treatment effects conditional on $Y(0)$ and S. Specifically, $\Pr(Y(1) - Y(0) \leq t | Y(0), S)$ can be sharply bounded between estimable functions $F^L(t|Y(0), S)$ and $F^U(t|Y(0), S)$. The expressions for these bounds given in Frandsen and Lefgren (2015) simplify to the following under rank similarity:

$$F^L(t|Y(0), S) := \begin{cases} 0 & , \quad F_1(Y(0) + t) < U(0) \\ \frac{F_{U|S}(F_1(Y(0)+t)|S) - F_{U|S}(U(0)|S)}{1 - F_{U|S}(U(0)|S)} & , \quad F_1(Y(0) + t) \geq U(0) \end{cases} \tag{5}$$

and from above by

$$F^U(t|Y(0), S) := \begin{cases} \frac{F_{U|S}(F_1(Y(0)+t)|S)}{F_{U|S}(U(0)|S)} & , \quad F_1(Y(0) + t) \leq U(0) \\ 1 & , \quad F_1(Y(0) + t) \geq U(0) \end{cases}, \tag{6}$$

where $F_{U|S}$ is the cdf of ranks conditional on S. These pointwise bounds can be integrated to obtain bounds on the overall distribution of treatment effects to obtain bounds on, for example, the fraction of individuals hurt by treatment. The examples below show these bounds can be substantially tighter than the classical bounds based on Frechet-Hoeffding limits in Williamson and Downs (1990). Frandsen and Lefgren (2015) show that the assumption of rank exchageability can further tighten bounds on the overall distribution of treatment effects in certain cases.

5 Monte Carlo Simulations

5.1 Finite-sample size and power of the rank similarity test

The asymptotic theory implies that for sufficiently large samples the proposed testing procedures will have approximately correct size. This section shows that the tests also have good size and power properties in finite samples.

The simulations generate iid samples of size n from the following data generating pro-

cess. An observed rank-shifting variable is generated as $S_i \sim N\left(0, \sigma_S^2\right)$. The untreated potential outcome is generated as $Y_i(0) = \beta S_i + \varepsilon_i$, where $\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right)$, independently of S_i. For fixed σ_S^2 and σ_ε^2, β parameterizes the informativeness of the rank-shifting variable S_i. The treated potential outcome is constructed as $Y_i(1) = (1-\omega) Y_i(0) + \omega\eta_i$, where $\eta_i \sim N\left(0, \sigma_\eta^2\right)$ independently of S_i and ε_i and $\omega \in [0,1]$. The parameter ω controls the degree of departure from rank invariance (or similarity). The special case $\omega = 0$ corresponds to rank invariance, while $\omega = 1$ corresponds to a setting where one's treated rank is orthogonal to one's untreated rank. The instrument Z_i is assigned as in a totally randomized experiment where half of the subjects are assigned $Z_i = 0$ and the other half are assigned $Z_i = 1$, independently of $(S_i, \varepsilon_i, \eta_i)$. Treatment status exhibits negative selection: potential treatment status for $Z_i = z$ is $D_i(z) = 1(\rho Y_i(0) + \zeta_i \leq \gamma(z - 1/2))$, where $\zeta_i \sim N\left(0, \sigma_\zeta^2\right)$ independently of $(S_i, \varepsilon_i, \eta_i, Z_i)$, γ is a constant that governs the strength of the instrument, and $\rho \in [0,1]$. The special case of $\rho = 0$ corresponds to an exogenous treatment. In this setup, the local average treatment effect (LATE) is equal to zero, as is the average treatment effect (ATE). Unless stated otherwise, the simulations set $n = 1,000$, $\sigma_S^2 = .5$, $\beta = .75$, $\sigma_\varepsilon^2 = .75$, $\sigma_\eta^2 = 1$, $\rho = 0$, $\sigma_\zeta^2 = 1$, and $\gamma = 3$.

The first set of simulations examines the finite-sample size of the test for sample sizes ranging from $n = 100$ to $n = 1,000$ for an exogenous and endogenous treatment. The degree of departure from rank invariance or rank similarity is naturally set at $\omega = 0$. The degree of endogeneity is set at $\rho = 0$ for the exogenous treatment and $\rho = .3$ for the endogenous treatment. The simulations show that even for relatively small sample sizes the test maintains accurate size under both exogeneity and endogeneity. Figure 2 plots the simulated rejection rate for tests of nominal size $\alpha = .05$ by sample size. The rejection rate hovers around five percent over the entire range.

The next set of simulations explores the test's power to detect departures from rank invariance or similarity. The simulations vary the degree of departure between $\omega = 0$ and $\omega = 1$. The simulations show that the rejection probability increases rapidly as a function of the degree of departure from rank invariance or rank similarity. Figure 3 plots the simulated

rejection rates by ω for tests of nominal size $\alpha = .05$. At the far left of the plot, where $\omega = 0$, the rejection rate is approximately .05, indicating correct size. As violations of rank invariance are introduced, however, the rejection rate increases steeply. The test rejects 75 percent of the time when $\omega = .5$ and essentially 100 percent of the time when $\omega = 1$.

The next set of simulations examines the test's power when the sample size ranges from $n = 100$ to $n = 1,000$. The degree of departure from rank invariance is set at $\omega = .75$. The simulations show the test is lower-powered for small sample sizes, but has excellent power for sample sizes common in empirical analysis. Figure 4 plots the simulated rejection rates by n for tests of nominal size $\alpha = .05$. The rejection rate is about 40 percent for the smallest sample size of $n = 100$, but increases to over 99 percent for $n = 1,000$.

The next set of simulations examines how the test's power varies with the informativeness of the rank-shifting variable S_i. The simulations vary the β parameter from zero to one, corresponding to an R^2 between the control outcome and S_i of 0 to .3. The degree of departure from rank invariance is set at $\omega = .75$. The simulations show the test's power depends heavily on the informativeness of the rank-shifting variable. Figure 5 plots the simulated rejection rates by R^2 for tests of nominal size $\alpha = .05$. The rejection rate is .05 when S_i is completely uninformative, but reaches essentially 100 percent when the R^2 reaches .3.

The final set of simulations shows how the strength of the instrument affects the test's power. The simulations vary the instrument strength parameter from $\gamma = 0$ to $\gamma = 6$, corresponding to compliance rates from zero to 100 percent. The simulations show the test's power increases substantially with the strength of the instrument. Figure 6 plots simulated rejection rates by instrument strength as measured by compliance rate for tests of nominal size $\alpha = .05$. The figure shows the power is near the size for a completely weak instrument, and thus the test is essentially uninformative, as expected. Power increases significantly, however, as the instrument becomes stronger, and reaches essentially 100 percent for compliance rates near 100 percent.

6 Empirical Examples

6.1 Class Size and Student Achievement

In our first example, we test for rank similarity in the context of the Project STAR class size reduction experiment. Starting in 1985, the state of Tennessee randomized kindergarten students either into small classes of 13 to 17 students or into larger classes of 22 to 25 students. Some larger classrooms were randomly assigned a teacher's aide. Given the importance of the question and strength of the research design, data from this study have been analyzed many times. These studies include works by Folger and Breda (1989), Finn and Achilles (1990), Word et al. (1990) and Krueger (1999).

For this analysis, we consider only the effect of assignment to a small class. Prior research (e.g. Krueger, 1999) suggests the impact of an aide was minimal. We restrict our sample to students whose assignment occurred in kindergarten and were still in the sample at the end of first grade. This includes 1,368 treated students and 3,040 control students. We examine math performance at the end of first grade. On average, students assigned to the smaller classes performed 0.20 standard deviations better than students assigned to the larger classes. This positive effect is statistically significant at the 1 percent level.

To test for rank similarity, we compute math performance rank in both the treatment and control groups. Because we do not observe math performance prior to classroom assignment, we choose eligibility for free lunch in kindergarten as our rank shifting variable, S. Figure 7 shows how the distribution of ranks in the control distribution varies according to whether subjects were eligible (lower income) or ineligible (higher income) for free lunch. The density of ranks is shifted clearly to the right for control subjects with higher family income relative to subjects with lower family income.

We now examine the distributions of ranks in the treated and control distributions for different income levels. In Figure 8, we compare the distribution of ranks for treated and untreated subjects who had lower family incomes. Figure 9 represents the corresponding figure for subjects who had higher family incomes. Under the null hypothesis of rank

similarity, we expect the rank distributions between treated and control observations to be similar in both figures. Examining the rank distributions of students who had lower family incomes, we see that the bottom of the distribution has less mass in the treatment group relative to the control group. Similarly, if we examine the rank distribution of students with higher family incomes, the bottom of the distribution has more mass in the treatment group than the control group. This suggests that treatment may have been more efficacious for struggling students with low incomes than for struggling students with higher incomes. This visual evidence suggests that the assumption of rank similarity may be unfounded.

To test more formally for rank similarity and establish the power of the rank shifter, we estimate regression specification (2) via ordinary least squares:

$$\hat{U}_i = \alpha_0 + \alpha_1 D_i + \alpha_2 S_i + \delta S_i D_i + \varepsilon_i$$

Table 1 shows the results of this estimation. We see that the estimate for α_2 is -0.208, which implies that students who are eligible for free lunch are located nearly 21 percentage points lower in the control distribution than students who are ineligible for free lunch. This is highly significant and together with Figure 7 suggests that our rank shifter has strong power. The estimate for δ is 0.054 and statistically significant at the 1 percent level suggesting that the rank disadvantage of being from a low income family is smaller in the treatment group than in the control group. This test suggests that we should reject the null hypothesis of rank similarity.

6.2 Job Training Program and Earnings

In our second empirical example we examine the National Job Training Partnership Act (JTPA) Study. This study was funded by the U.S. Department of Labor to evaluate the extent to which programs funded by the Job Training Partnership Act of 1982 improved employment outcomes among participants. The act funded both private and public entities to provide a variety of services including classroom instruction in occupational skills, job

search assistance, and on-the-job training. While the program was first funded in 1983, the analysis sample came from training that was provided between 1987 and 1989. Published estimates of the impact of the estimated impacts of the JTPA can be found in Bloom et al. (1997).

