Testing for Rank Invariance or Similarity in Program Evaluation[*]

Yingying Dong[†]
University of California, Irvine

Shu Shen[‡]
University of California, Davis

First version, February 2015; this version, October 2015

Abstract

This paper discusses testable implications of rank invariance or rank similarity, assumptions that are common in program evaluation and quantile treatment effect (QTE) models. We nonparametrically identify and test the counterfactual distribution of individual potential ranks, or features of the distribution. The tests allow treatment to be endogenous, while essentially not requiring any additional assumptions other than those used to identify and estimate QTEs. We focus on testing ranks in the unconditional distribution of potential outcomes, and briefly discuss testing for invariance or similarity of conditional ranks. The proposed tests are applied to the JTPA training program and Project STAR. For the former, we investigate whether job training causes individuals to systematically change their ranks in the earnings distribution; for the latter, we analyze the impacts of small classes and teacher aides on the gender performance gap in early childhood education. Also investigated is how these educational treatments interact with teacher experience in affecting students' relative performance.

JEL Codes: C12, C14, J31
Keywords: Treatment effect heterogeneity, Distributional treatment effect, Rank invariance, Rank similarity, Quantile treatment effect (QTE), IVQR, LQTE

1 Introduction

In the last decade or so, the program evaluation literature has increasingly sought to identify and estimate distributional effects. This trend reflects growing awareness that program impacts can be heterogeneous, and many interesting questions regarding the political economy of any program require knowledge of the distribution. Recent examples include Heckman, Smith, and Clements (1997), Bitler, Gelbach, and Hoynes (2006, 2008), Dammert (2008),

[*]The authors would like to thank Alberto Abadie for providing the data.
[†]Email: yyd@uci.edu
[‡]Email: shushen@ucdavis.edu

Djebbari and Smith (2008), Eren and Ozbeklik (2014), and Bitler and Hoynes (2014). These studies show that there can be substantial heterogeneity in the program effects and that a focus on mean effects may mask meaningful, and policy-relevant, heterogenous impacts across the outcome distribution.

The distributional effects of a program may be of interest in their own right. Empirical researchers have long recognized that without rank preservation, the distributional effects are not equivalent to the distribution of program impacts. Rank preservation requires that an individual's potential rank remains the same regardless of being treated or not. This is also known as rank invariance in the quantile treatment effect (QTE) literature. Without such an assumption, it is possible that the differences of the potential outcome distributions at any quantiles are zero, but due to individuals moving up and down in the distribution, the true treatment effects are not zero.

Knowledge of the distribution of program impacts can be important from the policy perspective. Heckman, Smith, and Clements (1997) note that undesirable distributional aspects of programs cannot always be offset by transfers governed by a social welfare function. Some outputs of programs (such as test scores, other forms of human capital or non-transferable payments in kind) simply cannot be valued, summed, and then redistributed. Consequently, one may wish to know whether a welfare program helps some individuals as well as that it does not hurt others. To gain insight into the distribution of program impacts for a training program, Heckman, Smith, and Clements (1997) explore the rank preservation assumption and find that it is not plausible.

Rank preservation or rank invariance requires an individual's rank in the potential outcome distribution to be the same across treatment states (Doksum, 1974; Lehmann, 1974), while rank similarity only requires that an individual's potential rank has the same probability distribution (Chernozhukov and Hansen, 2005). In various QTE models, rank invariance or the less restrictive condition rank similarity is required either for the identification or the interpretation of the identified treatment effects. There is a growing literature on identification and estimation of QTEs since the pioneering work of Koenker and Bassett (1978)[1]. For example, Chernozhukov and Hansen (2005, 2006, 2008) instrumental variable quantile regression (IVQR) model restricts the evolution of individual ranks across treatment states and thereby identifies QTEs for the whole population. Rank invariance or less restrictively rank similarity is one of the key identifying assumptions. Similarly, the nonparametric IV quantile regression models of Chernozhukov, Imbens, and Newey (2007) and Horowitz and Lee (2007) implicitly impose rank invariance by imposing a scalar disturbance. In contrast, the LQTE framework of Abadie, Angrist, and Imbens (2002) permits essential heterogeneity in treatment effects by not restricting how individuals' ranks change across treatment states and, therefore, identifies QTEs only among compliers. Frolich and Melly (2013) adopt the LQTE framework to identify unconditional QTEs while allowing for covariates. Rank invariance is not required for identification in the LQTE framework but is still required for the interpretation of the identified QTEs as individual causal effects. The same holds true for other studies that identify QTEs as horizontal differences of the marginal distributions of potential outcomes in different treatment states (see, e.g., Firpo, 2007 or Imbens and Newey, 2009). Comparison between the LQTE framework and the IVQR model can be found in Wüthrich (2014).

In light of their empirical and theoretical importance, this paper discusses testable implications of rank invariance or rank similarity and proposes nonparametric tests for rank similarity. The same test also works for rank invariance. The discussion focuses on individuals' ranks in the unconditional distribution of potential outcomes, which are arguably more policy relevant. For example, welfare of the unconditionally poor or performance of

[1]See, for recent examples, Chesher (2003, 2005) for identification of quantile effects in nonseparable models with endogeneity; Firpo (2007) for estimation of unconditional QTEs under the unconfoundedness assumption; Frolich and Melly (2013), for estimation of unconditional LQTEs, Firpo, Fortin, and Lemieux (2007) for identification of the effect of a marginal change in an exogenous explanatory variable on the unconditional quantiles of an outcome; Rothe (2010) and Imbens and Newey (2009) for identification and estimation of unconditional QTEs of a continuous endogenous treatment; and Powell (2014) for estimation of unconditional QTEs in a general framework.

the unconditionally lower-ranking students attracts more public attention than, say, welfare of highly educated people with relatively low wages or performance of relatively low-ranking students from high-income families. Heckman, Smith, and Clements (1997), Bitler, Gelbach, and Hoynes (2006, 2008), Bitler and Hoynes (2014), and Eren and Ozbeklik (2014) are some of the recent empirical studies that look at the unconditional QTEs of active labor market programs. Recent theoretical papers focusing on identification and estimation of unconditional QTEs include Firpo (2007), Firpo, Fortin, and Lemieux (2007), Rothe (2010), Imbens and Newey (2009), and Frolich and Melly (2013). Testing for unconditional distributional effects is discussed in Abadie (2002). In addition to the benchmark tests on unconditional ranks, we also discuss testing invariance or similarity of conditional ranks. Such tests can be useful when one is interested in sub-group distributional effects.

At first, one may think that testing for rank invariance or similarity is not feasible since the same individual cannot be observed under treatment and also under no treatment. Let Y_1 and Y_0 be the potential outcome under treatment and no treatment, respectively. It is well known that without functional form restrictions, one cannot identify the joint distribution of Y_0 and Y_1. This makes a direct test for rank invariance or similarity seemingly impossible. Bitler, Gelbach, and Hoynes (2006, 2008) investigate plausibility of rank preservation in their contexts by directly comparing covariate means at similar quantiles of the outcome distribution between the treatment and the control group, given that treatment is exogenous. See also Dammert (2008) for the same strategy. Empirical researchers focus on rank preservation to justify the distributional effects as individual causal effects. Less well known is that a weaker condition, rank similarity, can also guarantee that the distribution of covariates is identical at the same quantile of the potential outcome distributions and hence justify a causal interpretation of the quantile difference.

Intuitively, rank similarity means that individuals with the same characteristics should have the same probability distribution of potential ranks with or without treatment. We therefore nonparametrically identify and test the counterfactual distribution of potential ranks (or features of the distribution) among observationally equivalent individuals. That is, our tests draw on implications of the conditional distribution of Y_1 conditional on observable covariates as well as that of Y_0 conditional on observable covariates. This is similar to Bitler, Gelbach, and Hoynes (2006, 2008). We address endogeneity and summarize the implied information into test statistics with well-behaved asymptotic distributions. In an independent and contemporaneous work, Frandsen and Lefgren (2015) also leverage observable covariates and propose a rank similarity test. They focus on a parametric test for the equality of mean ranks under treatment and no treatment. In contrast, this paper focuses on nonparametric identification and testing of the entire distribution of potential ranks.

Except for mild regularity conditions, the proposed tests do not require any additional assumptions other than those used to identify and estimate QTEs (or LQTEs). The tests allow treatment to be endogenous. Covariates are permitted in estimating the unconditional QTEs, so the tests can handle instrumental variables regardless of whether they are valid conditional on covariates or not. Note that although the frequently used constant treatment effect models imply rank invariance, our tests do more than testing for constant treatment effects. Rank invariance or rank similarity can hold even when treatment effects are fully heterogenous. An obvious example of this is when Y_1 is weakly monotonically increasing in Y_0. Note also that we do not view the proposed tests as tests for the identifying assumption of the Chernozhukov and Hansen (2005, 2006, 2008) IVQR model or the nonparametric IV quantile regression model of Chernozhukov, Imbens, and Newey (2007) and Horowitz and Lee (2007), since these models impose rank invariance or similarity only after conditioning on all relevant covariates.[2]

Monte Carlo simulations show that the proposed tests have good size and power prop-

[2]Similarity or invariance of conditional ranks is more plausible when conditioning on a rich set of covariates, as discussed in Chernozhukov and Hansen (2006). The proposed tests utilize the predictive power of observable covariates for potential ranks. The tests can be useful in testing conditional ranks only when not all covariates are included in the conditioning set of the conditional QTEs.

erties in finite samples. The tests are applied to two empirical applications. One focuses on evaluating the large publically funded training program under the Job Training Partnership Act (JTPA), and the other on analyzing a large-scale randomized experiment in education, Project STAR. In the first empirical application, we show that rank similarity can be strongly rejected for both male and female trainees. In contrast, falsification tests using age as the outcome variable fail to reject rank similarity for either group. Further, while male trainees systematically change their ranks throughout the earnings distribution, for female trainees, ranks change mainly at the lower tail of the earnings distribution. Overall, the evidence suggests that the impacts of the JTPA training program are more complicated than what would be suggested by standard QTE estimates. One should, therefore, be cautious in equating the distributional effects of training with the effects of training on individual trainees.

In the second empirical application, we find that attending a small class or having a teacher aide is beneficial for boys and consequently narrows the gender gap in test scores. We also find that small classes substantially narrow the performance gaps among students taught by teachers with different levels of experience. The largest improvement is observed for students assigned to teachers with 6-10 years' experience. On the other hand, contrary to what one might believe, assigning a teacher aide to a 'greed-handed' teacher is relatively inefficient and negatively affects students' ranks.

The rest of the paper is organized as follows. Section 2 discusses testable implications of rank invariance and rank similarity. Section 3 provides identification of the counterfactual distribution of potential ranks and further the implied conditional moment restrictions implied by rank similarity. Section 4 discusses the primary test statistic along with its asymptotic distribution. Section 5 presents Monte Carlo simulations. Section 6 presents the empirical applications. Section 7 discusses various extensions. Concluding remarks are provided in Section 8.

2 Testable Implications of Rank Invariance or Similarity

We first define rank invariance and rank similarity, and then discuss their testable implications. We focus on unconditional ranks, or individuals' ranking in the unconditional distributions of potential outcomes. Discussion of conditional rank invariance or similarity is deferred to Section 7.

Let T be a binary treatment indicator that equals one if an individual is treated and zero if not. Given potential outcomes Y_1 and Y_0, the observed outcome is then $Y = Y_0(1 - T) + Y_1 T$. We use $F_t(.) : \mathbb{R} \to [0,1]$ and $q_t(.) : [0,1] \to \mathbb{R}$ to denote the unconditional cumulative distribution function and unconditional quantile function of Y_t for $t = 0, 1$. Following Doksum (1974) and Lehmann (1974), unconditional QTEs are defined in this paper as horizontal differences between the marginal distributions of potential outcomes. Specifically, $QTE(\tau) = q_1(\tau) - q_0(\tau)$, for all $\tau \in [0,1]$.

Let $U_t = F_t(Y_t)$ be potential ranks. $U_t \sim U(0,1)$ for $t = 0, 1$ by construction. Unless stated otherwise, this paper uses rank invariance as well as rank similarity to refer to conditions imposed on unconditional ranks.

Definition 1. *Rank invariance is the condition* $U_0 = U_1$*.*

Under rank invariance, U_0 and U_1 are the same random variable, so an individual's potential rank with or without treatment remains exactly the same. In practice, we never observe both U_0 and U_1 for the same individual, so we do not actually know whether the same individual remains at the same rank or not across treatment states. Let $\mathbf{X}$ be a vector of observables and V be unobservables. V may be be a vector. The dimension of V does not matter here. Suppose that for $t = 0, 1$, $Y_t = g_t(\mathbf{X}, V) : \mathcal{W} \to \mathbb{R}$, where $\mathcal{W}$ is the support of $(\mathbf{X}, V)$. For example, Y_t, $t = 1, 0$ could be potential earnings with or without training, $\mathbf{X}$ could be education and demographic characteristics, and V could be ability. $U_t = F_t(g_t(\mathbf{X}, V))$ is then a function that maps from $\mathcal{W}$ to $[0,1]$ and is deterministic given

$\mathbf{X}$ and V. Rank invariance holds if and only if $U_0 \mid (\mathbf{X} = \mathbf{x}, V = v) = U_1 \mid (\mathbf{X} = \mathbf{x}, V = v)$ for all $(\mathbf{x}, v) \in \mathcal{W}$. Immediately rank invariance implies $U_0 \mid (\mathbf{X} = \mathbf{x}) \sim U_1 \mid (\mathbf{X} = \mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ is the support of $\mathbf{X}$.

Rank invariance may be restrictive in practice. Consider the following thought experiment. A test is given to a random sample of students and a cloned sample that consists of the same students. The outcome of interest is the test score. The treatment is simply the cloning indicator, so treatment effects are (supposed to be) zero for everyone. However, due to random chance or luck, a student and her clone may not have the same test score or same rank in the test score distribution. Nevertheless, if we repeat this experiment infinitely many times, the student and her clone will have the same distribution of ranks and hence any features (e.g. mean, median) of their rank distributions will be the same.[3]

Rank similarity relaxes rank invariance by allowing for random deviations, or "slippages" in one's rank away from some common level, so that the exact rank for an individual may not be the same in different treatment states (Chernozhukov and Hansen, 2005). Assume that for $t = 0, 1$, $Y_t \mid \mathbf{X} = \mathbf{x}, V = v$ and hence $U_t \mid \mathbf{X} = \mathbf{x}, V = v$ is not deterministic as in the case of rank invariance. In particular, let $Y_t = g_t(\mathbf{X}, V, S_t)$, where $\mathbf{X}$ and V are observables (education or demographic characteristics) and unobservables (ability) that determine the common rank level of an individual, and S_t (e.g. luck) is an idiosyncratic shock. S_t is responsible for random slippages from the common rank level in the treatment state t. Note that $Y_t = g_t(\mathbf{X}, V, S_t)$ is a representation rather than a real restriction. The difference between S_t and V is that unlike V, S_t is realized only after the treatment is chosen. If one were to specify a treatment model, then V would enter the treatment model while S_t would not. Similar to Chernozhukov and Hansen (2005), we also "...implicitly make the assumption that one selects the treatment without knowing the exact potential outcomes..." Rank similarity can then be analogously defined as follows.

[3]Note that luck is plausibly mutually i.i.d for each student and her clone.

Definition 2. *Rank similarity is the condition* $U_0 \mid (\mathbf{X}=\mathbf{x}, V=v) \sim U_1 \mid (\mathbf{X}=\mathbf{x}, V=v)$ *for any* $(\mathbf{x}, v) \in \mathcal{W}$.

Instead of requiring that U_0 and U_1 are the same random variable, rank similarity just assumes that the conditional distributions of U_0 and U_1, conditional on $\mathbf{X}$ and V, are the same. Chernozhukov and Hansen (2005) consider a weaker condition, conditional rank similarity, which assumes that conditional ranks conditional on observables $\mathbf{X}$ are identically distributed across treatment status conditional on $\mathbf{X}$ and unobservables (V here).[4]

Rank invariance is a special case of rank similarity, where S_t is a null set of random variables and the distribution of $U_t \mid \mathbf{X}, V$ is degenerate. Rank invariance implies that the QTE at the τ quantile is the individual treatment effect for anyone who is at quantile τ. It is worth clarifying what rank similarity implies. The following discussion focuses on rank similarity. All the conclusions hold trivially for rank invariance.

Lemma 1.

1. *Given rank similarity,* $F_{\mathbf{X},V|U_0}(\mathbf{x}, v|\tau) = F_{\mathbf{X},V|U_1}(\mathbf{x}, v|\tau)$, *for all* $\tau \in (0,1)$ *and* $(\mathbf{x}, v) \in \mathcal{W}$.

2. *Given rank similarity,* $E[Y_1 - Y_0|\mathbf{X}=\mathbf{x}, V=v] = \int_0^1 QTE(\tau) dF_{U|\mathbf{X},V}(\tau|\mathbf{x}, v)$ *for all* $(\mathbf{x}, v) \in \mathcal{W}$, *where* $QTE(\tau) \equiv q_1(\tau) - q_0(\tau)$ *and* $F_{U|\mathbf{X},V}(.|\mathbf{x}, v) \equiv F_{U_t|\mathbf{X},V}(.|\mathbf{x}, v)$ *for* $t = 0, 1$ *and all* $(\mathbf{x}, v) \in \mathcal{W}$.

3. *(Main Testable Implication) Given rank similarity,* $F_{U_0|\mathbf{X}}(\tau|\mathbf{x}) = F_{U_1|\mathbf{X}}(\tau|\mathbf{x})$ *for all* $\tau \in (0,1)$ *and* $\mathbf{x} \in \mathcal{X}$.

Lemma 1 summarizes several implications of rank similarity. The proof is given in the appendix. The first part of the lemma says that given rank similarity, the distribution of

[4]Unconditional rank similarity implies conditional rank similarity. To see the difference between the two more clearly, by the Skorohod representation of potential outcomes, one can write $Y_t = q(t, U_t)$, where $q(t, U_t)$ is the quantile function of Y_t, e.g., $q(t, \tau)$ is τ unconditional quantile of Y_t. U_t is then the unconditional potential rank when $T = t$. In contrast, Chernozhukov and Hansen (2005, 2006, 2008) define the conditional rank using $Y_t \equiv q(t, \mathbf{x}, \widetilde{U}_t)$, so $\widetilde{U}_t$, a function of V and S_t, represents the conditional rank conditional on $\mathbf{X} = \mathbf{x}$. $\widetilde{U}_t$ is responsible for the heterogeneity of outcomes among individuals with the same observed characteristics $\mathbf{X} = x$ and treatment state t. Their rank invariance then requires $\widetilde{U}_0 = \widetilde{U}_1$ conditional on $\mathbf{X}$, while their rank similarity requires $\widetilde{U}_0 \sim \widetilde{U}_1$ conditional on $\mathbf{X}$ and the treatment model unobservables V.

observable and unobservable covariates is the same at the same quantile of the potential outcome distribution with or without treatment. The statement follows from the definition of rank similarity and the Bayes' rule. This result implies that comparing the same quantile of the marginal distributions of Y_1 and Y_0 is a comparison on average made for the same underlying individuals characterized by $\mathbf{X}$ and V.

The second part of the lemma states that for any individual defined by observables $\mathbf{X} = \mathbf{x}$ and unobservables $V = v$, her average treatment effect is a weighted average of the unconditional QTEs, where the weights are the individual's probabilities of being at different quantiles. It is known that under rank similarity, one loses the ability to point identify treatment effects for particular individuals, or $Y_1 - Y_0|\mathbf{X} = \mathbf{x}, V = v$ (see, e.g., Imbens and Newey, 2009). This result of Lemma 1.2 implies that under rank similarity, one may instead identify individual expected treatment effects, providing that one can identify $F_{U|\mathbf{X},V}(.|.)$. In practice, given rank similarity, individual expected treatment effects, rather than exact treatment effects, may be of greater policy interest, since the random shock S_t, luck or any counterpart of it, is not manipulable by nature.

Identification of individual expected treatment effects relies on researchers' ability to identify the QTEs and the individual's probability distribution of ranks. The next section provides identification of the potential rank distribution among observationally equivalent individuals, under essentially the same conditions as those used for identifying QTEs (or LQTEs). Further, if unobservables do not play a role in determining potential ranks, i.e. $F_{U_t|\mathbf{X},V}(.|\mathbf{x}, v) = F_{U_t|\mathbf{X}}(.|\mathbf{x}) \equiv F_{U_t|\mathbf{X}}(\tau|\mathbf{x})$ for $t = 0, 1$, then one can point identify an individual's expected treatment effect under rank similarity as $E[Y_1 - Y_0|\mathbf{X} = \mathbf{x}, V = v] = \int_0^1 (q_1(\tau) - q_0(\tau))\, dF_{U|\mathbf{X}}(\tau|\mathbf{x})$. Otherwise, bounds might be constructed, depending on assumptions on unobservables.

The last part of Lemma 1 is what we focus on. It follows immediately from the definition of rank similarity. It states that the distribution of potential ranks among those with $\mathbf{X} = \mathbf{x}$

are the same across treatment states. In other words, under rank similarity, treatment should not affect the distribution of ranks for observationally equivalent individuals.

Note that this testable implication of rank similarity in Lemma 1 only considers observables. It is a necessary but not sufficient condition for rank similarity. Below we provide an assumption under which the testable implication is also a sufficient condition for rank similarity.

Assumption 1. $F_{V|\mathbf{X},U_0}(v|\mathbf{x},\tau) = F_{V|\mathbf{X},U_1}(v|\mathbf{x},\tau)$ *for all* $\tau \in (0,1)$ *and* $\mathbf{x} \in \mathcal{X}$.

Assumption 1 assumes that conditional on observables $\mathbf{X}$, the distribution of unobservables is the same at the same rank of the potential outcome distribution. Alternatively, the assumption states that once the distribution of the observables is the same at the same potential rank, the distribution of unobservables will also be the same. Note that if unobservables do not play a role in determining potential ranks, i.e., $F_{U_t|\mathbf{X},V}(\tau|\mathbf{x},v) = F_{U_t|\mathbf{X}}(\tau|\mathbf{x})$ for $t = 0, 1$, it follows straightforwardly that rank similarity holds if and only if $F_{U_0|\mathbf{X}}(\tau|\mathbf{x}) = F_{U_1|\mathbf{X}}(\tau|\mathbf{x})$. However, Assumption 1 does not assume away unobservables, i.e., Assumption $F_{V|\mathbf{X},U_0}(.|\mathbf{x},\tau) = F_{V|\mathbf{X},U_1}(.|\mathbf{x},\tau)$ for all $\tau \in (0, 1)$ and $\mathbf{x} \in \mathcal{X}$ does not imply $F_{U_t|\mathbf{X},V}(.|\mathbf{x},v) = F_{U_t|\mathbf{X}}(.|\mathbf{x})$ for both $t = 0, 1$ and $(\mathbf{x}, v) \in \mathcal{W}$.

Assumption 1 resembles in spirit the unconfoundedness assumption that is popular in program evaluation, e.g., in various matching estimators (Rubin, 1990). Assumption 1 is essentially not testable, just like unconfoundedness.

Lemma 2. *Given Assumption 1, rank similarity holds if and only if* $F_{U_0|\mathbf{X}}(\tau|\mathbf{x}) = F_{U_1|\mathbf{X}}(\tau|\mathbf{x})$ *for all* $\tau \in (0, 1)$ *and* $\mathbf{x} \in \mathcal{X}$.

Given Assumption 1, for all $\tau \in (0, 1)$ and $\mathbf{x} \in \mathcal{X}$, $F_{\mathbf{X},V|U_0}(\mathbf{x},v|\tau) = F_{\mathbf{X},V|U_1}(\mathbf{x},v|\tau)$ if and only if $F_{\mathbf{X}|U_0}(\mathbf{x}|\tau) = F_{\mathbf{X}|U_1}(\tau|\mathbf{x})$. Then, by Bayes' rule $F_{U_0|\mathbf{X},V}(\tau|\mathbf{x}, v) = F_{U_1|\mathbf{X},V}(\tau|\mathbf{x},v)$ if and only if $F_{U_0|\mathbf{X}}(\tau|\mathbf{x}) = F_{U_1|\mathbf{X}}(\tau|\mathbf{x})$. It follows that $F_{U_0|\mathbf{X}}(\tau|\mathbf{x}) = F_{U_1|\mathbf{X}}(\tau|\mathbf{x})$ for all $\tau \in (0, 1)$ and $\mathbf{x} \in \mathcal{X}$ is not only a necessary but also a sufficient condition for rank similarity.

Note that without Assumption 1, a test based on the testable implication in Lemma 1 can only detect whether observationally equivalent individuals have the same distribution of potential ranks under treatment and no treatment. Further if identification is only feasible for compliers or one is only interested in compliers, Assumption 1 can be made conditional on compliers and hence we have that Corollary 2 holds among compliers.

3 Identification

For any $\tau \in (0,1)$ and $\mathbf{x} \in \mathcal{X}$, $F_{U_1|\mathbf{X}}(\tau|\mathbf{x}) - F_{U_0|\mathbf{X}}(\tau|\mathbf{x})$ is a measure of the difference between the conditional distributions of potential ranks under treatment or no treatment. Intuitively, this shows how the probability of staying at the same rank changes with treatment among those with $\mathbf{X} = \mathbf{x}$. This section discusses identification of $F_{U_t|\mathbf{X}}(\tau|\mathbf{x})$ for $t = 0, 1$ and hence $F_{U_1|\mathbf{X}}(\tau|\mathbf{x}) - F_{U_0|\mathbf{X}}(\tau|\mathbf{x})$, which serves as the basis for the proposed tests.

3.1 Exogenous Treatment

If T is exogenous, as in randomized experiments with perfect compliance, identification of $F_{U_t|\mathbf{X}}(\tau|\mathbf{x})$ for $t = 0, 1$ is trivial. In this case for any $\tau \in [0,1]$ and $\mathbf{x} \in \mathcal{X}$, we have

$$
\begin{aligned}
& F_{U_t|\mathbf{X}}(\tau|\mathbf{x}) = E\left[\mathbf{1}(U_t \leq \tau)|\mathbf{X} = \mathbf{x}\right] = E\left[\mathbf{1}(Y_t \leq q_t(\tau)|\mathbf{X} = \mathbf{x}\right] \\
= \quad & E\left[\mathbf{1}(Y \leq q_t(\tau))|\mathbf{X} = \mathbf{x}, T = t\right],
\end{aligned}
$$

where $q_t(\tau)$ for $t = 0, 1$ is directly identified from the sub-samples with $T = t$. Our emprical application of Project STAR illustrates this case.

3.2 Endogenous Treatment

When T is endogenous, a valid IV is required to identify QTEs and further the impact of treatment T on the distribution of ranks given $\mathbf{X}$. We adopt the LQTE framework. The LQTE identifying assumption is particularly suitable for our empirical application using the training data from the JTPA program (see discussion in Abadie, Angrist and Imbens 2002).

The LQTE framework permits essential heterogeneity in treatment effects and identifies distributional effects for compliers only, which are the largest sub-population for which QTEs can be point identified. Testing for rank similarity is then only relevant to compliers. However, if assumptions are made to identify unconditional QTEs for the whole population, one could analogously test for rank similarity for the whole population. For example, assuming unconfoundedness, one could adopt the estimator in Firpo (2007) to estimate unconditional QTEs for the whole population and then test for rank similarity for the whole population. As is clear from Theorem 1 and its proof, essentially what is required is a valid IV (or the unconfoundedness assumption) that allows the identification of unconditional QTEs in the first stage and then the effect of treatment on the distribution of ranks in the second stage.

Let Z be an IV for the endogenous treatment T. For simplicity, assume that the instrument Z takes on two values, 0 and 1, although identification does not rely on a binary IV. Further, let T_z be the potential treatment status if $Z = z$. The observed treatment status can then be written as $T = T_0(1 - Z) + T_1 Z$. Define compliers as individuals with $T_1 > T_0$ (Angrist, Imbens, and Rubin, 1996), and let C denote the set of compliers. Define the distribution function of the potential outcome among compliers as

$$F_{t|C}(y) = \Pr[Y_t \leq y | T_1 > T_0] \text{ for } t = 0, 1.$$

We are interested in testing for rank similarity among compliers. For notational convenience, unless stated otherwise we use U_t to refer to potential ranks among compliers only, i.e., $U_t \equiv U_{t|C} = F_{t|C}(Y_t)$ for $t = 0, 1$. Let $\mathcal{X}_C = \{\mathbf{x} \in \mathcal{X} : \Pr[T_1 > T_0 | \mathbf{X} = \mathbf{x}] > 0\}$. Similar to the last part of Lemma 1, rank similarity among compliers implies

$$F_{U_1|C,\mathbf{X}}(.|\mathbf{x}) = F_{U_0|C,\mathbf{X}}(.|\mathbf{x}) \text{ for all } \mathbf{x} \in \mathcal{X}_C.$$

Note that $F_{U_t|C,\mathbf{X}}(.|\mathbf{x})$ for $t = 0, 1$ is only defined on $\mathbf{x} \in \mathcal{X}_C$, and that values $\mathbf{x} \in \mathcal{X}_C$

represent compliers' observable characteristics, even though individuals with $\mathbf{X} = \mathbf{x} \in \mathcal{X}_C$ are not necessarily all compliers. The above condition states that the distribution of potential (complier) ranks remains the same across treatment states among compliers with the same observed characteristics. The following gives the identifying assumption for $F_{U_t|C,\mathbf{X}}(.|\mathbf{x})$ for both $t = 0, 1$ and all $\mathbf{x} \in \mathcal{X}_C$.

Assumption 2. *Let* (Y_t, T_z, X, Z), $t, z = 0, 1$ *be random variables mapped from the common probability space* $(\Omega, \mathsf{F}, \mathsf{P})$. *The following conditions hold jointly with probability one.*

1. *Independence:* $(Y_0, Y_1, T_0, T_1) \perp Z|\mathbf{X}$.

2. *First stage:* $E(T_1) \neq E(T_0)$.

3. *Monotonicity:* $\Pr(T_1 \geq T_0) = 1$.

4. *Nontrivial assignment:* $0 < \Pr(Z = 1|\mathbf{X} = \mathbf{x}) < 1$ *for all* $\mathbf{x} \in \mathcal{X}$.

Assumption 2 is the standard LQTE identifying assumption used in Abadie, Angrist, and Imbens (2002) and Abadie (2003), except that we allow for a weaker first stage. In particular we require $E(T_1) \neq E(T_0)$ to hold without conditioning on $\mathbf{X}$, so compliers do not have to exist at every value of $\mathbf{X}$. This is because we test whether rank similarity is violated at any values of $\mathbf{X}$ and because we identify and estimate the unconditional QTE instead of conditional QTE (for the latter point, see the discussion in Frolich and Melly 2013). Assumption 2.1 subsumes two related requirements: exclusion restriction and IV independence of the first stage error (Angrist, Imbens, and Rubin, 1996). Assumption 2.3 rules out defiers, which can be weakened by the assumption that there are conditionally more compliers than defiers (see, e.g., de Chaisemartin, 2014). Assumption 2.4 is also known as a common support assumption requiring $Supp(\mathbf{X}|Z = 0) = Supp(\mathbf{X}|Z = 1)$.[5]

[5]Vytlacil (2002) establishes equivalence of these independence and monotonicity assumptions to a latent index threshold-crossing model.

When Z is a random assignment of the treatment, as in our empirical application, the independence restriction is valid without conditioning on covariates $\mathbf{X}$; however, including covariates can remove any chance association between T and $\mathbf{X}$ or improve efficiency (Frolich and Melly, 2013). Let $q_{t|C}(\tau)$ for $t = 0, 1$ be the τ quantile of Y_t distribution among compliers.

Theorem 1. *Let* $I(\tau) \equiv \mathbf{1}\left(Y \leq \left(Tq_{1|C}(\tau) + (1-T)\, q_{0|C}(\tau)\right)\right)$*. Given Assumption 2, for all* $\tau \in (0,1)$*,* $\mathbf{x} \in \mathcal{X}_C$*, and* $t = 0, 1$*,* $F_{U_t|C,\mathbf{X}}(\tau|\mathbf{x})$ *is identified and is given by*

$$F_{U_t|C,\mathbf{X}}(\tau|\mathbf{x}) = \frac{E\left[I(\tau)\mathbf{1}\left(T=t\right)|Z=1,\mathbf{X}=\mathbf{x}\right] - E\left[I(\tau)\mathbf{1}\left(T=t\right)|Z=0,\mathbf{X}=\mathbf{x}\right]}{E[\mathbf{1}\left(T=t\right)|Z=1,\mathbf{X}=\mathbf{x}] - E[\mathbf{1}\left(T=t\right)|Z=0,\mathbf{X}=\mathbf{x}]}. \quad (1)$$

$F_{U_1|C,\mathbf{X}}(.|\mathbf{x}) = F_{U_0|C,\mathbf{X}}(.|\mathbf{x})$ *for* $\mathbf{x} \in \mathcal{X}_C$ *if and only if for all* $\tau \in (0,1)$ *and* $\mathbf{x} \in \mathcal{X}$

$$E\left[I(\tau)|Z=1,\mathbf{X}=\mathbf{x}\right] = E\left[I(\tau)|Z=0,\mathbf{X}=\mathbf{x}\right]. \quad (2)$$

$I(\tau)$ is referred to as a rank indicator. T is binary, so $I(\tau)$ can also be written as $I(\tau) = \mathbf{1}\left(Y \leq q_{1|C}(\tau)\right)T + \mathbf{1}\left(Y \leq q_{0|C}(\tau)\right)(1-T)$. Equation (2) in Theorem 1 states that given rank similarity, the instrument Z has no impact on the distribution of ranks conditional on $\mathbf{X} = \mathbf{x}$. Equation (2) is a necessary condition for rank similarity. Further if Assumption 1 holds, it will be a necessary and sufficient condition for rank similarity.

Note that although the conditional distribution $F_{U_t|C,\mathbf{X}}(.|\mathbf{x})$ is only defined for $\mathbf{x} \in \mathcal{X}_C$, under rank similarity equation (2) holds for all values of $\mathbf{x} \in \mathcal{X}$, since it holds trivially for any $\mathbf{x} \in \mathcal{X}/\mathcal{X}_C$. Theorem 1 nests exogenous treatment as a special case, as in this case, $Z = T$ and everyone is a complier.

$q_{t|C}(\tau)$ for $t = 0, 1$ is identified following Frolich and Melly (2013). In particular, given Assumption 2, they can be identified as the inverse function of $F_{t|C}(y)$, where $F_{t|C}(y)$ is given by

$$F_{t|C}(y) = \frac{\int_{\mathcal{X}} E\left(\mathbf{1}(Y \leq y)\mathbf{1}(T=t)|Z=1,\mathbf{X}\right) - E\left(\mathbf{1}(Y \leq y)(T=t)|Z=0,\mathbf{X}\right) dF_{\mathbf{X}}}{\int_{\mathcal{X}} E\left(T=t|Z=1,\mathbf{X}\right) - E\left(T=t|Z=0,\mathbf{X}\right) dF_{\mathbf{X}}}.$$

Alternatively, they can be identified by minimizing a weighted check function

$$\left(q_{0|C}(\tau), q_{1|C}(\tau)\right) = \arg\min_{q_0,q_1} E\left[\rho_\tau(Y - q_0(1-T) - q_1 T)\omega\right],$$

where $\rho_\tau(u) = u\left(\tau - \mathbf{1}(u < 0)\right)$ is the standard check function, $\omega = \left(\frac{Z}{\pi(\mathbf{X})} - \frac{1-Z}{1-\pi(\mathbf{X})}\right)(2T-1)$ and $\pi\left(\mathbf{X}\right) = \Pr\left(Z = 1|\mathbf{X}\right)$ is the instrument probability.

Theorem 1 suggests that one can test for rank similarity by a two-step procedure: first estimate the unconditional quantiles $q_{0|C}(\tau)$ and $q_{1|C}(\tau)$, and then test whether Equation (2) in Theorem 1 holds for all $\tau \in (0,1)$ and $\mathbf{x} \in \mathcal{X}$, replacing $q_{0|C}(\tau)$ and $q_{1|C}(\tau)$ with their estimates. If desired, one can also test a particular quantile or a subset of quantiles of interest. Features of the potential rank distribution, such as the median rank or the mean rank can also be tested. Similar to Abadie, Angrist, and Imbens (2002) in testing for distributional treatment effects in IV models, we test the reduced form Equation (2) rather than Equation (1). Upon rejection of rank similarity, researchers can further utilize Equation (1) to quantify the degree of violation in rank similarity at different values of τ and $\mathbf{x}$.

In practice, one needs the covariates $\mathbf{X}$ to be non-trivial, i.e., $F_{t|C,\mathbf{X}}\left(\mathbf{x}\right) = F_{t|C}$ does not hold for both $t = 0, 1$ and all $\mathbf{x} \in \mathcal{X}_C$, in order for a test based on Equation (2) to have any power. If $F_{t|C,\mathbf{X}}\left(\mathbf{x}\right) = F_{t|C}$ for both $t = 0, 1$ and all $\mathbf{x} \in \mathcal{X}_C$, then Equation (2) would always hold by construction.[6]

3.3 Testing the Moments of Potential Ranks

Instead of testing whether the distribution of potential ranks remains the same across treatment states, one may also test any functionals of the potential rank distribution. Recall that rank similarity among compliers implies that for all $\mathbf{x} \in \mathcal{X}_C$, $F_{U_1|C,\mathbf{X}}(.|\mathbf{x}) = F_{U_0|C,\mathbf{X}}(.|\mathbf{x})$.

[6] That is because if $F_{t|C,\mathbf{X}}(q_{t|C}(\tau)|\mathbf{x}) = F_{t|C}(q_{t|C}(\tau))$, $F_{t|C,\mathbf{X}}(q_{t|C}(\tau)|\mathbf{x}) = \tau$. Then the testable implication of rank similarity holds by construction, and hence Equation (2) holds by construction.

This further implies that, for instance,

$$E[U_1^p|C, \mathbf{X} = \mathbf{x}] = E[U_0^p|C, \mathbf{X} = \mathbf{x}],$$

for all $\mathbf{x} \in \mathcal{X}_C$ and some $p > 0$. When $p = 1$, this equation represents equality of the mean potential ranks. One can similarly test any higher order raw moments.

Let $U \equiv TU_1 + (1-T)U_0 = \int_0^1 \mathbf{1}\left(\left(Tq_{1|C}(\tau) + (1-T)q_{0|C}(\tau)\right) < Y\right) d\tau = 1 - \int_0^1 I(\tau)d\tau$. U is identified because $I(\tau)$ is identified. Analogous to Theorem 1, one can test whether the instrument Z has an impact on the conditional moments of potential ranks conditional on $\mathbf{X}$, i.e., test whether

$$E\left[U^p|Z = 1, \mathbf{X} = \mathbf{x}\right] = E\left[U^p|Z = 0, \mathbf{X} = \mathbf{x}\right]$$

for some $p > 0$ holds for all $\mathbf{x} \in \mathcal{X}$. In addition, the change in the conditional moments of potential ranks for compliers with $\mathbf{X} = \mathbf{x} \in \mathcal{X}_C$ is identified by

$$E[U_1^p|C, \mathbf{X} = \mathbf{x}] - E[U_0^p|C, \mathbf{X} = \mathbf{x}] = \frac{E\left[U^p|Z = 1, \mathbf{X} = \mathbf{x}\right] - E\left[U^p|Z = 0, \mathbf{X} = \mathbf{x}\right]}{E\left[T|Z = 1, \mathbf{X} = \mathbf{x}\right] - E\left[T|Z = 0, \mathbf{X} = \mathbf{x}\right]}.$$

4 Testing

This section discusses the test statistic along with its asymptotic properties, given the identification results in the previous section. Treatment is endogenous here, with exogenous treatment following as a special case. $\mathbf{X}$ is assumed to be discrete with finite support. Continuous covariates are considered in Section 7. Discrete $\mathbf{X}$ with finite support is a reasonable assumption, since typically one has a limited number of covariates and one can always discretize covariates. Denote $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_J\}$ as the support of $\mathbf{X}$.