This example is particularly interesting from a methodological perspective. Participation in a program was not directly controlled by the research design; rather, subjects were randomly assigned eligibility, but could choose whether to take up treatment. Thus, this example illustrates the test in a setting with imperfect compliance where instrumental variables methods are required. Since the test suggests rank similarity is plausible here (as we shall see) it also highlights the additional identifying power rank similarity and related assumptions garner.

We include in our analysis sample 4,337 adult men with non-missing prior year earnings. We use as our outcome variable combined earnings in the thirty months after treatment. Using treatment assignment as an instrument for program participation, we find that assignment to treatment increased the probability of treatment by 62.3 percentage points (from a baseline of 1.1 percentage points). Assuming monotonicity, the LATE for this combined sample of men was \$2,287, which is significant at the five percent level. These findings are similar to those calculated in Bloom et al. (1997) and Abadie et al. (2002).

We perform the rank similarity test under the monotonicity assumption that eligibility did not induce any subject *not* to participate. Given that only about one percent of individuals in the control group participated, this seems like a very reasonable assumption. Monotonicity allows us to use Abadie's κ-weighting method to estimate the conditional distribution of ranks among compliers. Thus, strictly speaking, the test here tests for rank similarity among compliers. Since rank similarity among compliers is a necessary condition for overall rank similarity, it remains a valid test.

Examining Figure 10, we see how the distribution of ranks differ in the control complier distribution according to prior earnings. We see that high prior earnings is associated with a strong shift to the right in the rank distribution. This demonstrates that our rank shifter

has strong statistical power, making it more likely that we can detect deviations from rank similarity.

Figure 11 shows estimated rank distributions of subjects with low prior earnings in treatment and control distributions. We would expect these to be very similar if rank similarity holds. We see that the distributions do look quite similar. Figure 12 also suggests that the treated and control distributions appear similar for subjects with high prior earnings. Together, this graphical evidence suggests that we are unlikely to be able to reject the null hypothesis of rank similarity.

We test this explicity by estimating equation (2) via a κ-weighted regression. The estimates of this regression are found in Table 2. In the first column, our rank shifter is a binary variable that takes on a value of one if the subject had above median prior earnings and zero otherwise. In this specification, the main effect of above-median prior earnings is 0.174, which implies that workers with above median earnings in the prior year are located approximately 17 percentage points higher in the control distribution than workers with below median earnings. This provides strong numeric evidence regarding the power of our rank shifter. The estimated interaction between treatment and above-median earnings is -0.028 and statistically insignficant. Together α_2 and δ suggest that high prior earnings increased subjects' rank in the treatment distribution by only 15 percentage points. However, the difference between the treatment and control group is not large enough for us to reject the null hypothesis of rank similarly. The standard errors are small enough that we would have been able to detect a significant difference if the effect of high prior earnings for the treatment group was more than 25 percent larger or smaller than for the control group. Column 2 shows very similar results when the rank shifter is a continuous measure of prior earnings. Collectively, the evidence does not suggest that the null hypothesis of rank similarity should be rejected.

Given that rank similarity holds, we can use Chernozhukov and Hansen to estimate the effect of treatment on the quantiles the outcome distribution. In Figure 13, we report the estimated treatment effects for each quantile. We see that the treatmente effects appear

very small for the bottom quantiles in the outcome distribution. However, the relationship is upward sloping with economically and statistically significant benefits accruting to the higher quantiles of the outcome distribution. For quantiles above the median, the effect of treatment appears to increase earnings by between two and six thousand dollars. If in addition to assuming rank similarity, we are further willing to assume rank invariance, these estimates can be interpreted as the treatment effects for individuals in those quantiles. Integrating across the quantiles, we obtain ATE, which is estimated as $2,135 with a standard error of $1,210. This estimate is very similar to the LATE reported above.

Our test of rank similarity suggests that individual ranks in the treated and control distribution are shifted (in distribution) in the same way by prior earnings. This is consistent with the assumption of stochastic increasingness in that increasing the rank in the control distribution is associated with improvements in the expected rank in the treated distribution. If one is willing to assume stochastic increasingness, using the method outlined by Frandsen and Lefgren (2015), one can bound the expected treatment effects for each level of the control outcome. One can also bound other features of the treatment effect distribution, such as the probability that an individual would be harmed by treatment. In Figures 14 we see bounds on the average treatment effect by quantile of the control distribution. We see that both upper and lower bounds are strictly above zero for the lowest quantiles. However, as the quantiles increase the lower bound becomes increasingly negative. The upper bound tends to be stable at about $20,000. Examining the bounds on the fraction of participants hurt, the lower bound is always zero. We see that for observations in the bottom 15 percent of the control distribution, the upper bound probability of being hurt is also zero. This upper bound rises, plateauing at about .8 between the 60th and 85th percentile of the control distribution before rising again at the very highest quantiles of the control distribution. Integrating across the pointwise bounds, we calculate bounds on the overall fraction of individuals injured by treatment. We find that the fraction of individuals harmed by the training programs can be bounded between 0 and 46 percent. This compares quite favorably to the classical bounds (Williamson and Downs, 1990) of 0 and

68 percent. Finally, our test is consistent not only with rank similarity but also with the slightly stronger assumption of rank exchangeability. Imposing this assumption allows us to tighten the upper bound on the fraction of individuals injured by treatment to 44 percent.

7 Conclusion

This paper proposed a test for rank invariance or rank similarity in a treatment effects setting, and showed how to construct bounds on the distribution of individual-level treatment effects when rank similarity holds. The test is based on standard least squares or instrumental variables estimation methods, and applies to exogenous and endogenous treatments. The test is consistent and simulations show it has good size and power properties in finite samples.

Empirical examples from job training programs and classroom interventions show rank preservation deserves scrutiny in real-life data. Rank similarity was rejected in test scores from the Tennessee STAR class-size experiment. The test showed that the JTPA data are consistent with rank similarity, allowing us to identify ATE in that setting and bound features of the distribution of treatment effects.

The proposed test should prove useful in examining the identification and interpretation of program evaluations, clinical trials, and other treatment effect estimates. The procedure applies to binary treatments, but the framework can be applied more generally. Specifying the test and constructing bounds for multi-valued or continuous treatments is left to future research.

References

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113:231–263, 2003.

Alberto Abadie, Joshua D. Angrist, and Guido Imbens. Instrumental variables estimates of

the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70: 91–117, 2002.

Joshua D. Angrist, Victor Chernozhukov, and Ivan Fernandez-Val. Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica*, 74:539–563, 2006.

Susan Athey and Guido Imbens. Identification and inference in nonlinear difference-in-difference models. *Econometrica*, 74:431–497, March 2006.

Marianne P. Bitler, Jonah B. Gelbach, and Hilary W. Hoynes. Distributional impacts of the self-sufficiency project. Working Paper 11626, National Bureau of Economic Research, September 2005.

Howard S. Bloom, Larry L. Orr, Stephen H. Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M. Bos. The benefits and costs of JTPA Title II-A programs: Key findings from the National Job Training Partnership Act Study. *The Journal of Human Resources*, 32(3):549–576, 1997.

Victor Chernozhukov and Christian Hansen. An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261, January 2005.

Victor Chernozhukov and Christian Hansen. Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491 – 525, 2006.

Victor Chernozhukov and Christian Hansen. Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, 142(1):379–398, January 2008.

Victor Chernozhukov, Guido W. Imbens, and Whitney K. Newey. Instrumental variable estimation of nonseparable models. *Journal of Econometrics*, 139(1):4 – 14, 2007. Endogeneity, instruments and identification.

Andrew Chesher. Nonparametric identification under discrete variation. *Econometrica*, 73 (5):pp. 1525–1550, 2005.

Kjell Doksum. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, 2(2):pp. 267–277, 1974.

Yingying Dong and Shu Shen. Testing for rank invariance or similarity in program evaluation: The effect of training on earnings revisited. Unpublished working paper, 2015.

Jeremy D. Finn and Charles M. Achilles. Answers and questions about class size: A statewide experiment. *American Educational Research Journal*, 28:557–77, 1990.

John Folger and Carolyn Breda. Evidence from project star about class size and student achievement. *Peabody Journal of Education*, 67(1):17–33, 1989.

Brigham R. Frandsen and Lars J. Lefgren. Weak stochastic increasingness, rank exchangeability, and partial identification of the distribution of treatment effects. Unpublished manuscript, 2015.

James J. Heckman, Jeffrey Smith, and Nancy Clements. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535, October 1997.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Guido W. Imbens and Whitney K. Newey. Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):pp. 1481–1512, 2009.

Guido W. Imbens and Jeffrey M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):pp. 5–86, 2009.

Alan B. Krueger. Experimental estimates of education production functions. *Quarterly Journal of Economics*, 114:497–532, 1999.

Erich Leo Lehmann. *Nonparametrics: Stastistical Methods Based on Ranks*. Holden-Day, San Francisco, 1975.

Whitney K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):pp. 1349–1382, 1994.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):pp. 1565–1578, 2003.

Halbert White. Using least squares to approximate unknown regression functions. *International Economic Review*, 21:149–170, 1980.

R. C. Williamson and T. Downs. Probabilistic arithmetic. i. numerical methods for calculating convolutions and dependency bounds. *Int. J. Approx. Reasoning*, 4(2):89–158, March 1990.

Elizabeth Word, John Johnston, Helen Pate Bain, BD Fulton, Jayne Boyd Zaharias, Charles M Achilles, Martha Nannette Lintz, John Folger, and Carolyn Breda. The state of tennessee's student/teacher achievement ratio (star) project: Technical report 1985-1990. Technical report, Tennessee State Department of Education, Nashville, 1990.

Appendix

The following regularity conditions, adapted from Newey (1994), are assumed in Theorem 5, proved below.

A1 (Linearization) (i) There is a function $D\left(T_i, F'\right)$ that is linear in F' such that for all F' with $\|F' - F\|$ small enough,

$$\left\|m\left(T_i, F'\right) - m\left(T_i, F\right) - D\left(T_i, F' - F\right)\right\| \leq b\left(T_i\right)\left\|F' - F\right\|^2;$$

(ii) $E\left[b\left(T_i\right)\right]\sqrt{n}\left\|\hat{F} - F\right\|^2 \underset{p}{\to} 0.$

A2 (Stochastic Equicontinuity) $\sum_{i=1}^{n}\left[D\left(T_i,\hat{F}-F\right)-\int D\left(t,\hat{F}-F\right)dP_0\left(t\right)\right]/\sqrt{n}\underset{p}{\to}$ 0.

A3 (Mean-square Continuity) (i) There is $\alpha\left(T_i\right)$ and a measure $\hat{P}$ such that $E\left[\alpha\left(T_i\right)\right]=0$, $E\left[\left\|\alpha\left(T_i\right)\right\|^2\right]<\infty$, and for all $\left\|\hat{F}-F\right\|$ small enough, $\int D\left(t,\hat{F}-F\right)dP_0\left(t\right)=\int\alpha\left(t\right)d\hat{P}\left(t\right)$. (ii) For the empirical distribution $\tilde{P}\left(t\right)=n^{-1}\sum_{i=1}^{n}1\left(T_i\leq t\right)$, $\sqrt{n}\left[\int\alpha\left(t\right)d\hat{P}\left(t\right)-\int\alpha\left(t\right)d$ 0.

A4 There are $\varepsilon,\left\|F'\right\|,b\left(T_i\right),\tilde{b}\left(T_i\right)>0$ such that (i) for all $\theta\in\Theta$, $m\left(T_i,\theta,F\right)$ is continuous at θ with probability one, $\left\|m\left(T_i,\theta,F\right)\right\|<b\left(T_i\right)$; $\left\|m\left(T_i,\theta,F'\right)-m\left(T_i,\theta,F\right)\right\|\leq\tilde{b}\left(T_i\right)\left(\left\|F'-F\right\|\right)^{\varepsilon}$.

A5 $E\left[m\left(T_i,\theta,F\right)\right]=0$ has a unique solution on Θ at θ_0, and Θ is compact.

A6 (i) $\theta_0\in$interior(Θ); (ii) there is $\left\|F'\right\|,\varepsilon>0$, and a neighborhood $\mathcal{N}$ of θ_0 such that for all $\left\|F'-F\right\|<\varepsilon$, $m\left(T_i,\theta,F'\right)$ is differentiable in θ on $\mathcal{N}$; (iii) Assumption A4 is satisfied with $m\left(T_i,\theta,F'\right)$ there equal to each row of $\partial m\left(T_i,\theta,F'\right)/\partial\theta$.

Assumption A1 is satisfied trivially by the moment conditions defining the least squares, κ-weighted least squares, and instrumental variables quantile estimators proposed in Section 2, since in those cases $m\left(\cdot,\cdot,\cdot\right)$ is itself linear in F.

Assumption A2 is a standard stochastic equicontinuity condition that is satisfied by sufficiently smooth moment conditions, including those proposed in Section 2.

Assumption A3 gives the conditions required of the adjustment term, $\alpha\left(T_i\right)$. The adjustment term itself is a pathwise derivative of the moment condition with respect to a parametric path through the general class of distributions to which P_0 belongs. Specifically, let the data's joint distribution P_0 belong to some family of general distributions $\mathcal{P}$. Let P_ω denote a parametric family within $\mathcal{P}$, with $\omega=0$ corresponding to the true distribution P_0. Then as shown in Newey (1994), the adjustment term satisfies

$$\partial E\left[m\left(T_i,\theta_0,F\right)\right]/\partial\omega=E\left[\alpha\left(T_i\right)S\left(T_i\right)\right],$$

where $S(T_i) = \partial \ln P_\omega(T_i)/\partial\omega$. The form of the adjustment term will depend on the moment conditions corresponding to the chosen estimator for θ_0, and is derived below for least squares and κ-weighted least squares.

Assumptions A4-A5 are standard assumptions for identification and uniform convergence of $\hat{\theta}$. Assumption A6 is standard for assuring asymptotic normality of $\hat{\theta}$.

Proof of Theorem 5. Assumptions A1-A6 together with the conditions on m stated in the Theorem satisfy the conditions of Lemma 5.3 in Newey (1994), which establishes that $\sqrt{n}\left(\hat{\theta} - \theta_0\right)$ converges in distribution to a normal random variable with variance V defined in the Theorem. Denote $A_n := \sqrt{n}h'\hat{\theta}$ and $\Sigma := h'Vh$, and note that $h'\theta_0 = \mathbf{0}_{q\times 1}$. Then the continuous mapping theorem implies A_n converges in distribution to a multivariate normal with covariance matrix Σ:

$$\sqrt{n}\left(h'\hat{\theta} - h'\theta_0\right) = \sqrt{n}h'\hat{\theta} = A_n \underset{d}{\to} N\left(\mathbf{0}_{q\times 1}, \Sigma\right).$$

Since Σ is symmetric positive definite, it has the Cholesky decomposition

$$\Sigma = CC',$$

and likewise its inverse has decomposition

$$\Sigma^{-1} = C'^{-1}C^{-1}.$$

By the continuous mapping theorem the random vector $G_n = C^{-1}A_n$ converges to a multivariate normal with mean $\mathbf{0}_{q\times 1}$ and variance matrix $C^{-1}\Sigma C'^{-1} = C^{-1}CC'C'^{-1} = \mathbf{I}$. But the test statistic $\hat{\Delta} = n\hat{\theta}'h\left(h'\hat{V}h\right)^{-1}h'\hat{\theta} = A_n'\Sigma^{-1}A_n$ can be written as $G_n'G_n$, and so by another application of the continuous mapping theorem, it converges by definition to a χ^2 distribution with q degrees of freedom. ■

The limiting distribution of the test statistic depends on an adjustment term $\alpha(T_i)$ that takes into account the first-step estimation of potential ranks. The following result derives

the adjustment for the case when D_i is exogenous and ordinary least squares is used to estimate equation (2).

Lemma 6 *Let* $\hat{\theta}$ *satisfy* $\sum_{i=1}^{n} m\left(T_i, \hat{\theta}, \hat{h}\right)/n = 0$, *where* $m(t, \theta, h) = w(h(y,d) - w'\theta)$, $t' = (y, w')'$, *and* $h(y,d) = F_{Y|D}(y|d) = \Pr(Y \leq y | D = d)$. *Define* $g(y,d) = E[W_i | Y_i \leq y, D_i = d]$. *Then* $\partial E[m(T_i, \theta_0, h(Y_i, D_i, \omega))]/\partial\omega = E[\alpha(T_i) S(T_i)]$ *is satisfied by*

$$\alpha(Y_i, D_i) = g(Y_i, D_i) U_i - E[g(Y_i, D_i) U_i].$$

Proof. Begin with the pathwise derivative of $E[m(T_i, \theta_0, h(Y_i, D_i; \omega))]$ with respect to a parametric path indexed by ω:

$$\begin{aligned}
& \frac{\partial}{\partial\omega} E\left[W_i\left(h(Y_i, D_i; \omega) - W_i'\theta_0\right)\right] \\
= & \frac{\partial}{\partial\omega} \int_w \int_y \sum_d w \int_v 1(v \leq y) f_{Y|D}(v|d;\omega)\, dv f(w, y, d)\, dy dw \\
= & \frac{\partial}{\partial\omega} \int_w \int_y \sum_d w \int_v (1 - 1(y \leq v)) f_{Y|D}(v|d;\omega)\, dv f(w, y, d)\, dy dw \\
= & \frac{\partial}{\partial\omega} \int_w \int_y \sum_d w \underbrace{\int_v f_{Y|D}(v|d;\omega) dv}_{1} f(w, y, d)\, dy dw \\
& - \frac{\partial}{\partial\omega} \int_w \int_y \sum_d w \int_v 1(y \leq v) f_{Y|D}(v|d;\omega)\, dv f(w, y, d)\, dy dw.
\end{aligned}$$

Since $\int_v f_{Y|D}(v|d;\omega) = 1$ for any path ω, the first derivative term is zero. Changing the order of integration in the second term gives

$$= -\frac{\partial}{\partial\omega} \sum_d \int_v \underbrace{\int_w \int_y w 1(y \leq v) f_{WY|D}(w, y|d)\, dy dw}_{E[W|Y \leq v, D=d] F_{Y|D}(v|d)} \Pr(D = d) f_{Y|D}(v|d;\omega)\, dv,$$

which, noting that $\int_w \int_y w1\left(y \leq v\right) f_{WY|D}\left(w,y|d\right) dydw = E\left[W1\left(Y \leq v\right)|D=d\right] = g\left(v,d\right) F_{Y|D}\left(v|d\right)$, and exchanging differentiation and integration, yields

$$\begin{aligned}
&= -\sum_d \int_v g\left(v,d\right) F_{Y|D}\left(v|d\right) \frac{\frac{\partial}{\partial \omega} f_{Y|D}\left(v|d;\omega\right)}{f_{Y|D}\left(v|d\right)} f_{Y|D}\left(v|d\right) \Pr\left(D=d\right) dv \\
&= -\sum_d \int_v g\left(v,d\right) F_{Y|D}\left(v|d\right) S\left(v|d\right) f_{YD}\left(v,d\right) dv \\
&= -E\left[g\left(Y,D\right) F_{Y|D}\left(Y|D\right) S\left(T\right)\right] \\
&= -E\left[\left(g\left(Y,D\right) U - E\left[g\left(Y,D\right) U\right]\right) S\left(T\right)\right] \\
&= -E\left[\alpha\left(Y,D\right) S\left(T\right)\right],
\end{aligned}$$

where $S\left(v|d\right)$ denotes the score conditional on $D=d$. ■

The adjustment for the case when D_i is endogenous and Abadie κ-weighted regression is used to estimate equation (2) is similar except we define $g\left(y,d\right) = E\left[W_i|Y_i \leq y, D_i = d, D_i\left(1\right) > D_i\left(0\right)\right]$.