4.1 The Distributional Test for Rank Similarity

Null Hypothesis and Test Statistic

In this section we develop testing procedure to examine whether Equation (2) in Theorem 1 holds with a specific data application. Let $\Omega = \{\tau_1, \tau_2, ..., \tau_K\}$ be a set of unconditional quantiles of interest.[7] Recall that $I(\tau) = \mathbf{1}\left(Y \leq \left(Tq_{1|C}(\tau) + (1-T)\, q_{0|C}(\tau)\right)\right)$. For any $j = 1, ..., J$ and $k = 1, ..., K$, define for $z = 0, 1$

$$m_j^z(\tau_k) = E\left[I\left(\tau_k\right)|Z = z, \mathbf{X} = \mathbf{x}_j\right].$$

We are interested in the null hypothesis

$$H_0 : m_j^0(\tau_k) = m_j^1(\tau_k), \text{ for all } j = 1, ..., J-1 \text{ and } k = 1, ..., K.$$

Only $J-1$ values of X are included in the null hypothesis, since $\sum_{j=1}^{J} m_j^z(\tau) = 1$ for any $\tau \in (0,1)$ and $z = 0, 1$.

Let $\{Y_i, T_i, Z_i, \mathbf{X}_i\}_{i=1}^n$ be a sample of i.i.d. draws of size n from $(Y, T, Z, \mathbf{X})$, and $I_i(\tau) = \mathbf{1}\left(Y_i \leq \left(T_i q_{1|C}(\tau) + (1-T_i)\, q_{0|C}(\tau)\right)\right)$ be the rank indicator for individual i. If the unconditional quantiles $q_{0|C}(\tau_k)$ and $q_{1|C}(\tau_k)$ were known, the conditional expectation $m_j^z(\tau_k)$ can be estimated by the proportion of individuals with $I_i(\tau_k) = 1$ in the subs-ample with $Z_i = z$ and $\mathbf{X}_i = \mathbf{x}_j$. However, $q_{0|C}(\tau_k)$ and $q_{1|C}(\tau_k)$ are unknown, so we have to estimate them first. Here we adopt the $\sqrt{n}$-consistent estimator proposed in Frolich and Melly (2013). The estimator has the advantage of estimating unconditional quantiles but still allowing for

[7]The proposed test can be extended to test for rank similarity over a continuous range of quantiles. In that case, the limiting distribution in Theorem 2 becomes a centered Gaussian process. Then, a Kolmogorov-Smirnov-type L_∞-norm test could be constructed. We do not pursue this route here because the asymptotics become quite complicated and obtaining numerical critical values in such a case is computationally demanding. Also, standard empirical practice is just to consider a finite set of quantiles.

covariates. Denote the corresponding estimates as $\hat{q}_{0|C}(\tau_k)$ and $\hat{q}_{1|C}(\tau_k)$. We have

$$\left(\hat{q}_{0|C}(\tau_k), \hat{q}_{1|C}(\tau_k)\right) = \arg\min_{q_0,q_1} \frac{1}{n}\sum_{i=1}^{n} \rho_{\tau_k}(Y_i - q_0(1-T_i) - q_1 T_i)\hat{\omega}_i,$$

where $\hat{\omega}_i = \left(\frac{Z_i}{\widehat{\pi}(\mathbf{X}_i)} - \frac{1-Z_i}{1-\widehat{\pi}(\mathbf{X}_i)}\right)(2T_i - 1)$ and $\widehat{\pi}(\mathbf{x}_j)$ is a consistent estimator of the instrument probability $\pi(\mathbf{x}_j)$.[8] Given $\hat{q}_{0|C}(\tau_k)$ and $\hat{q}_{1|C}(\tau_k)$, the conditional expectation $m_j^z(\tau_k)$ for all $z = 0, 1$, $j = 1, ..., J-1$ and $k = 1, ..., K$ can be estimated by

$$\hat{m}_j^z(\tau_k) = \frac{1}{n_j^z} \sum_{Z_i = z, \mathbf{X}_i = \mathbf{x}_j} \mathbf{1}\left(Y_i \leq \left(T_i \hat{q}_{1|C}(\tau_k) + (1-T_i)\hat{q}_{0|C}(\tau_k)\right)\right),$$

where $n_j^z = \sum_{i=1}^n \mathbf{1}(Z_i = z, \mathbf{X}_i = \mathbf{x}_j)$.

For $j = 1, ..., J$, let $\hat{\mathsf{m}}_j^z = \left[\hat{m}_j^z(\tau_1)\ \hat{m}_j^z(\tau_2)\ \cdots\ \hat{m}_j^z(\tau_K)\right]'$ and $\mathsf{m}_j^z = \left[m_j^z(\tau_1)\ m_j^z(\tau_2)\ \cdots\ m_j^z(\tau_K)\right]'$ be $K \times 1$ vectors. Let $\hat{\mathbf{m}}^z = \left[(\hat{\mathsf{m}}_1^z)'\ (\hat{\mathsf{m}}_2^z)'\ \cdots\ (\hat{\mathsf{m}}_{J-1}^z)'\right]'$ and $\mathbf{m}^z = \left[(\mathsf{m}_1^z)'\ (\mathsf{m}_2^z)'\ \cdots\ (\mathsf{m}_{J-1}^z)'\right]'$ be $K(J-1) \times 1$ vectors. Let $\hat{\mathbf{V}}$ be a consistent estimator of the asymptotic variance-covariance matrix of $\sqrt{n}\left(\hat{\mathbf{m}}^1 - \hat{\mathbf{m}}^0 - (\mathbf{m}^1 - \mathbf{m}^0)\right)$. We propose to test the null hypothesis H_0 using a Wald-type test statistic

$$W \equiv n\left(\hat{\mathbf{m}}^1 - \hat{\mathbf{m}}^0\right)' \hat{\mathbf{V}}^{-1}\left(\hat{\mathbf{m}}^1 - \hat{\mathbf{m}}^0\right).$$

Asymptotics

To derive asymptotic properties of the nonparametric estimator $\hat{m}_j^z(\tau_k)$ for all $z = 0, 1$, $j = 1, ..., J-1$ and $k = 1, ..., K$ and those of the test statistic W, we make the following assumptions regarding the underlying distribution of the data.

Assumption 3. *1. i.i.d. data: the data $(Y_i, T_i, Z_i, \mathbf{X}_i)$ for $i = 1, ..., n$ is a random sample of size n from $(Y, T, Z, \mathbf{X})$.*

[8] The Stata command "ivqte" can be conveniently used to estimate $\hat{q}_{0|C}(\tau_k)$ and $\hat{q}_{1|C}(\tau_k)$. $\hat{\omega}_i$ in practice is replaced by projected weights projected onto Y and T to make sure that the weights are nonnegative.

2. *For all $\tau \in \Omega = \{\tau_1, \tau_2, ..., \tau_K\}$, the random variable Y_1 and Y_0 are continuously distributed with positive density in a neighborhood of $q_{0|C}(\tau)$ and $q_{1|C}(\tau)$ in the subpopulation of compliers.*

3. *For all $j = 1, ..., J$, $\hat{\pi}(\mathbf{x}_j)$ is consistent, or $\hat{\pi}(\mathbf{x}_j) \xrightarrow{p} \pi(\mathbf{x}_j)$.*

4. *Let $f_{Y|T,Z,\mathbf{X}}$ be the conditional density of Y given T, Z and $\mathbf{X}$. For all $t, z = 0, 1$, $j = 1, ..., J$ and $\tau \in \Omega$, $f_{Y|T,Z,\mathbf{X}}(y|t, z, \mathbf{x}_j)$ has a bounded first derivative with respect to y in a neighborhood of $q_{t|C}(\tau)$. Let $f_{Y|\mathbf{X}}(y|\mathbf{x})$ be the conditional density of Y given $\mathbf{X}$. For all $\tau \in \Omega$ and $j = 1, ..., J$, $f_{Y|\mathbf{X}}(.|\mathbf{x}_j)$ is positive and bounded in a neighborhood of $q_{t|C}(\tau)$.*

Assumptions 3.1-3.3 guarantee consistency of $\hat{q}_{0|C}(\tau_k)$ and $\hat{q}_{1|C}(\tau_k)$ for $k = 1, ..., K$. Assumption 3.4 ensures that the asymptotic variance-covariance matrix of $\hat{\mathbf{m}}^1 - \hat{\mathbf{m}}^0$ is bounded and has full rank. Let $f_{t|C}$ be the density of potential outcome Y_t among compliers, $p_{Z,\mathbf{X}}$ be the joint probability of Z and $\mathbf{X}$, $p_{T|Z,\mathbf{X}}(z, \mathbf{x}_j)$ be the probability of receiving treatment given instrument status, and $P_c = E[T|Z = 1] - E[T|Z = 0]$ be the proportion of compliers. For all $k = 1, ..., K$, $f_{t|C}\left(q_{t|C}(\tau_k)\right) > 0$ by Assumption 3.2. $p_{Z,X} > 0$ by Assumptions 2.4 and 3.3, and $P_c > 0$ by Assumptions 2.2 and 2.3. The following theorem discusses the asymptotic distribution of $\hat{\mathbf{m}}^1 - \hat{\mathbf{m}}^0$.

Theorem 2. *Given Assumptions 2 and 3,*

$$\sqrt{n}\left(\hat{\mathbf{m}}^1 - \hat{\mathbf{m}}^0 - \left(\mathbf{m}^1 - \mathbf{m}^0\right)\right) \Rightarrow N(0, \mathbf{V})$$

where $\mathbf{V}$ is the $K(J-1) \times K(J-1)$ asymptotic variance-covariance matrix. The $\left(\sum_{j=1}^{J-1} K(j-1) + k, \sum_{j'=1}^{J-1} K(j'-1) + k'\right)$-th element of $\mathbf{V}$ is equal to

$E\left[\left(\phi_j^1(\tau_k) - \phi_j^0(\tau_k)\right)\left(\phi_{j'}^1(\tau_{k'}) - \phi_{j'}^0(\tau_{k'})\right)\right]$ with

$$\begin{aligned}\phi_j^z(\tau_k) \equiv & \phi_j^z(\tau_k; Y, T, Z, \mathbf{X}) = \frac{I(\tau_k) - m_j^z(\tau_k)}{p_{Z,X}(z, \mathbf{x}_j)} \mathbf{1}(Z = z, \mathbf{X} = \mathbf{x}_j) \\ & - \frac{f_{Y|T,Z,\mathbf{X}}(q_{0|C}(\tau_k)|0, z, \mathbf{x}_j)(1 - p_{T|Z,\mathbf{X}}(z, \mathbf{x}_j))}{P_c f_{0|C}(q_{0|C}(\tau_k))} \psi_0(Y, T, Z, \mathbf{X}) \\ & - \frac{f_{Y|T,Z,\mathbf{X}}(q_{1|C}(\tau_k)|1, z, \mathbf{x}_j) p_{T|Z,\mathbf{X}}(z, \mathbf{x}_j)}{P_c f_{1|C}(q_{1|C}(\tau_k))} \psi_1(Y, T, Z, \mathbf{X}),\end{aligned}$$

where $\psi_0(Y, T, Z, \mathbf{X})$ and $\psi_1(Y, T, Z, \mathbf{X})$ are defined in the proof of Theorem 7 in Frolich and Melly (2007), and are restated in the proof of this theorem in the Appendix.

The last two terms of $\phi_j^z(\tau_k; Y, T, Z, \mathbf{X})$ come from the estimation error of $\hat{q}_{0|C}(\tau_k)$ and $\hat{q}_{1|C}(\tau_k)$. If $q_{0|C}(\tau_k)$ and $q_{1|C}(\tau_k)$ were known, $\phi_j^z(\tau_k; Y, T, Z, \mathbf{X})$ would reduce to $\frac{I(\tau_k) - m_j^z(\tau_k)}{p_{Z,\mathbf{X}}(z, \mathbf{x}_j)} \mathbf{1}(Z = z, \mathbf{X} = \mathbf{x}_j)$ and hence the $\left(\sum_{j=1}^{J-1} K(j-1) + k, \sum_{j'=1}^{J-1} K(j'-1) + k'\right)$-th element of $\mathbf{V}$ is equal to if $j = j'$, and 0 if $j \neq j'$.

Given the above theorem, it follows immediately that with a consistent estimator $\hat{\mathbf{V}}$ for the variance-covariance matrix $\mathbf{V}$, the test statistic W converges to a Chi-squared distribution with $K(J-1)$ degrees of freedom under the null hypothesis, where $K(J-1)$ is the number of moment restrictions in H_0 as well as the rank of $\mathbf{V}$ given Assumption 3.4. Under the alternative hypothesis, the test statistic W explodes. Due to the complicated nature of the asymptotic variance-covariance matrix resulting from the first stage estimation of the unconditional quantile functions, we recommend estimating $\mathbf{V}$ by bootstrapping. Note that the set $\mathcal{X}$ is finite here. Section 7 discusses an extension that allows J to increase with sample size. In that case, the estimation error of the unconditional quantile functions does not play a role in the asymptotic distribution of the test statistic. The variance-covariance matrix can be estimated analytically.

Let the critical value c_α of the test be the $(1-\alpha) \times 100$-th percentile of the $\chi^2(K(J-1))$

distribution. Define the decision rule of the test as

$$\text{“reject the null hypothesis } H_0 \text{ if } W > c_\alpha\text{”}.$$

The following Corollary summarizes the asymptotic properties of the proposed test.

Corollary 1. *Given Assumptions 2 and 3, the proposed test satisfies*

1. *if H_0 is true,* $\lim_{n\to\infty} P(W > c_\alpha) = \alpha$, *and*

2. *if H_0 is false,* $\lim_{n\to\infty} P(W > c_\alpha) = 1$.

Note that Assumption 3.4 guarantees that for all $j = 1, ..., J-1$ and $k = 1, ..., K$, $\phi_j^1(\tau_k; Y, T, Z, \mathbf{X}) - \phi_j^0(\tau_k; Y, T, Z, \mathbf{X})$ is not degenerate and hence the variance-covariance matrix $\mathbf{V}$ has full rank. In practice with a finite sample, it is possible that for some small cells defined by values of $\mathbf{X}$ and Z, both $\hat{m}_j^1(\tau_k)$ and $\hat{m}_j^0(\tau_k)$ are degenerate. In that case, $\hat{\mathbf{V}}$ would not have full rank. The effective number of moment restrictions in H_0 is then the rank of $\hat{\mathbf{V}}$, which should be used as the degrees of freedom for the the test statistic.

4.2 The Mean Test for Rank Similarity

As is discussed in Section 3.3, rank similarity implies equality of conditional moments of potential ranks. In this section we construct a mean test for rank similarity. Tests for other moments of potential ranks can be constructed similarly and are omitted to save space.

Let $\bar{m}_j^z = E[U|Z = z, \mathbf{X} = \mathbf{x}_j]$ for $z = 0, 1$. As discussed in Section 2, one can construct a mean test for rank similarity by testing the null hypothesis

$$H_{0,mean} : \bar{m}_j^0 = \bar{m}_j^1, \text{ for all } j = 1, ..., J-1.$$

Let $\mathsf{R}(y,t) = \int_0^1 \mathbf{1}\left(tq_{1|C}(\tau) + (1-t)q_{0|C}(\tau) \le y\right) d\tau$ be the rank function such that $U = \mathsf{R}(Y,T)$. We use simulation to generate the rank function. Let $\left(\tau^1, ...\tau^S\right)$ be S random numbers drawn from the uniform distribution $U(0,1)$ that is independent of the data. The rank of any interior point y can be estimated by

$$\hat{\mathsf{R}}(y,t) = \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\left(\left(t\hat{q}_{1|C}\left(\tau^s\right) + (1-t)\hat{q}_{0|C}\left(\tau^s\right)\right) \le y\right),$$

which converges to $\mathsf{R}(y,t)$ in probability as $S, n \to \infty$. Let $\hat{U}_i = \hat{\mathsf{R}}(Y_i, T_i)$ for $i = 1, ..., n$. For $z = 0, 1$ and $j = 1, ..., J$, define the estimator of $\bar{m}_j^z$ as

$$\ddot{m}_j^z = \frac{1}{n_j^z} \sum_{Z_i = z, \mathbf{X}_i = \mathbf{x}_j} \hat{U}_i.$$

Let $\ddot{\mathbf{m}}^z = \left[\ddot{m}_1^z \; \ddot{m}_2^z \; \cdots \; \ddot{m}_{J-1}^z\right]'$ and $\bar{\mathbf{m}}^z = \left[\bar{m}_1^z \; \bar{m}_2^z \; \cdots \; \bar{m}_{J-1}^z\right]'$ be $(J-1) \times 1$ vectors. Since

$$\begin{aligned}
\bar{m}_j^z &= E\left[\int_0^1 \mathbf{1}\left(\left(Tq_{1|C}(\tau) + (1-T)q_{0|C}(\tau)\right) \le Y\right) \middle| Z = z, \mathbf{X} = \mathbf{x}\right] \\
&= \int_0^1 \left(1 - m_j^z(\tau)\right) d\tau,
\end{aligned}$$

and

$$\begin{aligned}
\ddot{m}_j^z &= \frac{1}{n_j^z} \sum_{Z_i = z, \mathbf{X}_i = \mathbf{x}_j} \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\left(\left(T_i\hat{q}_{1|C}\left(\tau^s\right) + (1-T_i)\hat{q}_{0|C}\left(\tau^s\right)\right) \le Y_i\right) \\
&= \frac{1}{S}\sum_{s=1}^{S}\left(1 - \hat{m}_j^z(\tau^s)\right),
\end{aligned}$$

the asymptotic property of $\ddot{\mathbf{m}}^1 - \ddot{\mathbf{m}}^0$ can be readily derived following results of Theorem 2. It is summarized in the following Corollary.

Corollary 2. *Suppose that Assumption 3 holds for $\Omega = (0,1)$. Given Assumptions 2 and 3,* *under the null hypothesis* $\bar{\mathbf{m}}^1 = \bar{\mathbf{m}}^0$, *when* $S, n \to \infty$

$$\sqrt{n}\left(\ddot{\mathbf{m}}^1 - \ddot{\mathbf{m}}^0\right) \Rightarrow N(0, \mathbf{V}_{mean}),$$

where $\mathbf{V}_{mean}$ *is the* $(J-1)\times(J-1)$ *asymptotic variance-covariance matrix. The* (j,j')*-th element of* $\mathbf{V}_{mean}$ *is* $E\left[\left(\int_0^1 \phi_j^1(\tau)d\tau - \int_0^1 \phi_j^0(\tau)d\tau\right)\left(\int_0^1 \phi_{j'}^1(\tau)d\tau - \int_0^1 \phi_{j'}^0(\tau)d\tau\right)\right]$, *where*

$$\begin{aligned}\int_0^1 \phi_j^z(\tau)d\tau = & -\frac{U-\bar{m}_j^z}{p_{Z,X}(z,\mathbf{x}_j)}\mathbf{1}(Z=z,\mathbf{X}=\mathbf{x}_j)\\ & -\int_0^1 \frac{f_{Y|T,Z,\mathbf{X}}(q_{0|C}(\tau)|0,z,\mathbf{x}_j)}{f_{0|C}(q_{0|C}(\tau))}d\tau\frac{\left(1-P_{T|Z,\mathbf{X}}(z,\mathbf{x}_j)\right)\psi_0(Y,T,Z,\mathbf{X})}{P_c}\\ & -\int_0^1 \frac{f_{Y|T,Z,\mathbf{X}}(q_{1|C}(\tau)|1,z,\mathbf{x}_j)}{f_{1|C}(q_{1|C}(\tau))}d\tau\frac{P_{T|Z,\mathbf{X}}(z,\mathbf{x}_j)\psi_1(Y,T,Z,\mathbf{X})}{P_c}.\end{aligned}$$

Again, the last two terms of $\int_0^1 \phi_j^z(\tau)d\tau$ come from the estimation error in the first-step estimation of $q_{0|C}(.)$ and $q_{1|C}(.)$.[9] This corollary leads to the following Wald-type test statistic

$$W_{mean} \equiv n\left(\ddot{\mathbf{m}}^1 - \ddot{\mathbf{m}}^0\right)'\ddot{\mathbf{V}}^{-1}\left(\ddot{\mathbf{m}}^1 - \ddot{\mathbf{m}}^0\right),$$

where $\ddot{\mathbf{V}}$ is the estimated variance-covariance matrix. The test statistic W_{mean} converges to a Chi-squared distribution with $J-1$ degrees of freedom under the null hypothesis, and explodes under the alternative. Therefore, to conduct the mean test for rank similarity at the significance level α, one can reject $H_{0,mean}$ if W_{mean} exceeds the $(1-\alpha)\times 100$-th percentile of the $\chi^2(J-1)$ distribution.