Table 1: STAR Rank Similarity Test

Regressor	OLS coefficient
Constant	0.605**
	(0.007)
Lower income	-0.208**
	(0.010)
Treatment status	-0.024
	(0.012)
(Lower income) x (Treatment status)	0.054**
	(0.018)
Observations	4408

Notes: Ordinary least squares estimates and robust standard errors of a regression of within-treatment rank on the variables in the left-hand column.

Table 2: JTPA Rank Similarity Test

	Specification	
Regressor	Binary	Continuous
Constant	0.408**	0.329**
	(0.014)	(0.021)
Pre-treatment earnings rank	0.174**	0.326**
	(0.018)	(0.034)
Treatment status	0.019	0.014
	(0.016)	(0.025)
(Pre-treatment earnings rank) x (Treatment status)	-0.028	-0.013
	(0.023)	(0.040)
Observations	4,337	4,337

Notes: Abadie-κ-weighted least squares estimates and robust standard errors of a regression of within-treatment rank on the variables in the left-hand column. The instrument is treatment assignment. In the Binary specification Pre-treatment earnings rank is an indicator for whether earnings are above median.

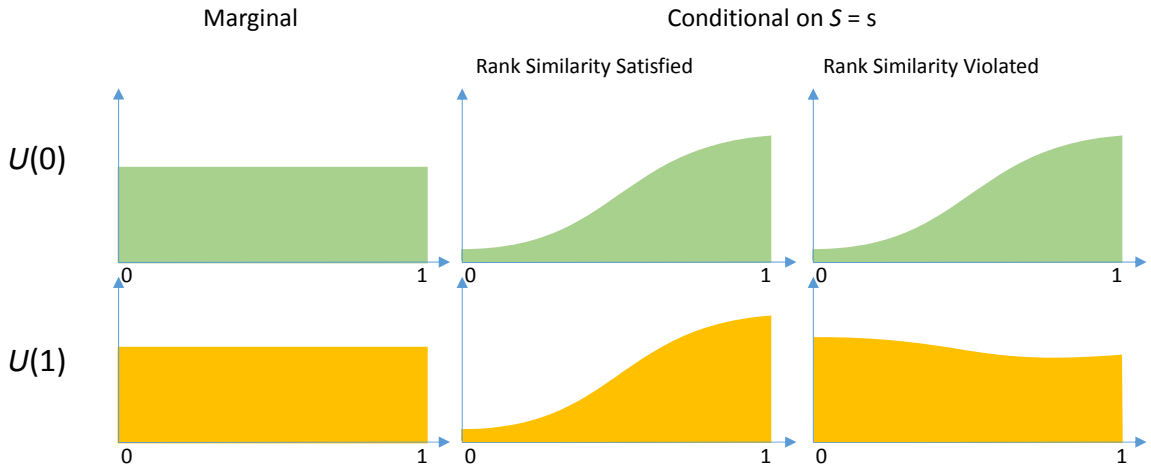

Figure 1: Illustrative distributions of potential ranks. The left column shows the marginal distributions. The middle column illustrates the conditional distribution of ranks when rank similarity is satisfied. The right column illustrates the conditional distribution of ranks when rank similarity is violated.

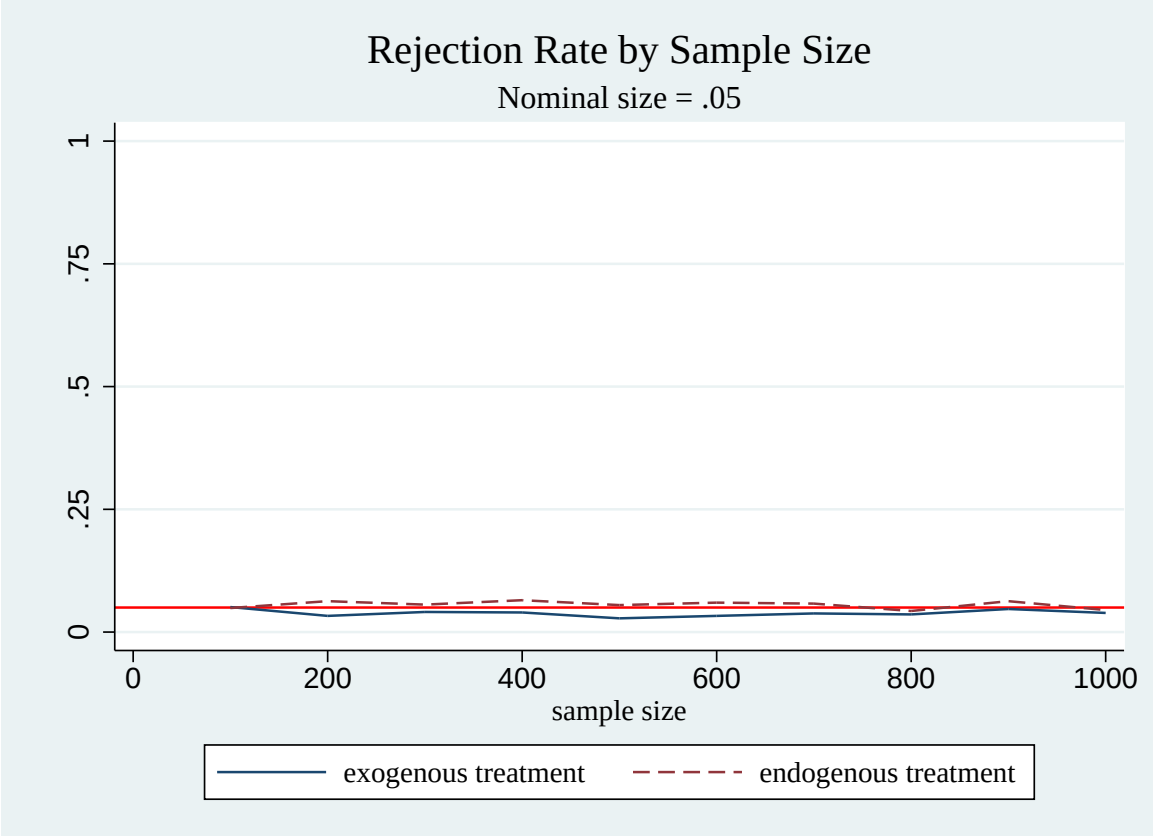

Figure 2: Monte Carlo simulation rejection rates from the rank similarity test as a function of the sample size (x-axis). Simulation model and parameters are described in the text. The nominal size of the tests is .05. Based on 1,000 iterations.

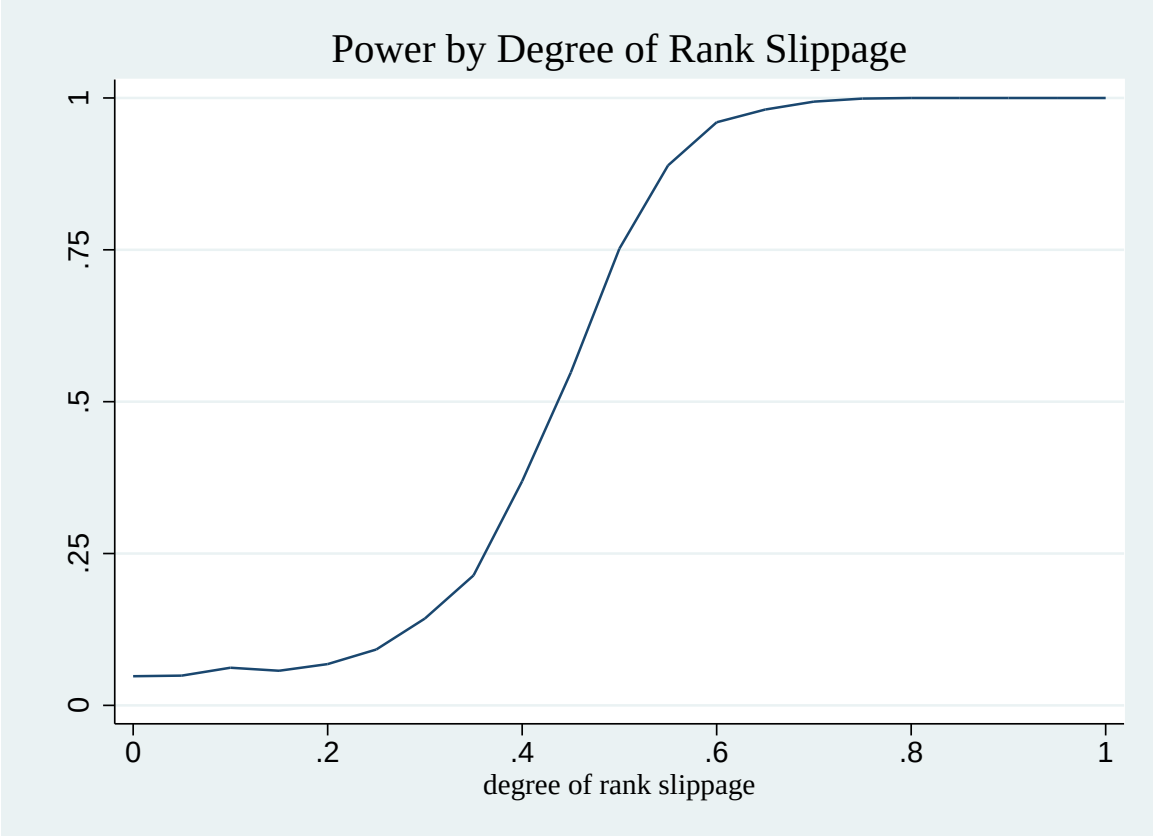

Figure 3: Monte Carlo simulation rejection rates from the rank similarity test as a function of ω, the degree of departure from rank similarity (x-axis). Simulation model and parameters are described in the text. The nominal size of the tests is .05. Based on 1,000 iterations with sample size $n = 1,000$.

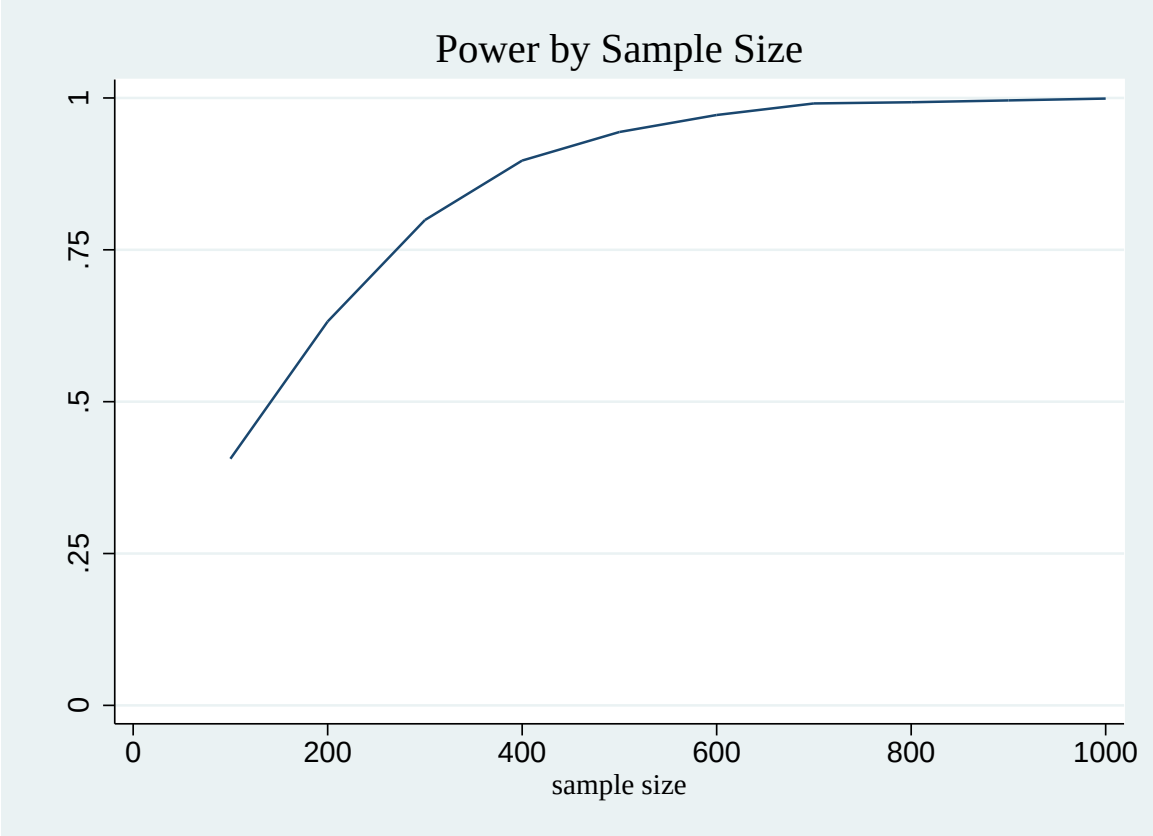

Figure 4: Monte Carlo simulation rejection rates from the rank similarity test as a function of the sample size (x-axis). Degree of departure from rank similarity is set at .75. Simulation model and parameters are described in the text. The nominal size of the tests is .05. Based on 1,000 iterations.

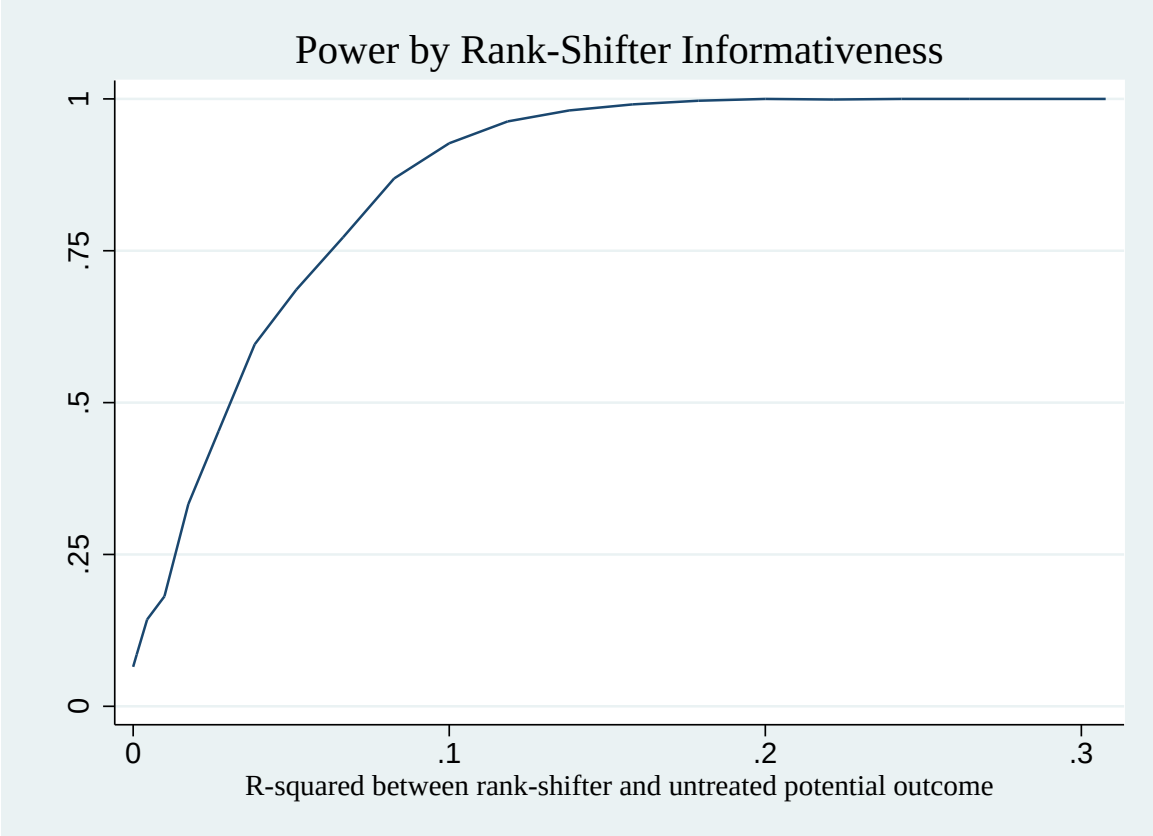

Figure 5: Monte Carlo simulation rejection rates from the rank similarity test as a function of the R^2 between the rank-shifting covariate and untreated potential outcomes (x-axis). Degree of departure from rank similarity is set at .75. Simulation model and parameters are described in the text. The nominal size of the tests is .05. Based on 1,000 iterations with sample size $n = 1,000$.

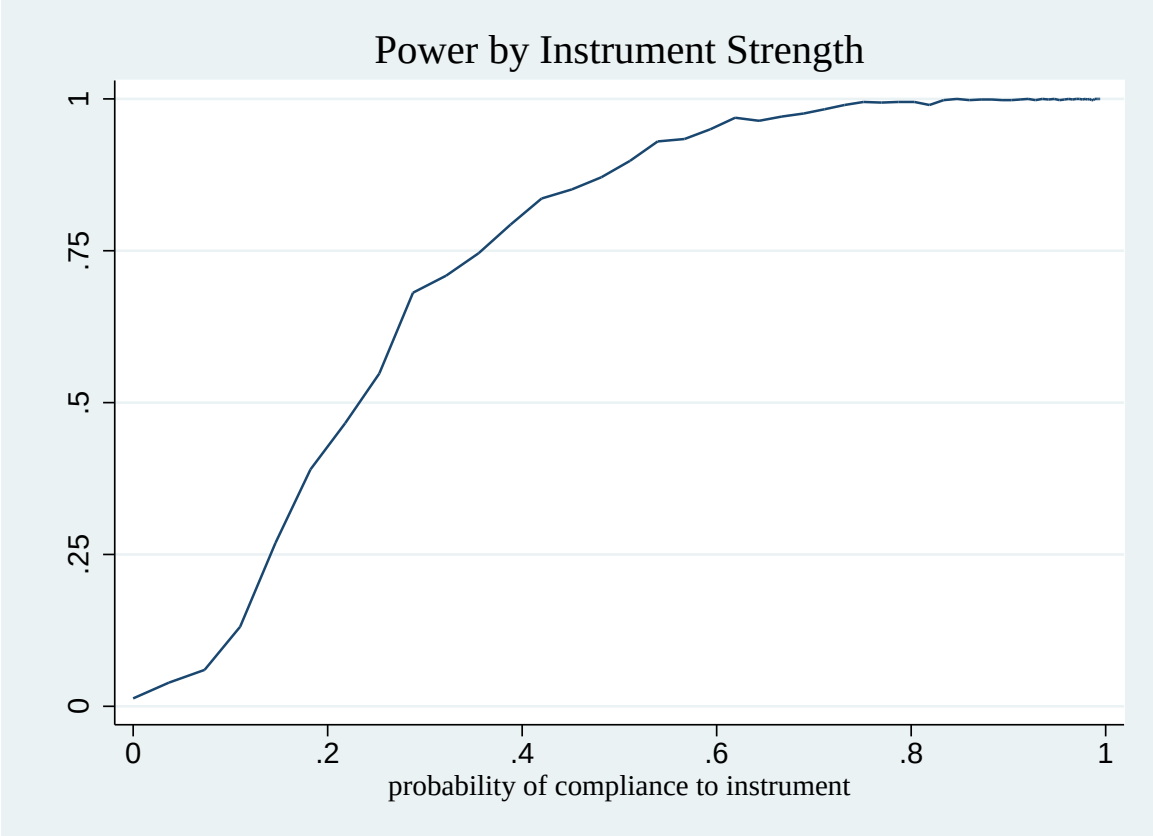

Figure 6: Monte Carlo simulation rejection rates from the rank similarity test as a function of the strength of the instrument as measured by the probability of compliance to assigned treatment (x-axis). Degree of departure from rank similarity is set at .75. Simulation model and parameters are described in the text. The nominal size of the tests is .05. Based on 1,000 iterations with sample size $n = 1,000$.