Note that rank similarity is a distributional concept. In practice, the mean test, which only tests a summary measure of the distribution, may have less power against different alternatives.[10] Results from both the Monte Carlo simulations and the JTPA application

[9] If $q_{0|C}(.)$ and $q_{1|C}(.)$ were known, $\int_0^1 \phi_j^z(\tau)d\tau$ would reduce to $-\frac{U-\bar{m}_j^z}{p_{Z,X}(z,\mathbf{x}_j)}\mathbf{1}(Z=z,\mathbf{X}=\mathbf{x}_j)$ and the off-diagonal elements in matrix $\mathbf{V}_{mean}$ would reduce to zero.

[10] Note also that the mean test is not computationally simpler than the distributional test, due to the estimation of each individual's rank.

confirms this point. However, the mean test can be useful when testing a large set of quantiles is not practical due to, e.g., a small sample size (so one can easily exhaust degrees of freedom).

5 Simulation

This section presents results from Monte Carlo simulations to illustrate the finite sample size and power properties of the proposed tests. We study both the distributional test and the mean test for rank similarity. We first consider exogenous treatment and then endogenous treatment.

For all Monte Carlo simulations, the observed covariate X is generated to take on five values with equal probability. In particular, $\Pr(X = 0.4j) = 0.2$, for $j = 1, ..., 5$. The unobserved covariate V and the idiosyncratic shocks S_0 and S_1 are generated as independent random variables and $V, S_0, S_1 \sim N(0, 1)$. In addition, $Y_0 = X + V + S_0$, $Y_1 = X + 1 - bXV + V + S_1$ and $Y = Y_0 (1 - T) + Y_1 T$. When $b = 0$, rank is invariant to treatment by construction; when $b \neq 0$, rank similarity does not hold. The value of b controls for the degree of violation in rank similarity. Greater value leads to greater violation. We consider $b = 0, 2$ and 3. DGP 1 focuses on exogenous treatment and DGP 2 focuses on endogenous treatment.

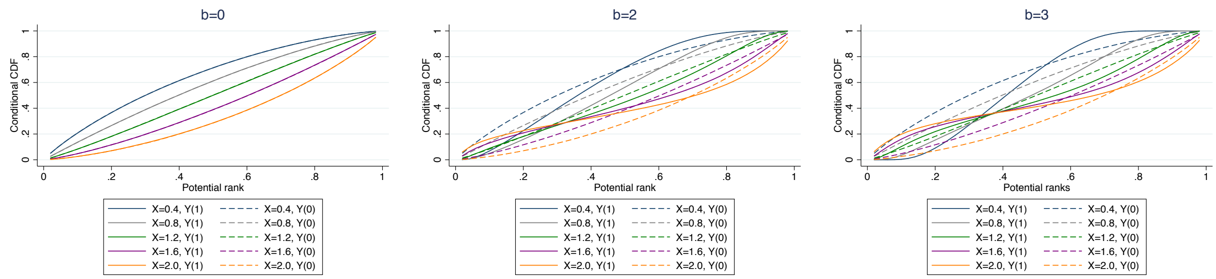
DGP 1: $\Pr(T = 0) = \Pr(T = 1) = 0.5$.

DGP 2: $\Pr(Z = 0) = \Pr(Z = 1) = 0.5$; $T = \mathbf{1}(0.15(Y_1 - Y_0) + Z - 0.5 > 0)$.

When treatment is endogenous, we have a generalized Roy model (Roy, 1951). Treatment is determined by an individual's benefit from the treatment $Y_1 - Y_0$ and additionally by an exogenous variable Z, which can be taken as a random assignment indicator for a program, representing program incentives. Subtracting 0.5 normalizes $Z - 0.5$ to be mean zero.

Figure 1 illustrates the design of DGP 1 for different values of b. In each graph, the solid lines represent the conditional distribution of potential ranks under treatment, i.e., $F_{U_1|X}(.|x)$ at $x = 0.4, 0.8, ..., 2$ while the dotted lines represent the corresponding conditional

Figure 1: Conditional distributions of potential ranks



distribution of potential ranks under no treatment $F_{U_0|X}(.|x)$. When $b = 0$, rank is invariant to treatment, so the two conditional CDFs overlap at all values of X. When $b \neq 0$, rank similarity is violated. Figure 1 shows that the degree of violation indeed increases with b. In addition, for $b = 2, 3$, rank similarity is violated more strongly at the lower quantiles.

The left half of Table 1 presents simulation results under DGP 1. We conduct the mean test as well as the distributional test with four different sets of quantiles at the 5% significance level. For each $b = 0, 2, 3$, we draw samples of size 500, 1000, 1500, 2000 and 2500. All test statistics are constructed using bootstrapped variance-covariance matrices with 1,000 bootstrap repetitions. 1,000 simulations are performed for each test, and the rejection rates are reported in Table 1.

Results for $b = 0$ under DGP 1 in Table 1 show that both the distributional test and the mean test control size well. Results for $b = 2, 3$ show that the power of the proposed tests increases with the sample size. The rejection rate goes to one rapidly with the increase of the sample size. In addition, our distributional tests are sensitive to the part of the distribution at which rank similarity is more seriously violated. When $b = 2, 3$, as shown in Figure 1, rank similarity is violated more seriously at the lower quantiles. Consequently, the distributional tests show greater power when these lower quantiles are included in the tests. In contrast, the mean test for rank similarity has much lower power due to the fact that rank similarity is violated only at part of the distribution.

Figure 2 demonstrates the power performance of the proposed tests visually. The left

Table 1: Small sample performance of the proposed tests

N	DGP 1 500	1000	1500	2000	2500	DGP 2 500	1000	1500	2000	2500
	$b=0$					$b=0$				
Test 1: $\Omega=\{0.5\}$	0.034	0.039	0.051	0.040	0.053	0.025	0.036	0.041	0.038	0.057
Test 2: $\Omega=\{0.2,0.3,0.4\}$	0.013	0.013	0.025	0.021	0.023	0.012	0.012	0.018	0.017	0.025
Test 3: $\Omega=\{0.5,0.6,0.7,0.8\}$	0.014	0.014	0.023	0.023	0.018	0.006	0.013	0.016	0.022	0.015
Test 4: $\Omega=\{0.2,0.3,...,0.8\}$	0.006	0.010	0.013	0.013	0.013	0.002	0.010	0.006	0.010	0.008
Test 5: Mean Test	0.051	0.044	0.048	0.041	0.067	0.054	0.050	0.051	0.045	0.057
	$b=2$					$b=2$				
Test 1: $\Omega=\{0.5\}$	0.074	0.150	0.232	0.303	0.388	0.084	0.242	0.379	0.522	0.615
Test 2: $\Omega=\{0.2,0.3,0.4\}$	0.269	0.776	0.968	0.994	1.000	0.170	0.589	0.870	0.965	0.993
Test 3: $\Omega=\{0.5,0.6,0.7,0.8\}$	0.151	0.581	0.857	0.962	0.991	0.021	0.150	0.340	0.600	0.764
Test 4: $\Omega=\{0.2,0.3,...,0.8\}$	0.287	0.910	0.996	1.000	1.000	0.053	0.431	0.823	0.960	0.993
Test 5: Mean Test	0.103	0.213	0.278	0.424	0.500	0.152	0.322	0.481	0.622	0.709
	$b=3$					$b=3$				
$\Omega=\{0.5\}$	0.143	0.335	0.512	0.640	0.800	0.113	0.293	0.441	0.617	0.700
$\Omega=\{0.2,0.3,0.4\}$	0.817	0.999	1.000	1.000	1.000	0.284	0.783	0.975	1.000	1.000
$\Omega=\{0.5,0.6,0.7,0.8\}$	0.306	0.880	0.992	1.000	1.000	0.020	0.198	0.450	0.704	0.865
$\Omega=\{0.2,0.3,...,0.8\}$	0.836	0.999	1.000	1.000	1.000	0.093	0.634	0.949	1.000	1.000
Mean Test	0.340	0.659	0.853	0.941	0.971	0.191	0.441	0.602	0.772	

graph presents evidence when the sample size is fixed at 1,000 and the value of b varies; the right graph presents evidence when b is fixed at $b=2$ and the sample size varies. As is clear from these graphs, the distributional tests at a wide range of quantile values in general have better small sample performance given the DGP under study.

Next we study the performance of the proposed tests under DGP 2, where treatment is endogenous. Figure 3 illustrates the design of DGP 2. In each graph, the solid lines represent the conditional distribution of compliers' ranks under treatment, i.e., $F_{U_1|C,X}(.|x)$ at

Figure 2: Small sample performance of the proposed tests: exogenous treatment

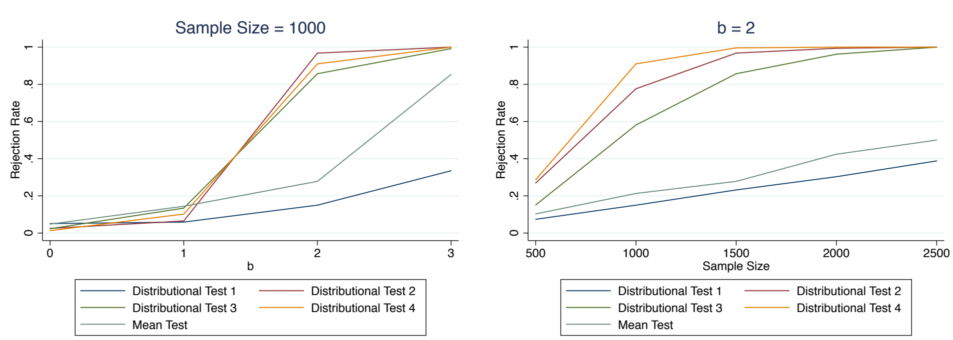


Figure 3: Conditional distributions of potential ranks among compliers

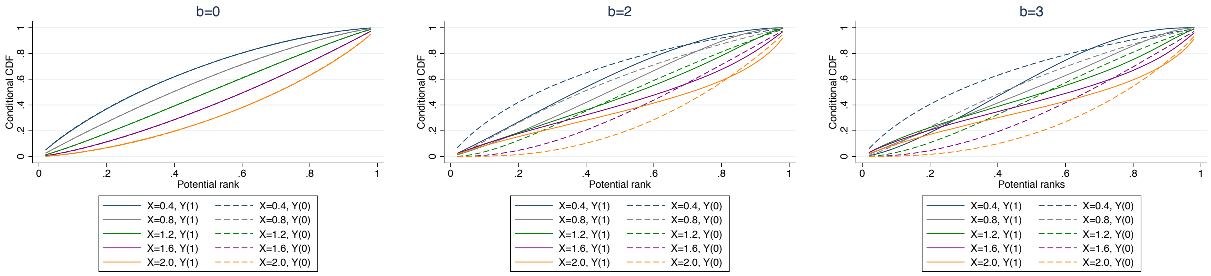
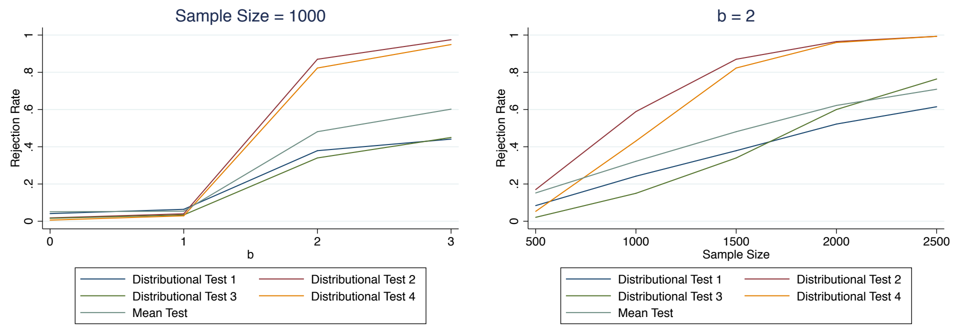


Figure 4: Small sample performance of the proposed tests: endogenous treatment



$x = 0.4, 0.8, ..., 2$ while the dotted lines represent the corresponding conditional distribution of compliers' potential ranks under no treatment $F_{U_0|C,X}(.|x)$. The violation of rank similarity is greater when the value of b is larger, especially at the lower quantiles.

The right half of Table 1 presents the simulation results under DGP 2. The proposed tests again control size well. The power of the tests increases quickly with the sample size. The distributional tests at a range of quantile values again outperform the mean and median tests in terms of power. Note that at any given sample size, the tests under DGP2 when treatment is exogenous has relatively lower power than those under DGP 1 when treatment is exogenous. This is not surprising, since 1) the effective sample size is determined by compliers, and 2) the ranks for always takers and never takers do not change by construction.

Finally, Figure 4 visually illustrates the small sample performance of the proposed tests with exogenous treatment. Again the distributional tests when a wide range of quantiles are included in the test generally perform better in small samples.

6 Empirical Applications

6.1 JTPA

The impact of job training programs on the earnings of trainees, especially those with low income, is of great interest to both policy makers and economists. Abadie, Angrist, and Imbens (2002) and Chernozhukov and Hansen (2008) utilize data from a randomized experiment conducted under the JTPA to estimate the impact of the JTPA training program on the distribution of trainee earnings. An interesting feature of the JTPA training experiment is that there are almost no always takers, so the estimated LQTEs can be seen as the QTEs for the treated or the trainees. Both Abadie, Angrist, and Imbens (2002) and Chernozhukov and Hansen (2008) focus on conditional QTEs.

Here we estimate unconditional QTEs, as we are interested in learning how training affects the unconditional distribution of earnings. We then test for rank similarity. If rank similarity were true, one could infer from the estimated distributional effects the causal impacts of training on individual trainees at different quantiles of the earnings distribution. Abadie, Angrist, and Imbens (2002) find that the JTPA training program has significant impacts at every quantile of the earnings distribution for females, with largest proportional effects at the low quantiles. In contrast, training does not raise the low quantiles of earnings for males. We find similar patterns for the unconditional QTEs. However, it still remains to be seen whether training induces individuals to systematically change their ranks in the earnings distribution. For example, can we conclude that training has no real impacts on male trainees at the lower tail of quantiles?

We use the same data as those used in Abadie, Angrist, and Imbens (2002), but we additionally obtain age in years (instead of 5 age categories) to perform falsification tests for our rank similarity tests. The sample consists of 5,102 observations for males and 6,102 observations for females. The data contain information on earnings (Y), training (T) and treatment assignment status (Z), and some pre-determined individual characteristics ($\mathbf{X}$). Earnings are measured as total earnings over the 30 month period following the assignment into the treatment or control group. The set of individual characteristics includes dummies for black or Hispanic applicants, a dummy for high-school graduates or GED holders, a dummy for married applicants, whether the applicant worked at least 12 weeks in the 12 months preceding random assignment, a dummy for AFDC receipt (for women only) and 5 age category dummies. Institutional details along with information regarding the experimental data collection and sample selection criteria can be found in Abadie, Angrist, and Imbens (2002).

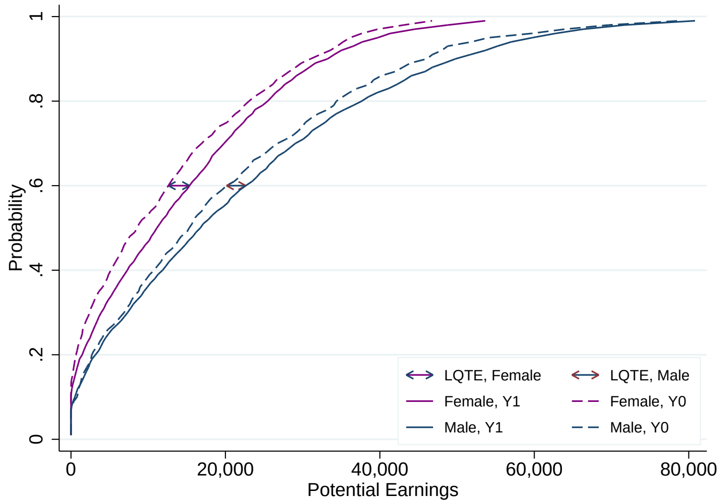
Table 2: First-stage estimates of unconditional QTEs of training on trainee earnings

	Female			Male		
Quantile	Y_0	QTE		Y_0	QTE	
0.15	195	291	(341.88)	1,462	249	(713.36)
0.20	723	714	(358.31)*	2,733	390	(723.01)
0.25	1,458	1,200	(372.08)***	4,434	489	(746.85)
0.30	2,463	1,380	(399.21)***	6,993	340	(891.74)
0.35	3,784	1,705	(497.01)***	8,836	594	(1,042.40)
0.40	5,271	1,974	(669.75)***	11,010	723	(1,104.63)
0.45	6,726	2,451	(766.25)***	13,104	1,069	(1,144.28)
0.50	8,685	2,436	(829.29)***	15,374	1,291	(1,234.59)
0.55	11,007	2,089	(877.56)**	17,357	2,239	(1,295.79)*
0.60	12,618	2,729	(886.96)***	20,409	2,118	(1,418.40)
0.65	14,682	2,943	(920.45)***	23,342	2,319	(1,557.00)
0.70	16,971	2,772	(1,027.14)***	27,169	1,780	(1,606.66)
0.75	20,252	2,106	(1,152.35)*	30,439	2,408	(1,641.47)
0.80	23,064	2,331	(1,149.71)**	34,620	2,800	(1,701.90)*
0.85	26,735	1,762	(1,179.91)	39,233	3,955	(1,886.98)**

Note: Standard errors are in the parentheses; All estimates control for covariates including dummies for black, Hispanic, high-school graduates (including GED holders), marital status, whether the applicant worked at least 12 weeks in the 12 months preceding random assignment, and AFDC receipt (for women only) as well as 5 age group dummies; * significant at the 10% level, ** significant at the 5% level, ***significant at the 1% level.