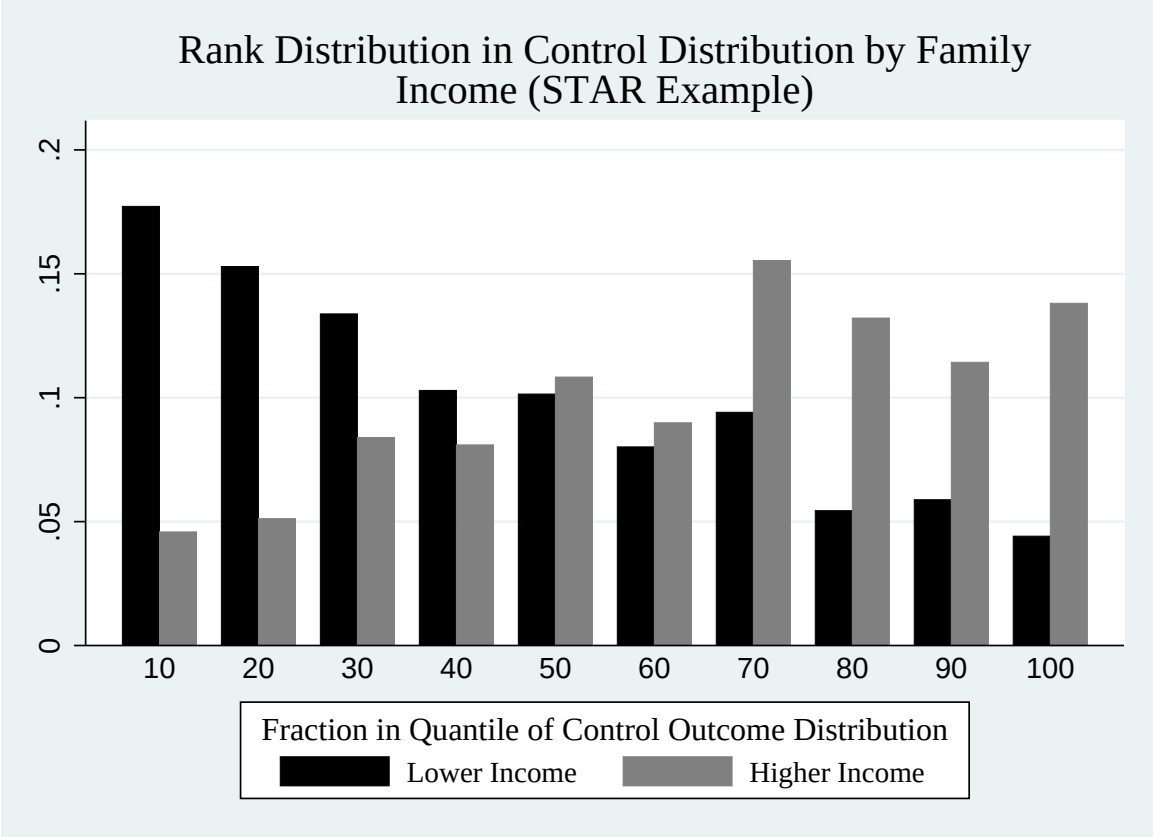

Figure 7: Histograms of sample ranks among control subjects in the Tennessee STAR class-size experiment. Lower income corresponds to eligibility for free or reduced-price lunch.

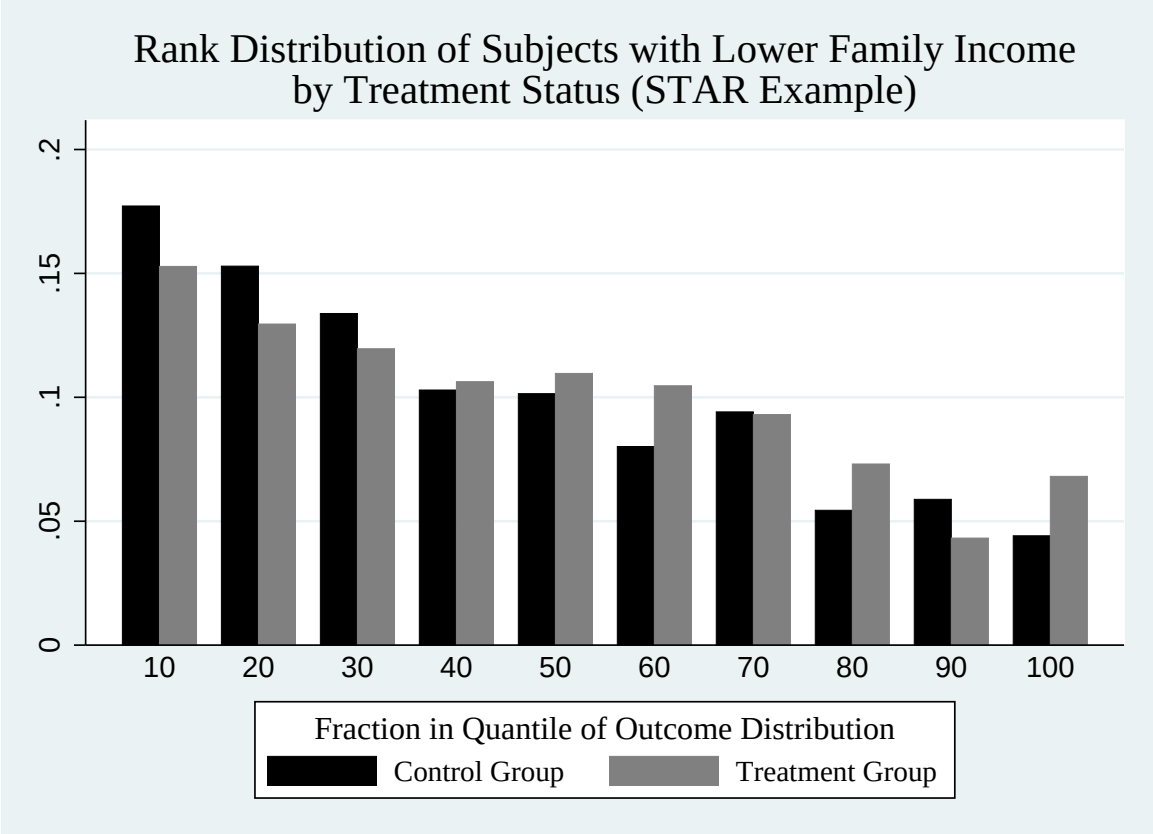

Figure 8: Histograms of sample ranks among students eligible for free or reduced-price lunch in the Tennessee STAR class-size experiment. Sample ranks are defined within treatment status.

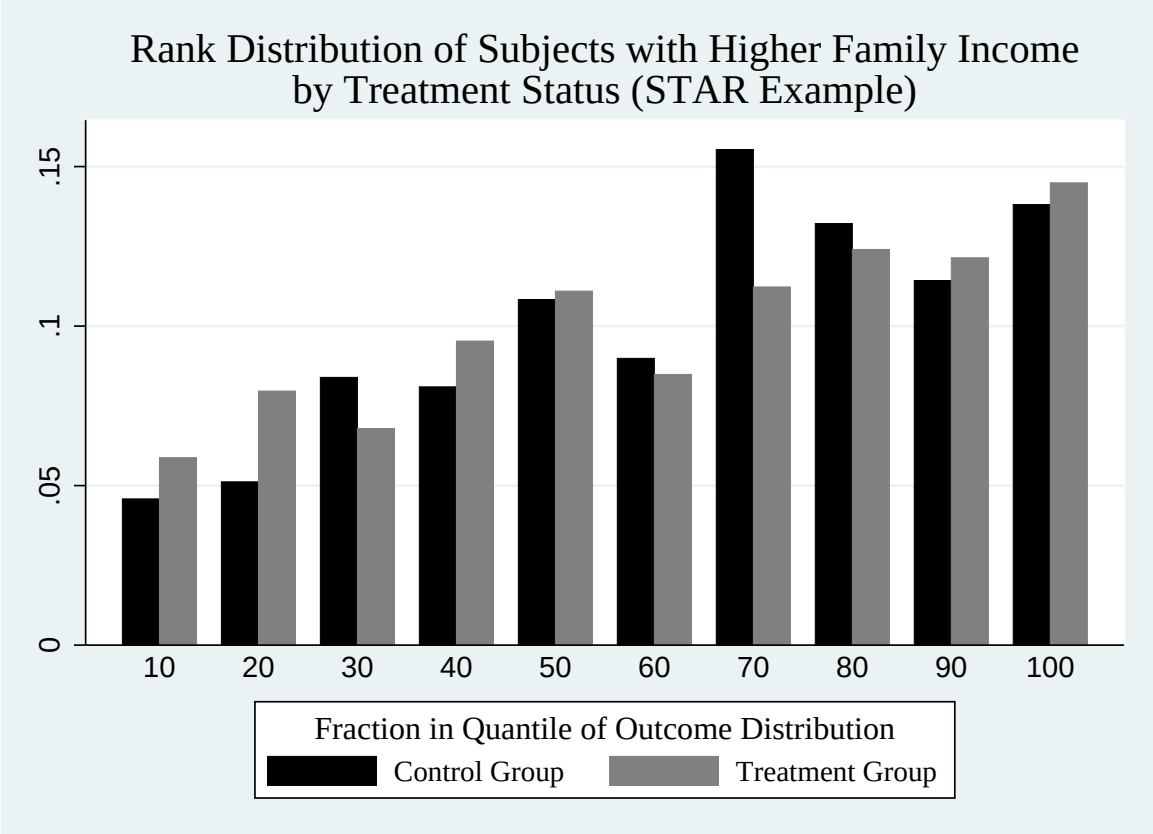

Figure 9: Histograms of sample ranks among higher-income students (those who do not qualify for free or reduced-price lunch) in the Tennessee STAR class-size experiment. Sample ranks are defined within treatment status.

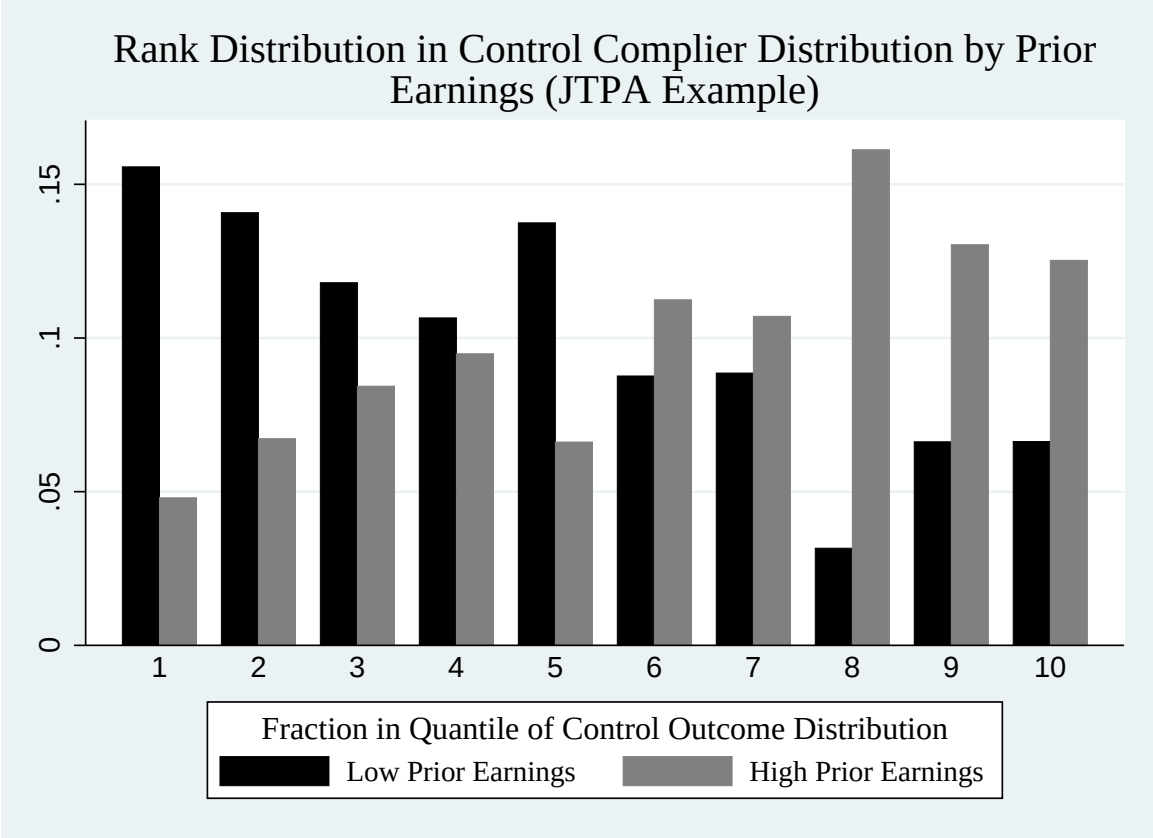

Figure 10: Histograms of sample ranks among control subjects in the JTPA experiment. Lower prior earnings corresponds to below median.

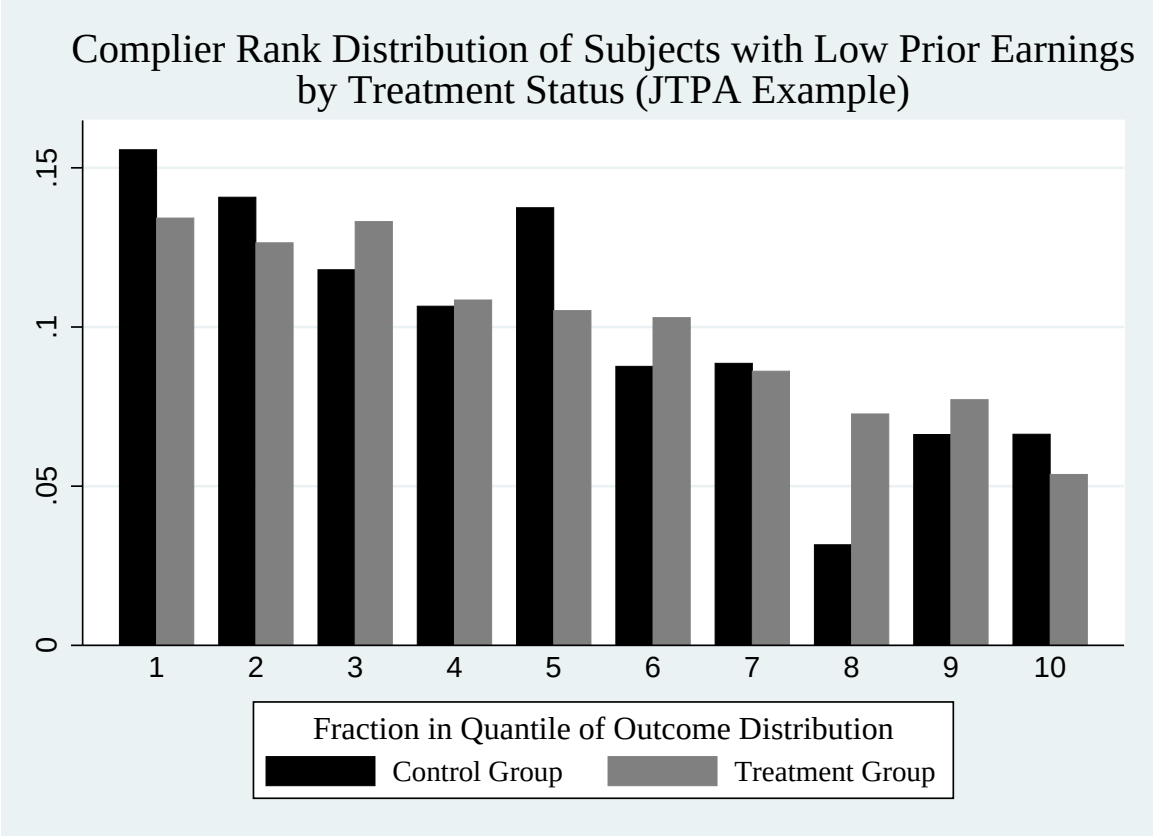

Figure 11: Histograms of sample ranks among low-prior-earning men in the JTPA experiment. Sample ranks are defined within treatment status.

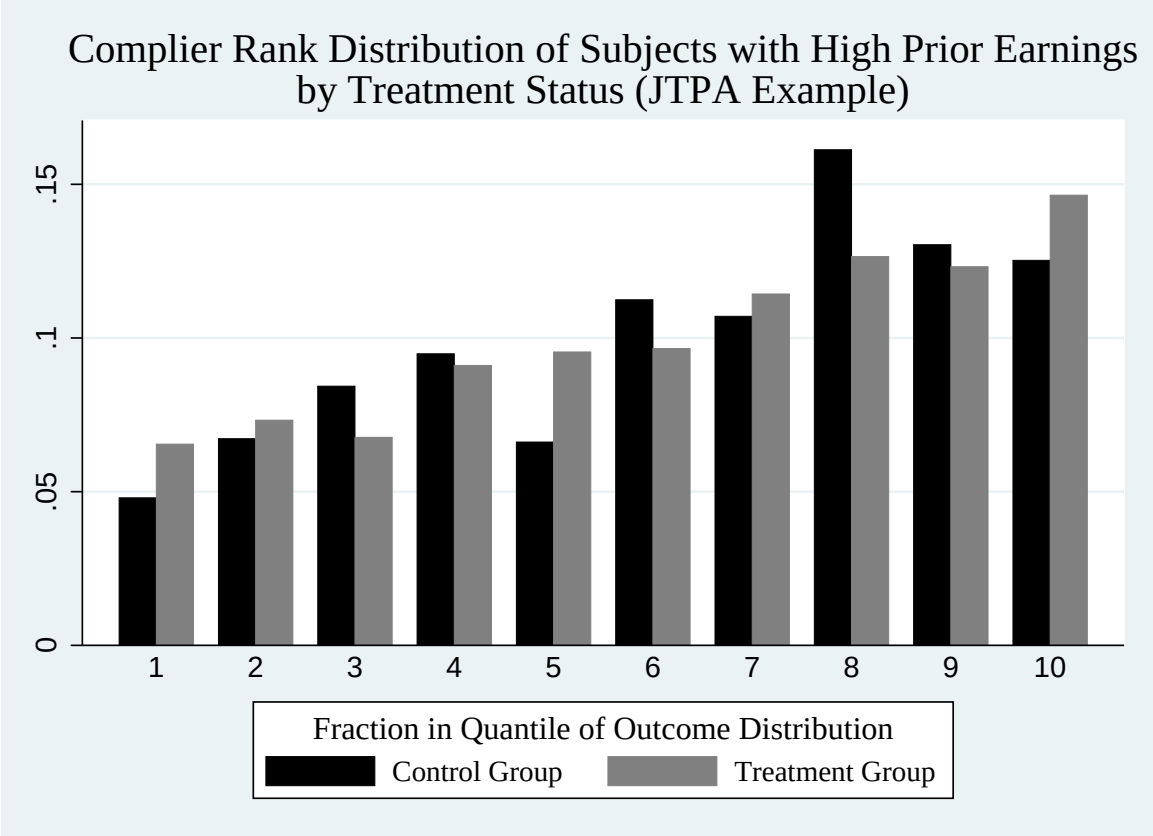

Figure 12: Histograms of sample ranks among high-prior-earning men in the JTPA experiment. Sample ranks are defined within treatment status.