Table 2 presents the estimated unconditional QTEs at equally-spaced quantiles from 0.15 to 0.85 with an increment of 0.05. Also presented are quantiles of the potential earnings without training. Thus, the ratio of the two numbers in each row gives the percentage change in earnings at each quantile. They would represent real impacts of training on earnings for individuals at each quantile if rank similarity were satisfied. Similar to the findings documented in Abadie, Angrist, and Imbens (2002) for the conditional distribution

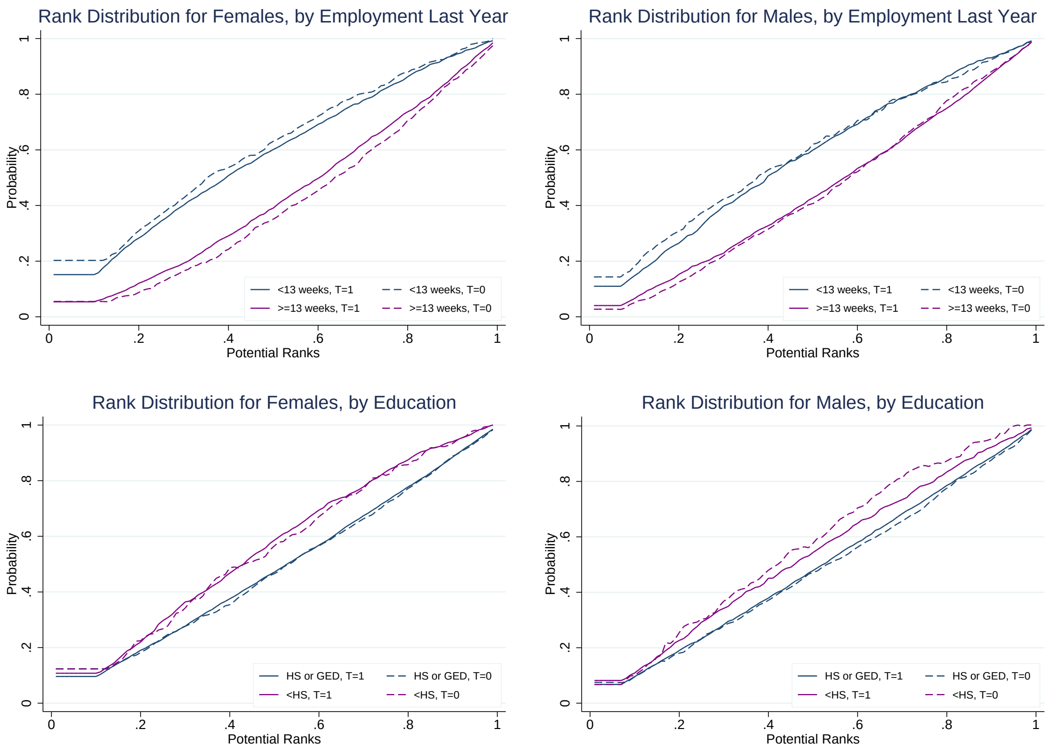
Figure 5: Distributions of potential earnings for trainees



of earnings, estimates in Table 2 show that the JTPA training program has significant impacts at almost every quantile for female trainees. The corresponding percentage changes are larger at lower quantiles due to females' very low potential earnings without training. In sharp contrast, the estimated unconditional QTEs are much smaller and insignificant among male trainees at the low quantiles. At the same time, male trainees have much higher potential earnings without training, leading to small and insignificant percentage changes at low quantiles. The estimated QTEs for males are much larger in absolute terms above the median, but still small in percentage terms. Also, only estimates right above the median and at the top quantiles are statistically significant. Figure 5 shows the counterfactual distribution of potential earnings for trainees under training or no training. Consistent with the estimates in Table 2, for male trainees the lower tail of the earnings distribution does not change much.

Figure 6 present the counterfactual distribution of potential ranks for sub-groups defined by education and by employment the year before the randomization. In each graph, the dotted lines represent the cumulative density functions (CDF) of potential ranks under no training, while the solid lines represent those under training. Those plots illustrate that different sub-groups may move up or down in ranks when receiving training services. Note that these figures present evidence at the aggregate level when sub-groups are defined by

Figure 6: Distributions of potential ranks for trainees



characteristics at a couple of dimensions. Below we provide formal test results for rank similarity.

Panel A of Table 3 reports results of the distributional tests for rank similarity at two different sets of quantiles. In columns I, the χ^2 tests are conducted jointly at $\Omega = \{0.15, 0.20, ...0.85\}$, while in columns II, the tests are conducted jointly at $\Omega = \{0.20, 0.3, ...0.80\}$. For both tests, we either control for covariates in the first-stage unconditional QTE estimation or not. To ensure the common support assumption (Assumption 2.4), in constructing the test statistics we do not use $\mathbf{X}$ values with less than 5 observations when either $Z = 1$ or $Z = 0$.

Panel B of Table 3 reports results from the same tests except that we replace the dependent variable earnings with age in years. Rank similarity holds trivially in this case, since training should not have causal effects on age. Further, individual characteristics are correlated with age, so these tests serve as valid falsification tests for this empirical application.

Table 3: The distributional tests for rank similarity - JTPA

	Female				Male			
	I		II		I		II	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
	Panel A: Dependent Var. Earnings							
χ^2	7,652.1	7,763.8	1,197.2	1,177.8	2,780.7	2,719.0	886.1	876.8
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
d.f.	1,544	1,544	723	723	1,218	1,218	570	570
	Panel B: Falsification test (Dependent Var. Age)							
χ^2	478.8	471.9	252.0	259.9	209.3	203.5	124.7	123.0
	(0.926)	(0.953)	(0.366)	(0.245)	(1.000)	(1.000)	(0.977)	(0.982)
d.f.	525	525	245	245	338	338	158	158

Note: Results are based on the Chi-squared test in Theorem 2; Variance-covariance matrices are bootstrapped with 2,000 replications; P-values are in the parentheses; Columns I report a joint test at equally-spaced 15 quantiles from 0.15 to 0.85; Columns II reports a joint test at equally-spaced 7 quantiles from 0.20 to 0.80; (1) controls for covariates in the first-stage unconditional QTE estimation, while (2) does not; X values with fewer than 5 observations when either $Z = 0$ or $Z = 1$ are not used in the test to ensure the common support assumption.

As we can see from Table 3, rank similarity can be strongly rejected for both female and male trainees. The results are very similar regardless of whether one controls for covariates or not in the first stage. This is due to the fact that assignment to treatment is well randomized in the JTPA experiment, and so the assignment indicator can serve as a valid IV conditional on covariates or not. Consistent with the findings here, Abadie, Angrist, and Imbens (2002) show that the base line covariates are roughly balanced by assignment status. In sharp contrast, when age is used as the dependent variable rank similarity can not be rejected for either females or males. Not surprisingly training itself does not cause individuals to systematically change their ranks in the distribution of ages.

To investigate how seriously rank similarity is violated at different parts of the potential earnings distribution. Panel A of Table 4 presents results of the rank similarity test at each quantile from 0.15 to 0.85. Panel B presents results of the corresponding falsification test at each quantile. For males, rank similarity can be rejected at almost all quantiles except for the 0.35 and 0.40 quantiles. For females, rank similarity can be rejected strongly at the lower tail of the distribution, but not much near the median or above. Again in sharp contrast, at the 5% significance level we fail to reject rank similarity at all quantiles for both females

Table 4: The individual quantile tests for rank similarity - JTPA

	Panel A: Dependent Var. Earnings				Panel B: Falsification test (Dependent Var. Age)			
	Female		Male		Female		Male	
Quantile	χ^2		χ^2		χ^2		χ^2	
0.15	134.4	(0.012)	103.8	(0.045)	43.9	(0.144)	19.4	(0.561)
0.20	143.0	(0.004)	113.3	(0.010)	37.9	(0.340)	22.1	(0.391)
0.25	126.2	(0.060)	107.8	(0.025)	26.0	(0.863)	13.9	(0.907)
0.30	131.9	(0.034)	104.7	(0.039)	26.9	(0.834)	15.0	(0.861)
0.35	147.2	(0.003)	95.8	(0.142)	22.1	(0.956)	17.9	(0.712)
0.40	118.3	(0.160)	88.6	(0.291)	31.1	(0.659)	23.2	(0.447)
0.45	107.5	(0.387)	110.7	(0.019)	32.1	(0.611)	22.4	(0.497)
0.50	110.9	(0.304)	113.6	(0.012)	32.3	(0.599)	19.2	(0.692)
0.55	112.6	(0.266)	110.9	(0.019)	30.8	(0.673)	19.6	(0.664)
0.60	112.1	(0.276)	112.3	(0.015)	32.7	(0.581)	22.3	(0.503)
0.65	121.7	(0.113)	105.0	(0.044)	29.4	(0.734)	18.4	(0.735)
0.70	108.0	(0.375)	106.1	(0.038)	36.7	(0.388)	24.0	(0.402)
0.75	130.4	(0.035)	109.7	(0.018)	45.4	(0.112)	16.5	(0.831)
0.80	118.4	(0.128)	116.5	(0.005)	47.7	(0.074)	17.1	(0.802)
0.85	92.3	(0.697)	118.7	(0.002)	44.7	(0.125)	18.7	(0.716)

Note: Results are based on the Chi-squared test in Theorem 2; Variance-covariance matrices are bootstrapped with 2,000 replications; P-values are in the parentheses; Covariates are controlled for in the first-stage unconditional QTE estimation. X values with fewer than 5 observations when either $Z = 1$ or $Z = 0$ are not used in the test to ensure the common support assumption.

and males when age in years is used as the dependent variable.[11]

We next conduct the mean test for rank similarity. These additional results are presented in Panel A of Table 5. Again we can strongly reject rank similarity for the male sample. The evidence is a bit weaker for the female sample. In particular, for females we cannot reject rank similarity at the 5% significance level. The test is marginally significant at the 10% level. This result may not be surprising, given that rank similarity is violated mainly at the lower tail of the earnings distribution for female trainees. As illustrated in the Monte Carlo simulations, the mean test for rank similarity may have less power in this case. We also conduct the mean test using age as the dependent variable. The results are reported in Panel B of Table 5. Again we fail to reject rank similarity with high p-values.

The above results show that rank similarity is seriously violated for male trainees. Recall that the estimated distributional effects for males are mostly small and insignificant. These test results are interesting, since they suggest that training causes men to systemically change

[11]The only time the test is rejected only at the 10% significance level is when looking at the 0.80 quantile for females. There is no systematic evidence of violation of rank similarity otherwise.

Table 5: The mean tests for rank similarity - JTPA

	Female				Male			
	(1)		(2)		(1)		(2)	
	Panel A: Dependent Var. Earnings							
χ^2	123.1	(0.098)	123.1	(0.098)	115.2	(0.009)	115.2	(0.009)
d.f.	104		104		82		82	
	Panel B: Falsification test (Dependent Var. Age)							
χ^2	30.6	(0.683)	30.6	(0.683)	18.4	(0.736)	18.4	(0.736)
d.f.	35		35		23		23	

Note: Results are based on the Chi-squared test for the mean ranks only; Variance-covariance matrices are bootstrapped with 2,000 replications; P-values are in the parentheses; (1) controls for covariates in the first-stage unconditional QTE estimation, while (2) does not; X values with fewer than 5 observations when either $Z = 1$ or $Z = 0$ are not used in the test to ensure the common support assumption.

their ranks in the distribution of earnings and that the distribution of program effects are more complicated than those suggested by the estimated QTE's for male trainees. For female trainees, training seems to cause more changes in ranks at the low quantiles of the earnings distribution.

Overall evidence suggests that the estimated QTEs can at best reflect how the earnings distribution changes with training. One should be cautious in equating the distributional effects of the JTPA training (as represented by QTEs) with the true impacts on individual trainees. For example, although training does not raise the lower tail of the earnings distribution for males, it does not necessarily mean that the JTPA training has no real impacts on male trainees at the bottom of the earnings distribution.

Finally, it is worth mentioning that our results are largely consistent with the findings in Heckman, Smith, and Clements (1997). Also using the JTPA experimental data, Heckman, Smith, and Clements (1997) conduct large-scale permutation exercises to investigate the program's impact distribution. They show that "heterogeneity is an important feature of impact distributions" and that "perfect positive dependence across potential outcome distributions – produces estimates of impact distributions that are not credible."

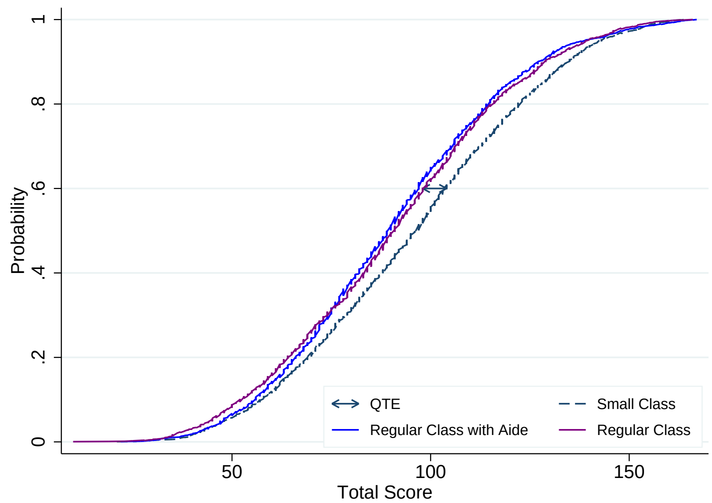
6.2 Project STAR

Project STAR (Student-Teacher Achievement Ratio) is a large-scale randomized experiment designed to study the effect of class size on students' academic performance. The experiment took place in Tennessee in the mid-1980's. Teachers and over 11,000 students in 79 public schools were randomly assigned to either a small class (13-17 students), a regular-size class (22-25 students), or a regular-size class with a full-time teacher aide from grade K to 3.

The STAR literature documents sizeable effects of attending a small class on students' performance and little effects of having a teacher aide (see, e.g., Krueger, 1999).[12] In addition, Whitmore (2005) evaluates the gender difference in the effects of attending small classes in Project STAR, as well as the peer effect of having a high percentage of female classmates (who are relatively better performing than boys). It is shown that attending a small class has positive effects on average for both boys and girls, and that there is no significant gender difference in the estimated average effects. Further attending a predominantly female class has positive impacts on a student's test scores. More recently, Armstrong and Shen (2015) find that the effect of attending a small class depends on teacher experience and that boys assigned to teachers with 6-10 years' experience show the biggest average improvement in test scores. Educators, policy makers and researchers have long been concerned that boys are falling behind in early childhood education, since early education may have long-run impacts. We therefore investigate the impacts of small classes and having a teacher aide on the gender gap in test scores in grade-K. We also investigate how class size or teacher aide interact with teachers' experience in affecting students' performance. Test scores are ordinal in nature, e.g., they cannot be summed and redistributed, so it is natural to focus on students' test score ranks and examine the distributions of students' ranks across class types and teacher experience.

[12] In addition, it is documented that attending small classes improves students' academic performance in the short run (Krueger, 1999) and increases students' probabilities of taking college-entrance exam (Krueger and Whitmore, 2001), attending college as well as earnings at age 27 (Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan, 2010).

Figure 7: Distributions of potential test scores for kindergarteners



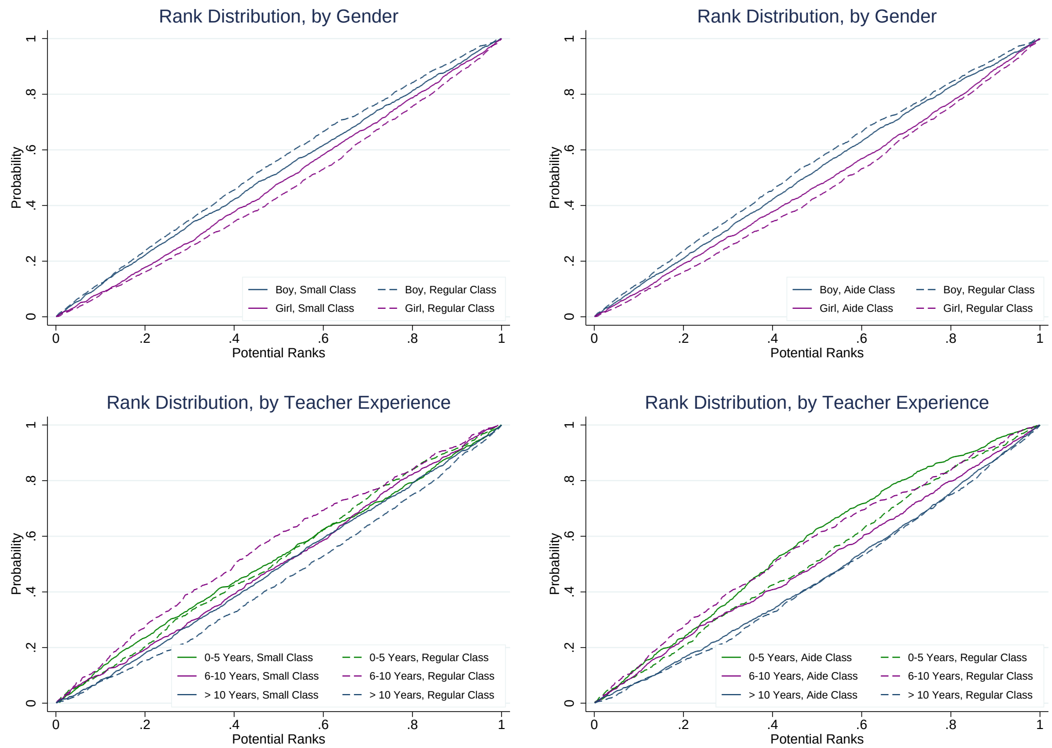
We use data from Achilles, Bain, Bellott, Boyd-Zaharias, Finn, Folger, Johnston, and Word (2008).[13] The sample consists of 5,692 grade K students with non-missing grade K total test scores (the outcome Y here). Among these 5,692 kindergartners, 1,718 are assigned to attend a small class, 1,989 are assigned to attend a regular class with a teacher aide, and the remaining 1,985 attend a regular class without a teacher aide. The treatment T is then attending a small or an aide class instead of a regular class. We look at class enrollment and test scores in kindergarten. Given that there is little non-compliance, we assume that the treatment is exogenous.[14] Figure 7 shows the distributions of test scores for students in the three types of classes. Attending a small class substantially improves the (unconditional) score distribution in grade K while attending a regular class with a teacher aide does not seem to have much impact on the (unconditional) score distribution.

Figure 8 presents the distributions of potential ranks across class types for boys and girls, as well as those for students assigned to teachers with different experience levels. In each graph, the dotted lines represent the CDF of potential ranks under control (regular class), while the solid lines represent those under treatment (small or aide class). The top left graph

[13] The data set can be downloaded from the Harvard Dataverse at http://hdl.handle.net/1902.1/10766.

[14] Krueger (1999) shows that only 0.3% of students in the experiment were not enrolled in the class type to which they were randomly assigned in kindergarten.

Figure 8: Distributions of potential ranks for kindergartners



shows that attending a small class helps boys catch up with girls and narrows the gender gap in ranks, i.e., the CDF curves move closer to the (invisible) 45 degree line. Interestingly the changes are smaller at the lower tail of the distribution. It seems that attending a small class is more beneficial for relatively better performing boys and improves their ranks relative to girls'. In the mean time, the top right graph shows that having a teacher aide also narrows the gender gap, but the effects are slightly smaller at the top part of the distribution. So having a teacher aide is more helpful for those relatively low performing boys.

The bottom left graph shows that a small class size greatly narrows the achievement gaps among students taught by teachers with differen experience levels. The greatest improvement is observed for students assigned to teachers with 6-10 years' experience. The bottom right graph, in contrast, shows that assigning an aide to a 'greed-handed' teacher is relatively inefficient and in fact negatively affects students' ranks. Again, teachers with mid-level experience (6-10 years) benefit most from having an aide. In the following we conduct the proposed tests and formally quantify the statistical precision of the rank changes.

Table 6: The distributional and mean tests for rank similarity - STAR

	Small v.s. Regular without Aide		Regular with v.s. without Aide		Small v.s. Regular	
	I	II	I	II	I	II
			Panel A: Dependent Var. Total Score			
χ^2	31.90	14.78	25.35	10.75	21.63	7.77
	(0.007)	(0.011)	(0.045)	(0.057)	(0.118)	(0.169)
d.f.	15	5	15	5	15	5
# of clusters	226	226	197	197	324	324
N	3,699	3,699	3,972	3,972	5,688	5,688
			Panel B: Falsification test (Dependent Var. Age)			
χ^2	8.69	3.54	11.43	4.76	11.67	2.35
	(0.893)	(0.617)	(0.722)	(0.445)	(0.704)	(0.799)
d.f.	15	5	15	5	15	5
# of clusters	226	226	197	197	324	324
N	3,699	3,699	3,972	3,972	5,688	5,688

Note: Results are based on the Chi-squared test in Theorem 2 for the special case with $T = Z$; Variance-covariance matrices are bootstrapped with 2,000 replications, clustered at classroom level; P-values are in the parentheses; Columns I report results from the distributional test at quantiles 0.25, 0.50, and 0.75; Columns II reports the results from the mean test.

Based on the above discussion, we use as covariates $\mathbf{X}$ the gender dummy and the three dummies representing whether a teacher has 0 - 5, 6 - 10, or over 10 years' experience. Students' test scores in the same classroom are likely to be correlated. We follow the practice of the STAR literature (see, e.g., Whitmore, 2005) to cluster the bootstrapped variance-covariance matrix at the classroom level. The number of clusters varies from 197 to 324, depending on the samples used. Given the small number of clusters, and hence small effective sample size, we conduct the distributional test based on the three quartiles to conserve degrees of freedom and to avoid extreme quantiles. We additionally test each individual quantile at the 15 equally-spaced quantiles from 0.15 to 0.85, as we did for the JTPA program, to show at what quantiles rank similarity is violated.

Columns I and II in the top Panel A of Table 6 report results from the distributional test and the mean test for rank similarity, respectively. As shown in Table 6, rank similarity can be rejected at the 5% significance level for the distributional tests when the treatment is either small class or aide class. The mean tests show similar results. When students in regular classes with and without aide are pooled together and compared with those in small classes, the evidence for violation of rank similarity is weaker, since students in the aide classes are also treated. Panel B of Table 6 reports results from falsification tests using age as the dependent variable.[15] All falsification tests fail to reject rank similarity.

Table 7: The individual quantile tests for rank similarity - STAR

	Panel A: Dependent Var. Test Score				Panel B: Falsification test (Dependent Var. Age)			
	Small v.s. Regular without Aide		Regular with v.s. without Aide		Small v.s. Regular without Aide		Regular with v.s. without Aide	
Quantile	χ^2		χ^2		χ^2		χ^2	
0.15	5.62	(0.345)	4.88	(0.431)	3.86	(0.570)	1.07	(0.957)
0.20	8.47	(0.132)	7.53	(0.184)	3.31	(0.652)	1.75	(0.883)
0.25	15.83	(0.007)	10.28	(0.068)	3.89	(0.566)	2.84	(0.724)
0.30	14.86	(0.011)	11.51	(0.042)	2.20	(0.820)	4.16	(0.527)
0.35	16.29	(0.006)	10.98	(0.052)	5.78	(0.328)	4.39	(0.495)
0.40	15.73	(0.008)	10.48	(0.063)	4.15	(0.528)	4.11	(0.534)
0.45	18.14	(0.003)	11.66	(0.040)	2.08	(0.838)	5.08	(0.406)
0.50	13.89	(0.016)	11.02	(0.051)	1.63	(0.898)	2.16	(0.827)
0.55	14.17	(0.015)	8.96	(0.111)	1.21	(0.944)	1.69	(0.891)
0.60	17.37	(0.004)	9.41	(0.094)	1.39	(0.925)	1.68	(0.891)
0.65	12.82	(0.025)	6.26	(0.281)	2.33	(0.801)	1.15	(0.949)
0.70	10.30	(0.067)	5.14	(0.399)	3.01	(0.699)	2.78	(0.735)
0.75	5.24	(0.387)	3.12	(0.682)	2.42	(0.788)	6.68	(0.245)
0.80	7.01	(0.220)	2.04	(0.844)	4.39	(0.495)	11.33	(0.045)
0.85	7.62	(0.179)	2.92	(0.713)	11.23	(0.047)	15.80	(0.007)

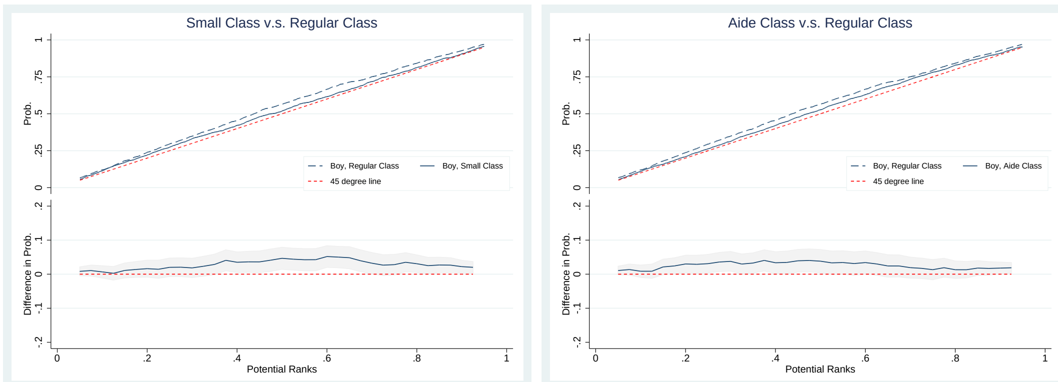
Note: Results are based on the Chi-squared test in Theorem 2 for the special case with $T = Z$; Variance-covariance matrices are bootstrapped with 2,000 replications, clustered at classroom level; P-values are in the parentheses.

Table 7 present results from the individual quantile tests. Panel A presents results for test scores, while Panel B presents results for ages (the falsification tests). When comparing small classes with regular classes, rank similarity can be rejected at most of the quantiles, particularly those in the middle. So although the top and bottom students are not significantly affected, students in the middle systematically change ranks when moving from regular-size classes to small classes. In contrast, evidence of potential rank changes is weaker when comparing aide classes with regular classes. There seem to be some rank changes below the median. Falsification tests again do not reveal systematic changes in ranks of age when the treatment is either small class or aide class.

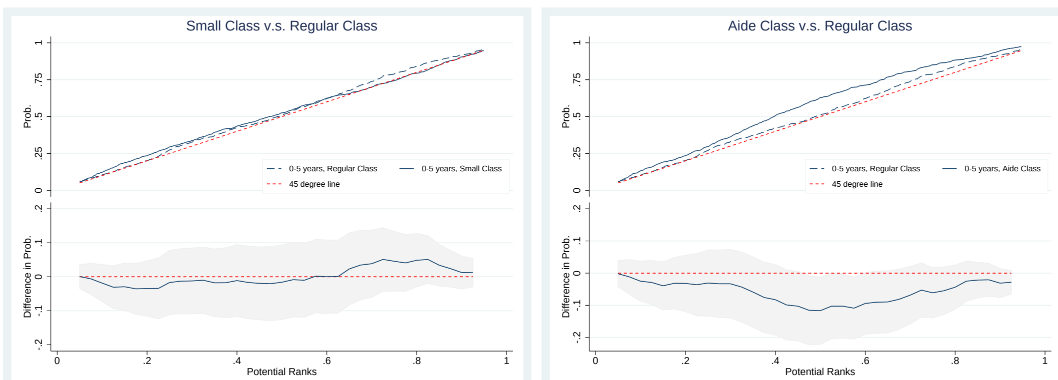
Figure 9 plots the distributions of potential ranks along with the changes between different types of classes for boys as well as those for students assigned to teachers with relatively low

[15]We construct students' exact age at December 31, 1985 from their date of birth information.

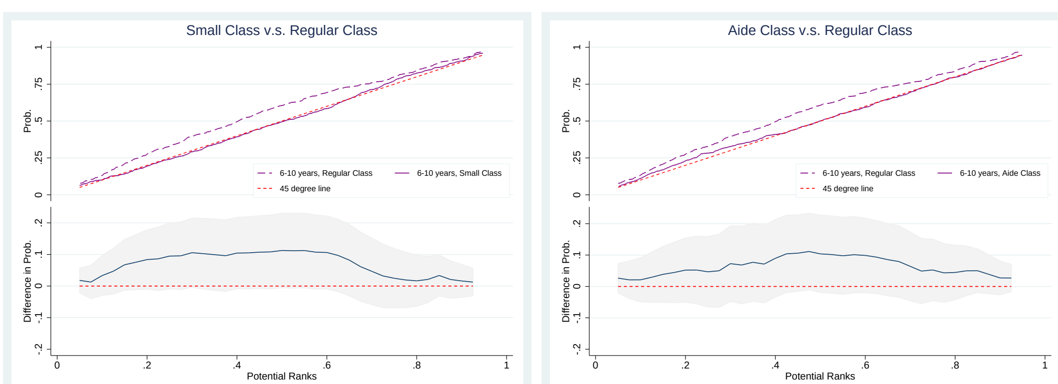
Figure 9: Changes in Potential Ranks for Sub-groups



(a) Boys



(b) Students Assigned to Teachers with 0-5 Years' Experience



(c) Students Assigned to Teachers with 6-10 Years' Experience

experience (0-5 or 6-10 years v.s. over 10 years' experience).[16] These figures confirm that attending a small class or having a teacher aide is beneficial for boys and narrows the gender performance gap (moving the CDFs closer to the 45 degree line). Students in small classes or in regular classes with a teacher aide also fare better when the teachers are somewhat experienced (6-10 years). On the other hand, contrary to what one might believe, an aide is relatively inefficient in improving students' ranks when assigned to an inexperienced teacher.

7 Extensions

7.1 Covariates with Infinite Support

If we allow J, the number of the unique values of $\mathbf{X}$, to go to infinity as the sample size n goes to infinity, then the second-stage conditional means would be estimated at a slower rate. The estimation error for the unconditional quantiles in the first stage, which is of order $\sqrt{n}$, is small enough relative to the estimation error in the second stage and hence can be ignored. Here we discuss a test statistic allowing $J \to \infty$ as $n \to \infty$. In this case the asymptotic distribution in Theorem 2 no longer holds because the density $p_{Z,X}$ in the denominator goes to zero in the limit. Recall that $n_j^z = \sum_{i=1}^n \mathbf{1}(Z_i = z, \mathbf{X}_i = \mathbf{x}_j)$ for $z = 0, 1$. We make the following assumptions on the data.

Assumption 4. *1. i.i.d. data: the data $\{Y_i, T_i, Z_i, \mathbf{X}_i\}$ for $i = 1, ..., n$ is a random sample of size n of $(Y, T, Z, \mathbf{X})$.*

2. For all $\tau \in \Omega = \{\tau_1, \tau_2, ..., \tau_K\}$, the random variable Y_1 and Y_0 are continuously distributed with positive density in a neighborhood of $q_{0|C}(\tau)$ and $q_{1|C}(\tau)$ in the subpopulation of compliers.

3. Let $n_j = \sum_{i=1}^n \mathbf{1}(\mathbf{X} = \mathbf{x_j})$. $n_j \asymp n/J$ uniformly over j, i.e. there exist $0 < c \leq C < \infty$ such that $c\frac{n}{J} \leq n_j \leq C\frac{n}{J}$ for all $j = 1, ..., J$.

[16] Note that figures for girls would mirror those for boys, so we do not present figures for all groups.

4. $\hat{\pi}(\mathbf{x}_j)$ *is uniformly consistent, or* $\sup_{j=1,\dots,J} |\hat{\pi}(\mathbf{x}_j) - \pi(\mathbf{x}_j)| \xrightarrow{p} 0$ *as* $n, J \to \infty$.

5. *For all* $t, z = 0, 1,\ j = 1, \dots, J$ *and* $\tau \in \Omega$, $f_{Y|T,Z,\mathbf{X}}(.|t, z, \mathbf{x}_j)$ *is bounded in a neighborhood of* $q_{t|C}(\tau)$. *For all* $\tau \in \Omega$ *and* $j = 1, \dots, J$, $f_{Y|\mathbf{X}}(.|\mathbf{x}_j)$ *is positive and bounded in a neighborhood of* $q_{t|C}(\tau)$.

The following corollary gives the asymptotic distribution of the estimator $\hat{\mathsf{m}}_j^1 - \hat{\mathsf{m}}_j^0$ for $\mathsf{m}_j^1 - \mathsf{m}_j^0$ for $j = 1, .., J-1$ when $J \to \infty$ as $n \to \infty$.

Corollary 3. *Given Assumptions 2 and 4 we have*

$$\sqrt{\frac{n_j^1 n_j^0}{n_j^1 + n_j^0}} \left(\hat{\mathsf{m}}_j^1 - \hat{\mathsf{m}}_j^0 - \left(\mathsf{m}_j^1 - \mathsf{m}_j^0 \right) \right) \Rightarrow \mathbf{Z}_j \sim N(0, \mathbf{V}_j),$$

where $\mathbf{Z}_j$ *for* $j = 1, \dots, J-1$ *follow independent multivariate normal distributions with mean zero and variance-covariance matrix* $\mathbf{V}_j$, *and the* (k, k')*-th element of* $\mathbf{V}_j$ *is*

$$V_{j;k,k'} = \pi(\mathbf{x}_j) m_j^1(\tau_k \wedge \tau_{k'}) \left(1 - m_j^1(\tau_{k'})\right) + (1 - \pi(\mathbf{x}_j)) m_j^0(\tau_k \wedge \tau_{k'}) \left(1 - m_j^0(\tau_{k'})\right).$$

For each $j = 1, \dots, J$, we can define the Wald-type statistic

$$w_j = \frac{n_j^1 n_j^0}{n_j^1 + n_j^0} \left(\hat{\mathsf{m}}_j^1 - \hat{\mathsf{m}}_j^0 \right)' \hat{\mathbf{V}}_j^{-1} \left(\hat{\mathsf{m}}_j^1 - \hat{\mathsf{m}}_j^0 \right)$$

where $\hat{\mathbf{V}}_j$ is a consistent estimator of $\mathbf{V}_j$. Let $\hat{V}_{j;k,k'}$ be the (k, k')-th element of V_j,

$$\hat{V}_{j;k,k'} = n_j^0/(n_j^0 + n_j^1) \hat{m}_j^1(\tau_k \wedge \tau_{k'}) \left(1 - \hat{m}_j^1(\tau_{k'})\right) + n_j^1/(n_j^0 + n_j^1) \hat{m}_j^0(\tau_k \wedge \tau_{k'}) \left(1 - \hat{m}_j^0(\tau_{k'})\right).$$

The test statistic under the null hypothesis of rank similarity can be constructed as

$$W_{largeJ} = \frac{\sum_{j=1}^{J-1} w_j - K(J-1)}{\sqrt{2K(J-1)}}.$$

Under the null, $W_{largeJ} \Rightarrow N(0,1)$ as $J \to \infty$. Let c_α be the $(1-\alpha) \times 100$-th percentile of the $N(0,1)$ distribution. The *one-sided* decision rule of the test is to

$$\text{“reject the null hypothesis } H_0 \text{ if } W_{largeJ} > c_\alpha\text{”.}$$

Note that the asymptotic theory of this test relies on having both J and the sample size at each value of J go to infinity, which might be a reasonable assumption when one has big data at hand.

7.2 Continuous Covariates

This section considers the case where one has continuous covariates instead of discrete ones. Let $\mathbf{X} = (X_1, ..., X_L)$ be a L dimensional vector of continuous variables. For any $\mathbf{x} \in \mathcal{X}$ and $z = 0, 1$, define $m_k^z(\mathbf{x}) = E[I_i(\tau_k)|Z_i = z, \mathbf{X} = \mathbf{x}]$. Let $m_k(.) = m_k^1(.) - m_k^0(.)$, we are interested in testing the following null hypothesis.

$$H_0 : m_k(\mathbf{x}) = 0 \textit{ for all } \mathbf{x} \in \mathcal{X} \textit{ and } k = 1, ..., K.$$

The following discusses estimation of $m_k(\mathbf{x})$ as well as a Kolmogorov-Smirnov type test statistic. Similar as before let $\widehat{q}_{0|C}(\tau_k)$ and $\widehat{q}_{1|C}(\tau_k)$ be $\sqrt{N}$-consistent estimators of $q_{0|C}(\tau_k)$ and $q_{1|C}(\tau_k)$, respectively. With product kernel functions $\mathcal{K}_{h_z}(\mathbf{X}_i - \mathbf{x}) = \frac{1}{h_z}\mathcal{K}\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_z}\right)$ for $z = 0, 1$, and bandwidths $h_0, h_1 \to 0$, we can define the local linear estimators for $\hat{m}_k^0(\mathbf{x})$ and $\hat{m}_k^1(\mathbf{x})$ as the intercepts a_0 and a_1 in the following minimization problems

$$\min_{a_0, b_{01}, \ldots b_{0L}} \sum_{Z_i=0} \left[\mathbf{1}\left(Y_i \leq \hat{q}_{1|C}(\tau_k)T_i + \hat{q}_{0|C}(\tau_k)(1-T_i)\right) - a_0 - \sum_{l=1}^{L} b_{0l}(X_{i,l} - x_l) \right]^2 \mathcal{K}_{h_0}(\mathbf{X}_i - \mathbf{x}),$$

$$\min_{a_1, b_{11}, \ldots b_{1L}} \sum_{Z_i=1} \left[\mathbf{1}\left(Y_i \leq \hat{q}_{1|C}(\tau_k)T_i + \hat{q}_{0|C}(\tau_k)(1-T_i)\right) - a_1 - \sum_{l=1}^{L} b_{1l}(X_{i,l} - x_l) \right]^2 \mathcal{K}_{h_1}(\mathbf{X}_i - \mathbf{x}).$$

We can then estimate $m_k(\mathbf{x})$ by $\hat{m}_k(\mathbf{x}) = \hat{m}_k^0(\mathbf{x}) - \hat{m}_k^1(\mathbf{x})$. Let $s_k(\mathbf{x})$ be the standard error of $\hat{m}_k(\mathbf{x})$ for all $k = 1, ..., K$, which can be estimated using the asymptotic formula (Fan and Gijbels, 1996) or bootstrap. The test statistic can then be defined as

$$KS = \sup_{k,\mathbf{x}} \left| \frac{\hat{m}_k(\mathbf{x})}{s_k(\mathbf{x})} \right|.$$

With a significance level α, the null hypothesis is rejected if

$$KS > c_\alpha,$$

where c_α is the critical value that satisfies

$$\lim_{n\to\infty} P\left(\sup_{k,\mathbf{x}} \left| \frac{\hat{m}_k(\mathbf{x}) - m_k(\mathbf{x})}{s_k(\mathbf{x})} \right| > c_\alpha \right) \leq \alpha. \tag{3}$$

The following theorem establishes conditions under which one can simulate the critical value. The asymptotic results utilize the inference method proposed in Chernozhukov, Lee, and Rosen (2011).

Assumption 5. *1. i.i.d. data: the data $\{(Y_i, T_i, Z_i, \mathbf{X}_i)\}$ for $i = 1, ..., n$ is a random sample of size n of $(Y, T, Z, \mathbf{X})$.*

2. For all $\tau \in \Omega = \{\tau_1, \tau_2, ..., \tau_K\}$, the random variable Y_1 and Y_0 are continuously distributed with positive density in a neighborhood of $q_{0|C}(\tau)$ and $q_{1|C}(\tau)$ in the subpopulation of compliers.

3. $\mathbf{X}|Z = z$ has a conditional density that is bounded away from zero on $\mathcal{X}$ from above and from below away from zero for both $z = 0, 1$. $\hat{\pi}(\mathbf{x})$ is uniformly consistent, or $\sup_{\mathbf{x}\in\mathcal{X}} |\hat{\pi}(\mathbf{x}) - \pi(\mathbf{x})| \xrightarrow{p} 0$. $\mathcal{X}$ is convex.

4. $I_i(\tau_k) - m_k^z(\mathbf{x})|Z = z, \mathbf{X} = \mathbf{x}$ has a conditional density that is bounded from above and from below away from zero uniformly over $\mathbf{x} \in \mathcal{X}$, $z \in \{0, 1\}$ and $\tau_k \in \Omega$.