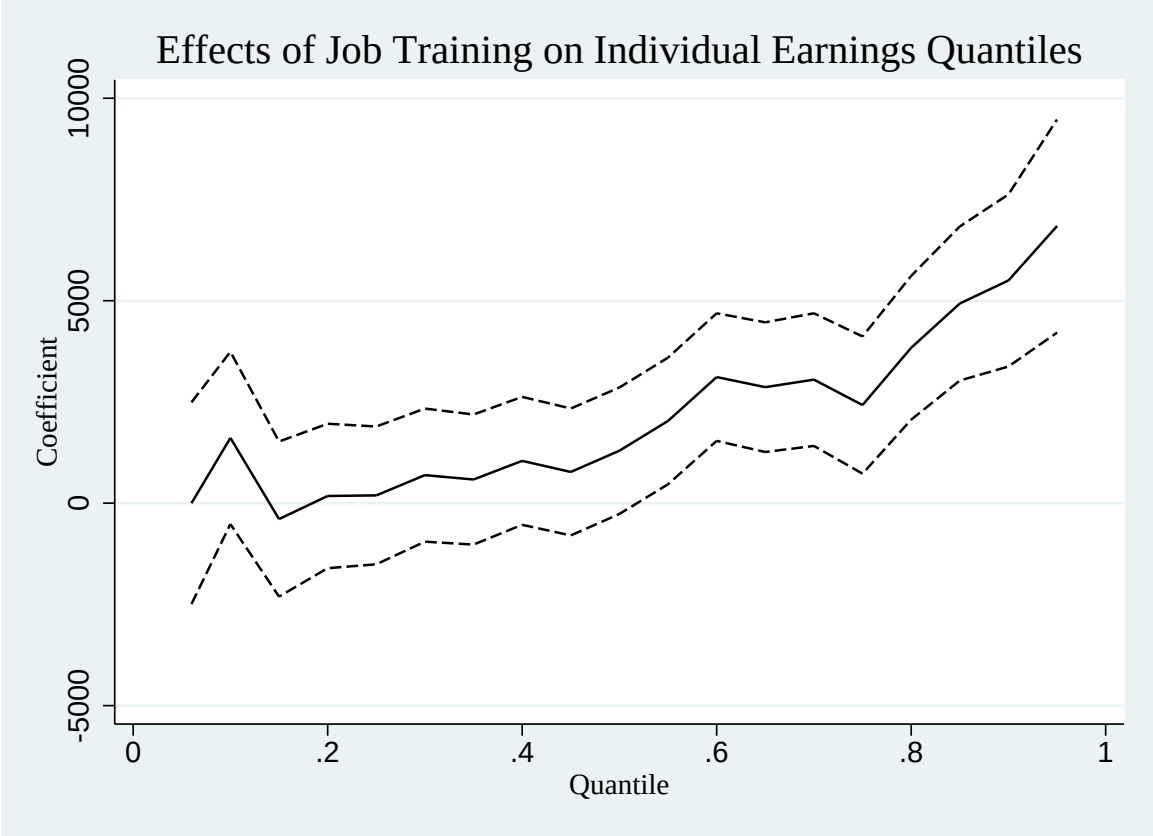

Figure 13: Quantile treatment effects estimates and 95-percent confidence intervals for the effects of participation in JTPA programs on trainee earnings. Estimates employ Chernozhukov and Hansen's (2006) procedure.

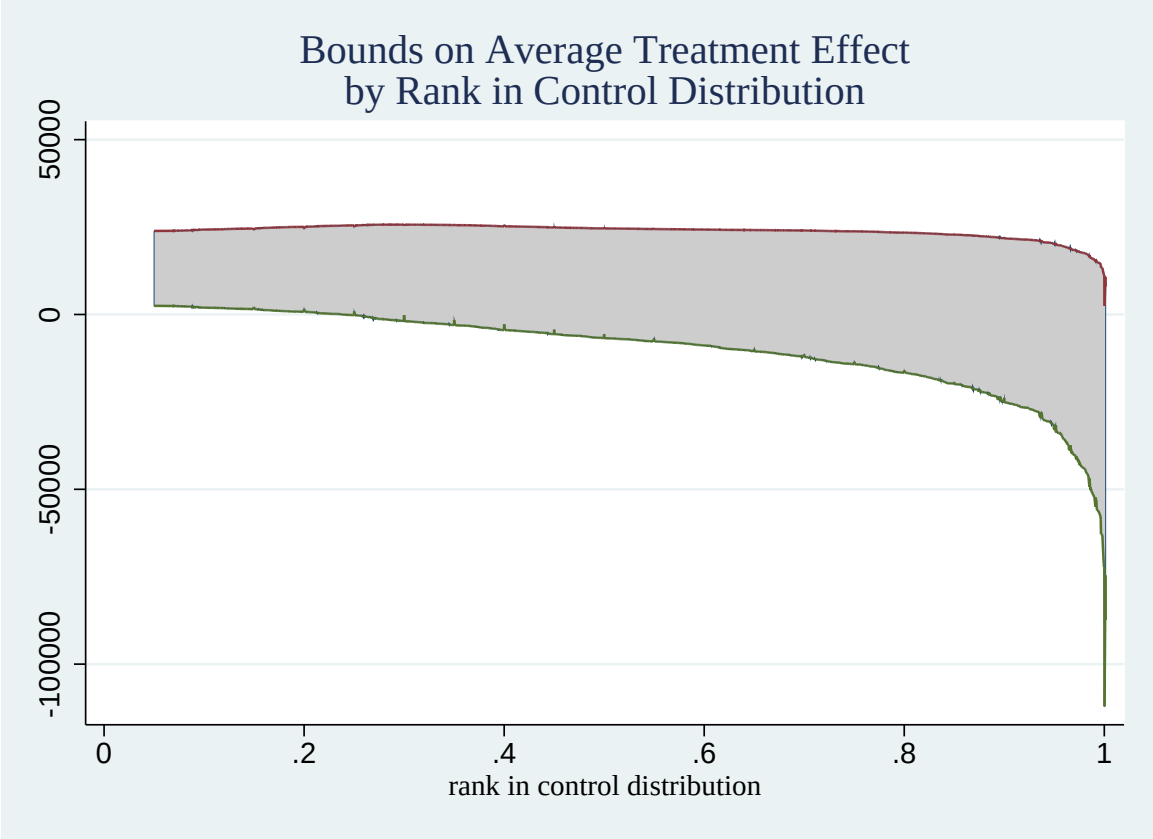

Figure 14: Estimated bounds on the expected individual-level effect of job training by rank in the control distribution.

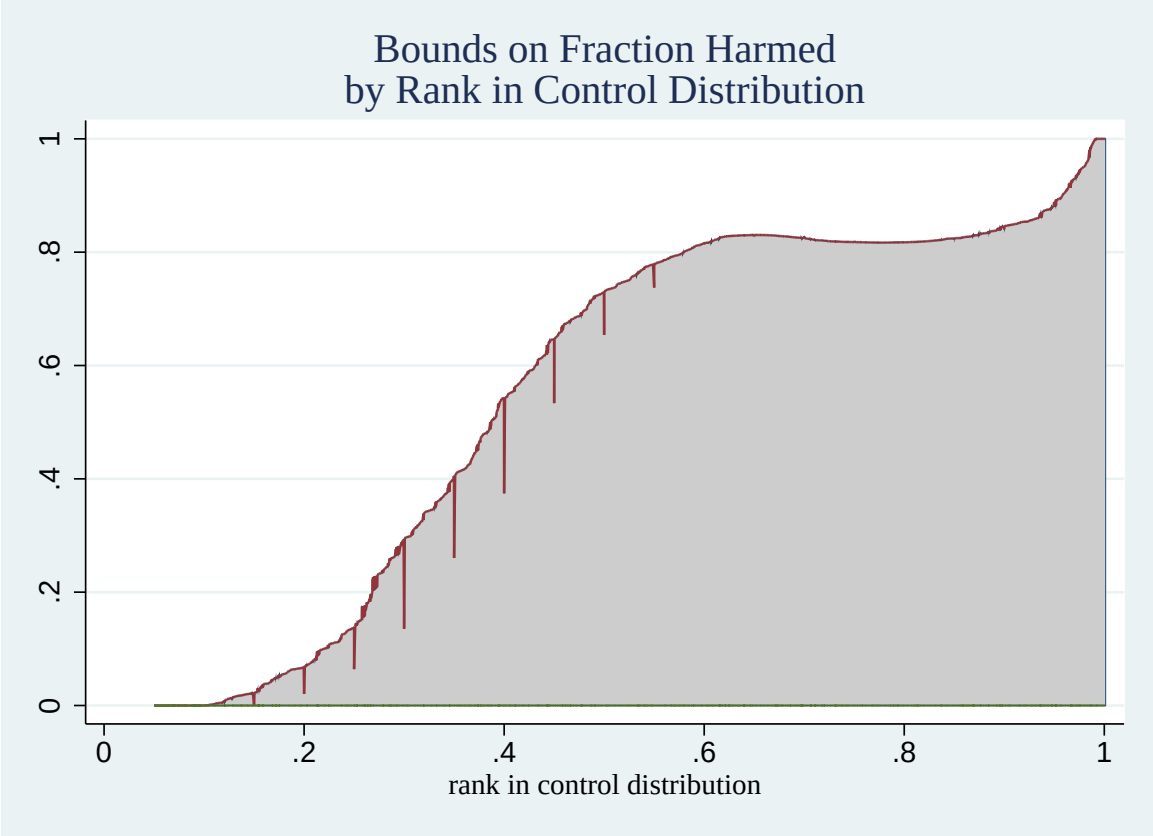

Figure 15: Estimated bounds on the probability of a negative individual-level effect of job training by rank in the control distribution.