5. *For all $k = 1, ..., K$, $m_k(\mathbf{x})$ is twice continuously differentiable. Its first derivative is bounded uniformly over $\mathbf{x} \in \mathcal{X}$ and $\tau_k \in \Omega$.*

6. *The kernel K has compact support and two continuous derivatives, and satisfies $\int uK(u)\,du = 0$ and $\int K(u)\,du = 1$.*

7. *The bandwidths, h_0 and h_1, satisfy that $nh_0^{L+2} \to \infty$, $nh_0^{L+2} \to \infty$, $nh_0^{L+4} \to 0$, and $nh_0^{L+4} \to 0$ as $n \to \infty$.*

Let $\hat{\epsilon}_{k,i} = I_i(\tau_k) - \hat{m}_k^1(\mathbf{x}_i)Z_i - \hat{m}_k^0(\mathbf{x}_i)(1 - Z_i)$ and $\hat{m}_k^*(\mathbf{x})$ be a multiplier process such that

$$\hat{m}_k^*(\mathbf{x}) = \frac{\sum_{Z_i=1} \eta_i \hat{\epsilon}_{k,i} \mathcal{K}_{h_1}(\mathbf{X}_i - \mathbf{x})}{\sum_{Z_i=1} \mathcal{K}_{h_1}(\mathbf{X}_i - \mathbf{x})} - \frac{\sum_{Z_i=0} \eta_i \hat{\epsilon}_{k,i} \mathcal{K}_{h_0}(\mathbf{X}_i - \mathbf{x})}{\sum_{Z_i=0} \mathcal{K}_{h_0}(\mathbf{X}_i - \mathbf{x})}$$

with $\{\eta_i\}_{i=1}^N$ simulated from i.i.d. $N(0,1)$, independent of data.

Define c_α as the $(1-\alpha) \times 100$-th percentile of the simulated process $\sup_{k,\mathbf{x}} \left| \frac{\hat{m}_k^*(\mathbf{x})}{s_k(\mathbf{x})} \right|$.

Corollary 4. *Given Assumptions 2 and 5, the multiplier bootstrap critical value c_α defined above satisfies the condition required in equation* (3)*.*

Consistency of the test under the null follows trivially from the corollary.

7.3 Testing Conditional Ranks

The discussion so far has focused on testing whether individuals' unconditional ranks remain the same with or without treatment. This section considers conditional ranks, which are relevant when one is interested in conditional QTEs or LQTEs. Previous tests can be readily extended to this case. Two main modifications are required. First, one estimates conditional quantiles conditional on some covariates of interest $\mathbf{X}_1$ in the first step. Second, use additional covariates $\mathbf{X}_2$ along with $\mathbf{X}_1$ to perform the test.

Note that additional covariates are needed for the test, other than those included in the conditioning set of conditional quantile estimation. Therefore, this test is feasible only when the conditioning set for the conditional quantiles is small or one has a large set of covariates. Take as an example this paper's empirical application, we estimate quantiles of potential earnings and perform the tests separately for male and female trainees, so the tests are essentially rank similarity tests for conditional ranks conditional on gender. We briefly describe the basic idea below and leave the full exploration for future work.

Let $q_{t|C,\mathbf{X_1}}(\tau|\mathbf{x}_1) = F^{-1}_{t|C,\mathbf{X_1}}(\tau|\mathbf{x}_1)$ for $t = 0, 1$ and $\tau \in (0, 1)$, i.e., $q_{t|C,\mathbf{X_1}}(\tau|\mathbf{x}_1)$ is the conditional quantile function of Y_t conditional on $\mathbf{X}_1 = \mathbf{x}_1$ among compliers. If Assumption 2 holds conditional on $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')'$, following Frolich and Melly (2013), one can identify $q_{t|C,\mathbf{X_1}}(\tau|\mathbf{x}_1)$ for $t = 0, 1$ by

$$\left(q_{0|C,\mathbf{X_1}}(\tau|\mathbf{x}_1), q_{1|C,\mathbf{X_1}}(\tau|\mathbf{x}_1)\right) = \arg\min_{q_0,q_1} E\left[\rho_\tau(Y - q_0(1-T) - q_1 T)\omega^{FM}|\mathbf{X}_1 = \mathbf{x}_1\right], \quad (4)$$

where $\omega^{FM} = \left(\frac{Z}{\pi(\mathbf{X})} - \frac{1-Z}{1-\pi(\mathbf{X})}\right)(2T-1)$ and as before $\pi(\mathbf{X}) = \Pr(Z = 1|\mathbf{X} = \mathbf{x})$ is the instrument probability.

If Assumption 2 holds conditional on $\mathbf{X}_1$, then one may still utilize the above equation (4) for identification. Alternatively, if a linear model for conditional quantiles is assumed, one may follow Abadie, Angrist, and Imbens (2002) and identify the conditional quantiles $\tilde{q}_{t|C}(\tau) \equiv q_{t|C,\mathbf{X_1}}(\tau|.)$ for $t = 0, 1$ by

$$\left(\tilde{q}_{0|C}(\tau), \tilde{q}_{1|C}(\tau)\right) = \arg\min_{q_0,q_1} E\left[\rho_\tau(Y - q_0(1-T) - q_1 T - \mathbf{X}_1'\gamma)\omega^{AAI}\right], \quad (5)$$

where $\omega^{AAI} = 1 - \frac{T(1-Z)}{1-\pi(\mathbf{x}_1)} - \frac{(1-T)Z}{1-\pi(\mathbf{x}_1)}$ and $\pi(\mathbf{x}_1) = \Pr(Z = 1|\mathbf{X} = \mathbf{x}_1)$. To ensure nonnegativity, ω^{FM} and ω^{AAI} can be replaced with their projections onto Y, T and $\mathbf{X}_1$.

Define the rank indicator among those with $\mathbf{X}_1 = \mathbf{x}_1$ for any $\tau \in (0, 1)$ and $\mathbf{x}_1 \in \mathcal{X}_1$, where $\mathcal{X}_1$ is the support of $\mathbf{x}_1$,

$$\tilde{I}(\tau, \mathbf{x}_1) \equiv \mathbf{1}\left(Y \leq \left(T q_{0|C,\mathbf{X_1}}(\tau|\mathbf{x}_1) + (1-T)\, q_{1|C,\mathbf{X_1}}(\tau|\mathbf{x}_1)\right)\right).$$

Let $\mathcal{X}_2$ be the support of $\mathbf{X}_2$ conditional on $\mathbf{X}_1 = \mathbf{x}_1$. Then analogous to Theorem 1, rank similarity for the conditional ranks conditional on $\mathbf{X}_1 = \mathbf{x}_1$ holds if and only if for all $\tau \in (0,1)$ and $\mathbf{x}_2 \in \mathcal{X}_2$,

$$E\left[\tilde{I}(\tau, \mathbf{x}_1)|Z=1, \mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2\right] = E\left[\tilde{I}\left(\tau, \mathbf{x}_1\right)|Z=0, \mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2\right]. \quad (6)$$

One can then test invariance or similarity of the conditional ranks by estimating $q_{t|C,\mathbf{x_1}}(\tau|\mathbf{x}_1)$ for $t = 0, 1$ first and then test whether equation (6) holds for all $\tau \in (0,1)$ and $\mathbf{x_2} \in \mathcal{X}_2$, replacing $q_{t|C,\mathbf{x_1}}(\tau|\mathbf{x}_1)$ with their estimates. Given a sample of i.i.d. data $\{Y_i, T_i, Z_i, \mathbf{X}_i\}_{i=1}^n$, $q_{t|C,\mathbf{x_1}}(\tau|\mathbf{x}_1)$ for $t = 0, 1$ can be estimated by the sample counterparts of equation (4) or (5). That is,

$$\left(\hat{q}_{0|C,\mathbf{X_1}}\left(\tau|\mathbf{x}_1\right), \hat{q}_{1|C,\mathbf{X_1}}\left(\tau|\mathbf{x}_1\right)\right) = \arg\min_{q_0,q_1} \frac{1}{n_1} \sum_{\mathbf{X}_{1i}=\mathbf{x_1}} \rho_\tau(Y_i - q_0(1-T_i) - q_1 T_i)\hat{\omega}_i^{FM},$$

where $n_1 = \sum_{i=1}^n \mathbf{1}\left(\mathbf{X}_{1i} = \mathbf{x}_1\right)$ and $\hat{\omega}_i^{FM} = \left(\frac{Z_i}{\hat{\pi}(\mathbf{X}_i)} - \frac{1-Z_i}{1-\hat{\pi}(\mathbf{X}_i)}\right)(2T_i - 1)$, or

$$\left(\check{q}_{0|C}\left(\tau\right), \check{q}_{1|C}\left(\tau\right)\right) = \arg\min_{q_0,q_1} \frac{1}{n} \sum_{\mathbf{i=1}}^n \rho_\tau(Y_i - q_0(1-T_i) - q_1 T_i - \mathbf{X}_{1i}'\gamma)\hat{\omega}_i^{AAI},$$

where $\hat{\omega}_i^{AAI} = 1 - \frac{T_i(1-Z_i)}{\hat{\pi}(\mathbf{X}_{1i})} - \frac{(1-T_i)Z_i}{1-\hat{\pi}(\mathbf{X}_{1i})}$.

When the IV Z is random assignment of treatment, Assumption 2 would hold without conditioning on any covariates. In this case, $\mathbf{X}_1$ can be the set of covariates one is interested in conditioning on, and $\mathbf{X}_2$ can be the additional covariates used for the test purpose. Either of the above estimators can be used for the first-stage estimation of conditional quantiles.

8 Conclusion

This paper proposes tests for rank invariance or rank similarity that are popular in program evaluation and various quantile treatment effects models. We nonparametrically identify and test the counterfactual distribution of potential ranks (or features of the distribution, such as moments, median or any particular quantile) among observationally equivalent individuals. The tests can be useful in examining whether particular sub-groups of interest change ranks in the outcome distribution under treatment. By testing any particular quantile of the potential rank distribution, the proposed tests are informative regarding at which part of potential outcome distribution rank similarity is violated.

The proposed tests allow treatment to be endogenous, with exogenous treatment following as a special case. The tests can handle instrument variables that are valid regardless of conditioning on covariates or not. Other than mild regularity conditions, the tests do not require any additional assumptions other than those used to identify and estimate the first-stage unconditional QTEs.

We show that with discrete covariates, under standard assumptions, the test statistics asymptotically follow either a Chi-squared distribution or a standard normal distribution. A Kolmogorov-Smirnov type test can be conducted with continuous covariates. When conditional QTEs are of interest, the proposed tests can be extended to test invariance or similarity of ranks in the conditional distribution of potential outcomes.

Simulation studies show good size and power of the proposed tests in small samples. We empirically apply the proposed tests to investigate the JTPA program, a large publically funded training program and to analyze Project STAR, a large-scale randomized experiment on class size in the US. We show that training causes individuals to systematically change their ranks in the earnings distribution. While male trainees change ranks throughout the earnings distribution, female trainees change ranks only at the lower tail of the distribution. Overall evidence suggests that the impact of training on earnings are more complicated than what would be suggested by the standard QTEs. The estimated QTEs of the JTPA training program therefore at best reflect the impacts of training on the distribution of trainee earnings instead of those on individual trainees. For Project STAR, we find that attending a small class or having a teacher aide is beneficial for boys and narrows the gender gap in test scores. We also find that small classes greatly narrow the achievement gaps among students taught by teachers with different levels of experience. Those students who are assigned to teachers with 6-10 years' experience show the greatest improvement.

We focus on testing for invariance or similarity of individual ranks in the unconditional distributions of potential outcomes. The proposed test can be extended to test for invariance or similarity of conditional ranks when additional covariates (other than those used in the first-stage conditional QTE estimation) are available for testing. The JTPA empirical application exemplifies this case – essentially we condition on gender to estimate QTEs and to further test for rank invariance or similarity among male or female trainees separately. An interesting future direction of research is then to develop further tests for invariance or similarity of conditional ranks. It might also be interesting to test stochastic dominance, in addition to equivalence, of the conditional distribution of potential ranks across treatment states. We leave both to future work.

Appendix

Proof of Lemma 1

Part 1 of the Lemma: by Bayes' Rule,

$$f_{\mathbf{X},V|U_t}(\mathbf{x},v|\tau) = \frac{f_{U_t|\mathbf{X},V}(\tau|\mathbf{x},v)f_{\mathbf{X},V}(\mathbf{x},v)}{\iint\limits_{\mathcal{W}} f_{U_t|\mathbf{X},V}(\tau|\mathbf{x},v)f_{\mathbf{X},V}(\mathbf{x},v)d\mathbf{x}dv} \text{for } t = 0,1,\ \tau \in (0,1) \text{ and } (\mathbf{x},v) \in \mathcal{W}.$$

Rank similarity is defined as $F_{U_0|\mathbf{X},V}(\tau|\mathbf{x},v) = F_{U_1|\mathbf{X},V}(\tau|\mathbf{x},v)$. Immediately $f_{U_t|\mathbf{X},V}(\tau|\mathbf{x},v) \equiv f_{U|\mathbf{X},V}(\tau|\mathbf{x},v)$ for $t = 0,1$. Then $f_{\mathbf{X},V|U_t}(\mathbf{x},v|\tau) \equiv f_{\mathbf{X},V|U}(\mathbf{x},v|\tau)$ for $t = 0,1$ and hence $F_{\mathbf{X},V|U_1}(\mathbf{x},v|\tau) \equiv F_{\mathbf{X},V|U_0}(\mathbf{x},v|\tau)$ for any $\tau \in (0,1)$ and $(\mathbf{x},v) \in \mathcal{W}$.

Part 2 of the Lemma: For both $t = 0,1$ and all values of $(\mathbf{x},v) \in \mathcal{W}$, $E[Y_t|\mathbf{X} = \mathbf{x}, V = v] = E[q_t(U_t)|\mathbf{X} = \mathbf{x}, V = v] = \int_0^1 q_t(\tau)dF_{U_t|\mathbf{X},V}(\tau|\mathbf{x}, v) = \int_0^1 q_t(\tau)dF_{U|\mathbf{X},V}(\tau|\mathbf{x}, v)$, where the last equality follows from the definition of rank similarity. Therefore, $E[Y_1 - Y_0|\mathbf{X} = \mathbf{x}, V =$

$v] = \int_0^1 (q_1(\tau) - q_0(\tau))\, dF_{U|\mathbf{X},V}(\tau|\mathbf{x}, v)$ for any $\tau \in (0,1)$ and $(\mathbf{x},v) \in \mathcal{W}$.

Part 3 of the Lemma: $f_{U_t|\mathbf{X}}(\tau|\mathbf{x}) = \int_{Supp(V|\mathbf{X})} f_{U_0|\mathbf{X},V}(\tau|\mathbf{x},v) dF_{V|\mathbf{X}}(v|\mathbf{x})$ for $t = 0, 1$, so $f_{U_0|\mathbf{X},V}(\tau|\mathbf{x},v) = f_{U_1|\mathbf{X},V}(\tau|\mathbf{x},v)$ implies $f_{U_0|\mathbf{X}}(\tau|\mathbf{x}) = f_{U_1|\mathbf{X}}(\tau|\mathbf{x})$ and hence $F_{U_0|\mathbf{X}}(\tau|\mathbf{x}) = F_{U_1|\mathbf{X}}(\tau|\mathbf{x})$ for any $\tau \in (0,1)$ and $(\mathbf{x},v) \in \mathcal{W}$.

Proof of Lemma 2

Lemma 1 shows that rank similarity $F_{U_0|\mathbf{X},V}(\tau|\mathbf{x},v) = F_{U_1|\mathbf{X},V}(\tau|\mathbf{x},v)$ implies $F_{U_0|\mathbf{X}}(\tau|\mathbf{x}) = F_{U_1|\mathbf{X}}(\tau|\mathbf{x})$. The following proves that $F_{U_0|\mathbf{X}}(\tau|\mathbf{x}) = F_{U_1|\mathbf{X}}(\tau|\mathbf{x})$ implies $F_{U_0|\mathbf{X},V}(\tau|\mathbf{x},v) = F_{U_1|\mathbf{X},V}(\tau|\mathbf{x},v)$ given Assumption 1.

Notice that $F_{U_0|\mathbf{X}}(\tau|\mathbf{x}) = F_{U_1|\mathbf{X}}(\tau|\mathbf{x})$ if and only if $f_{U_0|\mathbf{X}}(\tau|\mathbf{x}) = f_{U_1|\mathbf{X}}(\tau|\mathbf{x})$. Further, Assumption 1 implies that $f_{\mathbf{X},V|U_0}(\mathbf{x},v|\tau) = f_{\mathbf{X},V|U_1}(\mathbf{x},v|\tau)$. By Bayes' rule

$$f_{U_t|\mathbf{X},V}(\tau|\mathbf{x},v) = \frac{f_{\mathbf{X},V|U_t}(\mathbf{x},v|\tau) f_{U_t}(\tau)}{f_{\mathbf{X},V}(\mathbf{x},v)} = \frac{f_{\mathbf{X},V|U_t}(\mathbf{x},v|\tau) \int_{\mathcal{X}} f_{U_t|\mathbf{X}}(\tau|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}}{f_{\mathbf{X},V}(\mathbf{x},v)}.$$

Therefore, $f_{U_1|\mathbf{X},V}(\tau|\mathbf{x},v) = f_{U_0|\mathbf{X},V}(\tau|\mathbf{x},v)$. It follows that $F_{U_1|\mathbf{X},V}(\tau|\mathbf{x},v) = F_{U_0|\mathbf{X},V}(\tau|\mathbf{x},v)$.

Proof of Theorem 1

Identification of $F_{U_t|C,\mathbf{X}}(\tau|\mathbf{x})$ for $t = 0, 1$, $\tau \in (0,1)$ and $\mathbf{x} \in \mathcal{X}_C$:

Given Assumption 2, and for $\mathbf{x} \in \mathcal{X}_C$ $E(T|Z = 1, \mathbf{X} = \mathbf{x}) - E(T|Z = 0, \mathbf{X} = \mathbf{x}) \neq 0$, following the standard LATE arguments (see, e.g. Abadie 2003), $F_{U_t|C,\mathbf{X}}(\tau|\mathbf{x})$ for $t = 0, 1$ are identified as follows

$$\begin{aligned} & F_{U_1|C,\mathbf{X}}(\tau|\mathbf{x}) = E\left[\mathbf{1}(U_1 \le \tau)|C, \mathbf{X} = \mathbf{x}\right] \\ = \; & E\left[\mathbf{1}(Y_1 \le q_{1|C}(\tau))|C, \mathbf{X} = \mathbf{x}\right] \\ = \; & \frac{E\left[\mathbf{1}(Y \le q_{1|C}(\tau))T|Z = 1, \mathbf{X} = \mathbf{x}\right] - E\left[\mathbf{1}(Y \le q_{1|C}(\tau))T|Z = 0, \mathbf{X} = \mathbf{x}\right]}{E\left[T|Z = 1, \mathbf{X} = \mathbf{x}\right] - E\left[T|Z = 0, \mathbf{X} = \mathbf{x}\right]}, \end{aligned} \tag{7}$$

and

$$
\begin{aligned}
& F_{U_0|C,\mathbf{X}}(\tau|\mathbf{x}) = E\left[\mathbf{1}(U_0 \leq \tau)|C, \mathbf{X}=\mathbf{x}\right] \\
= & \ E\left[\mathbf{1}(Y_0 \leq q_{0|C}(\tau))|C, \mathbf{X}=\mathbf{x}\right] \\
= & \ \frac{E\left[\mathbf{1}(Y \leq q_{0|C}(\tau))(1-T)|Z=1, \mathbf{X}=\mathbf{x}\right] - E\left[\mathbf{1}(Y \leq q_{0|C}(\tau))(1-T)|Z=0, \mathbf{X}=\mathbf{x}\right]}{E\left[1-T|Z=1, \mathbf{X}=\mathbf{x}\right] - E\left[1-T|Z=0, \mathbf{X}=\mathbf{x}\right]}
\end{aligned}
\tag{8}
$$

Recall $I(\tau) \equiv \mathbf{1}\left(Y \leq Tq_{1|C}(\tau) + (1-T)q_{0|C}(\tau)\right)$. The above equations can be re-written as

$$
\begin{aligned}
& F_{U_t|C,\mathbf{X}}(\tau|\mathbf{x}) \\
= & \ \frac{E\left[I(\tau)\mathbf{1}(T=t)|Z=1, \mathbf{X}=\mathbf{x}\right] - E\left[I(\tau)\mathbf{1}(T=t)|Z=0, \mathbf{X}=\mathbf{x}\right]}{E\left[\mathbf{1}(T=t)|Z=1, \mathbf{X}=\mathbf{x}\right] - E\left[\mathbf{1}(T=t)|Z=0, \mathbf{X}=\mathbf{x}\right]} \text{ for } t=0,1.
\end{aligned}
$$

Derivation of equation (2):

Case 1: for any $\mathbf{x} \in \mathcal{X}_C$, $E\left[T|Z=1, \mathbf{X}=\mathbf{x}\right] - E\left[T|Z=0, \mathbf{X}=\mathbf{x}\right] \neq 0$. Equations (7) and (8) yield

$$
\begin{aligned}
& F_{U_1|C,\mathbf{X}}(\tau|\mathbf{x}) - F_{U_0|C,\mathbf{X}}(\tau|\mathbf{x}) \\
= & \ \frac{E\left[I(\tau)|Z=1, \mathbf{X}=\mathbf{x}\right] - E\left[I(\tau)|Z=0, \mathbf{X}=\mathbf{x}\right]}{E\left[T|Z=1, \mathbf{X}=\mathbf{x}\right] - E\left[T|Z=0, \mathbf{X}=\mathbf{x}\right]} = 0.
\end{aligned}
$$

Alternatively, first note that rank similarity implies $F_{U_0|C,\mathbf{X}}(\tau|.) = F_{U_1|C,\mathbf{X}}(\tau|.)$ for $\mathbf{x} \in \mathcal{X}_C$. It is equivalent to

$$
\begin{aligned}
& E\left[\mathbf{1}(U_1 \leq \tau)|C, \mathbf{X}=\mathbf{x}\right] - E\left[\mathbf{1}(U_0 \leq \tau)|C, \mathbf{X}=\mathbf{x}\right] \\
= & \ E\left[\mathbf{1}(Y_1 \leq q_{1|C}(\tau))|T_1 > T_0, \mathbf{X}=\mathbf{x}\right] - E\left[\mathbf{1}(Y_0 \leq q_{0|C}(\tau))|T_1 > T_0, \mathbf{X}=\mathbf{x}\right] \\
= & \ E\left[\mathbf{1}(Y_1 \leq q_{1|C}(\tau))|T_1 > T_0, Z=1, \mathbf{X}=\mathbf{x}\right] - E\left[\mathbf{1}(Y_0 \leq q_{0|C}(\tau))|T_1 > T_0, Z=0, \mathbf{X}=\mathbf{x}\right] \\
= & \ E\left[\mathbf{1}(Y_1 \leq q_{1|C}(\tau))|T_1 > T_0, T=1, \mathbf{X}=\mathbf{x}\right] - E\left[\mathbf{1}(Y_0 \leq q_{0|C}(\tau))|T_1 > T_0, T=0, \mathbf{X}=\mathbf{x}\right] \\
= & \ E\left[\mathbf{1}(Y \leq q_{1|C}(\tau))|C, T=1, \mathbf{X}=\mathbf{x}\right] - E\left[\mathbf{1}(Y \leq q_{0|C}(\tau))|C, T=0, \mathbf{X}=\mathbf{x}\right] \\
= & \ E\left[I(\tau)|C, T=1, \mathbf{X}=\mathbf{x}\right] - E\left[I(\tau)|C, T=0, \mathbf{X}=\mathbf{x}\right] = 0,
\end{aligned}
$$

where the second equality follows from the definition of compliers, the third equality follows from Assumption 1.1 Y_1, Y_1, T_1, $T_0 \perp Z|\mathbf{X}$, and the fourth equality follows from the fact that $T = T_1 Z + T_0 (1 - Z) = Z$ for compliers.

Given Assumption 2, for $\mathbf{x} \in \mathcal{X}_C$, $E(I(\tau)|C, T = 1, \mathbf{X} = \mathbf{x}) - E(I(\tau)|C, T = 0, \mathbf{X} = \mathbf{x})$ can be identified as the ratio of two intention-to-treat estimands (Imbens and Angrist, 1994), i.e.,

$$\begin{aligned} & E(I(\tau)|C, T = 1, \mathbf{X} = \mathbf{x}) - E(I(\tau)|C, T = 0, \mathbf{X} = \mathbf{x}) \\ = \; & \frac{E(I(\tau)|Z = 1, \mathbf{X} = \mathbf{x}) - E(I(\tau)|Z = 0, \mathbf{X} = \mathbf{x})}{E(T|Z = 1, \mathbf{X} = \mathbf{x}) - E(T|Z = 0, \mathbf{X} = \mathbf{x})} = 0. \end{aligned}$$

Case 2: for $\mathbf{x} \in \mathcal{X}/\mathcal{X}_C$, $T_0 = T_1$ by Assumption 2.3. Further

$$\begin{aligned} E[I(\tau)|Z = z, \mathbf{X} = \mathbf{x}] &= E[\mathbf{1}(Y \le q_1(\tau))T + \mathbf{1}(Y \le q_0(\tau))(1 - T)|Z = z, \mathbf{X} = \mathbf{x}] \\ &= E[\mathbf{1}(Y_1 \le q_1(\tau))T_z + \mathbf{1}(Y_0 \le q_0(\tau))(1 - T_z)|Z = z, \mathbf{X} = \mathbf{x}] \\ &= E[\mathbf{1}(Y_1 \le q_1(\tau))T_z + \mathbf{1}(Y_0 \le q_0(\tau))(1 - T_z)|\mathbf{X} = \mathbf{x}], \end{aligned}$$

where the second equality holds by the definition of Y and T_0, T_1 and the third equality holds by Assumption 2.1. $T_0 = T_1$ then means that Equation (2) holds trivially for $\mathbf{x} \in \mathcal{X}/\mathcal{X}_C$.

Therefore, given any $\tau \in (0, 1)$, $F_{U_0|C,\mathbf{X}}(\tau|.) = F_{U_1|C,\mathbf{X}}(\tau|.)$ for $\mathbf{x} \in \mathcal{X}_C$ holds *if and only if* for $\mathbf{x} \in \mathcal{X}$,

$$E[I(\tau)|Z = 1, \mathbf{X} = \mathbf{x}] - E[I(\tau)|Z = 0, \mathbf{X} = \mathbf{x}] = 0.$$

Proof of Theorem 2

Proof. Let $\hat{p}_j^z = \frac{n_j^z}{n}$ be the nonparametric estimator of $p_j^z = P_{Z,X}(z, \mathbf{x}_j)$. The nonparametric estimator

$$\begin{aligned}\hat{m}_j^z(\tau_k) =& \frac{1}{n}\sum_i^n \mathbf{1}\left(Y_i \le \hat{q}_{1|C}(\tau_k), T_i = 1, Z_i = z, \mathbf{X}_i = \mathbf{x}_j\right)/\hat{p}_j^z \\ &+ \frac{1}{n}\sum_i^n \mathbf{1}\left(Y_i \le \hat{q}_{0|C}(\tau_k), T_i = 0, Z_i = z, \mathbf{X}_i = \mathbf{x}_j\right)/\hat{p}_j^z.\end{aligned}$$

It's population counterpart is

$$m_j^z(\tau_k) = F_{Y,T|Z,\mathbf{X}}(q_{1|C}(\tau_k), 1|z, \mathbf{x}_j) + F_{Y,T|Z,\mathbf{X}}(q_{0|C}(\tau_k), 0|z, \mathbf{x}_j).$$

Let $\check{m}_j^z(\tau_k) = \frac{1}{n}\sum_i^n \mathbf{1}\left(Y_i \le \left(T_i q_{1|C}(\tau_k) + (1-T_i)q_{0|C}(\tau_k)\right), Z_i = z, \mathbf{X}_i = \mathbf{x}_j\right)/p_j^z$ and $\tilde{m}_j^z(\tau_k) = F_{Y,T|Z,\mathbf{X}}(\hat{q}_{1|C}(\tau_k), 1|z, \mathbf{x}_j) + F_{Y,T|Z,\mathbf{X}}(\hat{q}_{0|C}(\tau_k), 0|z, \mathbf{x}_j)$. Decompose $\sqrt{n}\left(\hat{m}_j^z(\tau_k) - m_j^z(\tau_k)\right)$ so that

$$\begin{aligned}\sqrt{n}\left(\hat{m}_j^z(\tau_k) - m_j^z(\tau_k)\right) &= \sqrt{n}\left(\hat{m}_j^z(\tau_k) - \check{m}_j^z(\tau_k)\right) + \sqrt{n}\left(\check{m}_j^z(\tau_k) - \tilde{m}_j^z(\tau_k)\right) + \sqrt{n}\left(\tilde{m}_j^z(\tau_k) - m_j^z(\tau_k)\right) \\ &= I + II + III.\end{aligned}$$

First, we can show that for all $z = 0, 1$, $j = 1, ..., J-1$, and $k = 1, ..., K$,

$$\begin{aligned}
I = \sqrt{n}\left(\hat{m}_j^z(\tau_k) - \check{m}_j^z(\tau_k)\right) =& \frac{1}{n}\sum_i^n \mathbf{1}\left(Y_i \leq \hat{q}_{1|C}(\tau_k), T_i = 1, Z_i = z, \mathbf{X}_i = \mathbf{x}_j\right) \frac{-\sqrt{n}(\hat{p}_j^z - p_j^z)}{\hat{p}_j^z p_j^z} \\
&+ \frac{1}{n}\sum_i^n \mathbf{1}\left(Y_i \leq \hat{q}_{0|C}(\tau_k), T_i = 0, Z_i = z, \mathbf{X}_i = \mathbf{x}_j\right) \frac{-\sqrt{n}(\hat{p}_j^z - p_j^z)}{\hat{p}_j^z p_j^z} \\
=& F_{Y,T|Z,\mathbf{X}}(\hat{q}_{1|C}(\tau_k), 1|z, \mathbf{x}_j)\frac{-\sqrt{n}(\hat{p}_j^z - p_j^z)}{p_j^z} \\
&+ F_{Y,T|Z,\mathbf{X}}(\hat{q}_{0|C}(\tau_k), 0|z, \mathbf{x}_j)\frac{-\sqrt{n}(\hat{p}_j^z - p_j^z)}{p_j^z} + o_p(1) \\
=& F_{Y,T|Z,\mathbf{X}}(q_{1|C}(\tau_k), 1|z, \mathbf{x}_j)\frac{-\sqrt{n}(\hat{p}_j^z - p_j^z)}{p_j^z} \\
&+ F_{Y,T|Z,\mathbf{X}}(q_{0|C}(\tau_k), 0|z, \mathbf{x}_j)\frac{-\sqrt{n}(\hat{p}_j^z - p_j^z)}{p_j^z} + o_p(1) \\
=& -\frac{m_j^z(\tau_k)}{p_j^z}\frac{1}{\sqrt{n}}\left(\sum_{i=1}^n \mathbf{1}(Z_i = z, \mathbf{X}_i = \mathbf{x}_j) - p_j^z\right).
\end{aligned}$$

The second equality follows from convergence of $\hat{p}_j^z$, and the fact that for fixed z, t and $\mathbf{x}_j$ the class of indicator functions $g_y = \mathbf{1}\left(Y_i \leq y, T_i = t, Z_i = z, \mathbf{X}_i = \mathbf{x}_j\right)$ is Glivenko-Cantelli so that $\left|\frac{1}{n}\sum_i^n \mathbf{1}\left(Y_i \leq \hat{q}_{t|C}(\tau_k), T_i = t, Z_i = z, \mathbf{X}_i = \mathbf{x}_j\right) - F_{Y,T,Z,\mathbf{X}}(\hat{q}_{t|C}(\tau_k), t, z, \mathbf{x}_j)\right| \overset{a.s.}{\to} 0$. The third equality follows from convergence of $\hat{q}_{t|C}(\tau_k)$, continuity of $F_{Y,T|Z,\mathbf{X}}(., t|z, \mathbf{x}_j)$ in a neighborhood of $q_{t|C}(\tau_k)$, and the Continuous Mapping Theorem.

Next, notice that due to consistency of $\hat{q}_{t|C}(\tau_k)$, continuity of $F_{Y,T,Z,\mathbf{X}}(., t, z, \mathbf{x}_j)$ in a neighborhood of $q_{t|C}(\tau_k)$, and the Continuous Mapping Theorem, we can show that for all

$z = 0, 1,\ j = 1, ..., J-1$, and $k = 1, ..., K$,

$$\begin{aligned}
&\int \left[\mathbf{1}(Y \leq \hat{q}_{1|C}(\tau_k), Z = z, \mathbf{X} = \mathbf{x}_j) - \mathbf{1}(Y \leq q_{1|C}(\tau_k), Z = z, \mathbf{X} = \mathbf{x}_j)\right]^2 dF(Y, 1, Z, \mathbf{X}) \\
&+ \int \left[\mathbf{1}(Y \leq \hat{q}_{0|C}(\tau_k), Z = z, \mathbf{X} = \mathbf{x}_j) - \mathbf{1}(Y \leq q_{0|C}(\tau_k), Z = z, \mathbf{X} = \mathbf{x}_j)\right]^2 dF(Y, 0, Z, \mathbf{X}) \\
=& F(\hat{q}_{1|C}(\tau_k), 1, z, \mathbf{x}_j) + F(q_{1|C}(\tau_k), 1, z, \mathbf{x}_j) - 2F(\min(\hat{q}_{1|C}(\tau_{kk}), q_{1|C}(\tau_k)), 1, z, \mathbf{x}_j) \\
&+ F(\hat{q}_{0|C}(\tau_k), 0, z, \mathbf{x}_j) + F(q_{0|C}(\tau_k), 0, z, \mathbf{x}_j) - 2F(\min(\hat{q}_{0|C}(\tau_{kk}), q_{0|C}(\tau_k)), 0, z, \mathbf{x}_j) \\
\xrightarrow{p}& 0.
\end{aligned}$$

Since for fixed z, t and $\mathbf{x}_j$ the class of indicator functions $g_y = \mathbf{1}\left(Y_i \leq y, T_i = t, Z_i = z, \mathbf{X}_i = \mathbf{x}_j\right)$ is also Donsker, by Lemma 19.24 of Van der Vaart (1998),

$$II = \sqrt{n}\left(\check{m}_j^z(\tau_k) - \tilde{m}_j^z(\tau_k)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{I_i(\tau_k)\mathbf{1}(Z_i = z, \mathbf{X}_i = \mathbf{X}_j)}{p_j^z} - m_j^z(\tau_k)\right] + o_p(1).$$

Last, by boundedness of the second derivative of $F_{Y|T,Z,\mathbf{X}}(y|t, z, \mathbf{x})$ with respect to y in a neighborhood of $q_{t|C}(\tau_k)$, we have that

$$\begin{aligned}
III =& \sqrt{n}\left(\tilde{m}_j^z(\tau_k) - m_j^z(\tau_k)\right) \\
=& \sqrt{n}\left(F_{Y,T|Z,\mathbf{X}}(\hat{q}_{1|C}(\tau_k), 1|z, \mathbf{x}_j) + F_{Y,T|Z,\mathbf{X}}(\hat{q}_{0|C}(\tau_k), 0|z, \mathbf{x}_j)\right) \\
&- \sqrt{n}\left(F_{Y,T|Z,\mathbf{X}}(q_{1|C}(\tau_k), 1|z, \mathbf{x}_j) - F_{Y,T|Z,\mathbf{X}}(q_{0|C}(\tau_k), 0|z, \mathbf{x}_j)\right) \\
=& f_{Y|T,Z,\mathbf{X}}(q_{0|C}(\tau_k)|0, z, \mathbf{x}_j)\left(1 - p_{T|Z,\mathbf{X}}(z, \mathbf{x}_j)\right)\sqrt{n}\left(\hat{q}_{0|C}(\tau_k) - q_{0|C}(\tau_k)\right) \\
&+ f_{Y|T,Z,\mathbf{X}}(q_{1|C}(\tau_k)|1, z, \mathbf{x}_j) p_{T|Z,\mathbf{X}}(z, \mathbf{x}_j)\sqrt{n}\left(\hat{q}_{1|C}(\tau_k) - q_{1|C}(\tau_k)\right) + o_p(1).
\end{aligned}$$

By Frolich and Melly (2013), we have

$$\sqrt{n}\left(\hat{q}_{t|C}(\tau_k) - q_{t|C}(\tau_k)\right) = -\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi_t(Y_i, T_i, Z_i, \mathbf{X}_i)}{P_c f_{t|C}(q_{t|C}(\tau_k))} + o_p(1)$$

where for $p_i = \Pr(Z_i = 1)$,

$$
\begin{aligned}
\psi_t(Y_i, T_i, Z_i, \mathbf{X}_i) = & \frac{Z_i T_i}{p_i}\left(\mathbf{1}(Y_i \leq q_{t|C}(\tau_k) - \tau_k\right) - E\left[\mathbf{1}(Y \leq q_{t|C}(\tau_k) - \tau_k)|T = t, Z = 1, \mathbf{X} = \mathbf{X}_i\right]) \\
& - \frac{(1 - Z_i)T_i}{1 - p_i}\left(\mathbf{1}(Y_i \leq q_{t|C}(\tau_k) - \tau_k\right) - E\left[\mathbf{1}(Y \leq q_{t|C}(\tau_k) - \tau_k)|T = t, Z = 0, \mathbf{X} = \mathbf{X}_i\right]) \\
& + \frac{Z_i T_i - E[T|\mathbf{X} = \mathbf{X}_i, Z = 1](Z_i - p_i)}{p_i} E\left[\mathbf{1}\left(Y \leq q_{t|C}(\tau_k)\right) - \tau_k|T = t, Z = 1, \mathbf{X} = \mathbf{X}_i\right] \\
& - \frac{(1 - Z_i)T_i - E[T|\mathbf{X} = \mathbf{X}_i, Z = 0](Z_i - p_i)}{1 - p_i} E\left[\mathbf{1}\left(Y \leq q_{t|C}(\tau_k)\right) - \tau_k|T = t, Z = 0, \mathbf{X} = \mathbf{X}_i\right].
\end{aligned}
$$

Combining all the results yields

$$
\begin{aligned}
\sqrt{n}\left(\hat{m}_j^z(\tau_k) - m_j^z(\tau_k)\right) = & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\left(I_i(\tau_k) - m_j^z(\tau_k)\right)\mathbf{1}(Z_i = z, \mathbf{X}_i = \mathbf{x}_j)}{p_j^z}\right] \\
& - f_{Y|T,Z,\mathbf{X}}(q_{0|C}(\tau_k)|0, z, \mathbf{x}_j)\left(1 - p_{T|Z,\mathbf{X}}(z, \mathbf{x}_j)\right)\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_0(Y_i, T_i, Z_i, \mathbf{X}_i)}{P_c f_{0|C}(q_{0|C}(\tau_k))} \\
& - f_{Y|T,Z,\mathbf{X}}(q_{1|C}(\tau_k)|1, z, \mathbf{x}_j) p_{T|Z,\mathbf{X}}(z, \mathbf{x}_j)\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_1(Y_i, T_i, Z_i, \mathbf{X}_i)}{P_c f_{1|C}(q_{1|C}(\tau_k))} + o_p(1) \\
= & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_j^z(\tau_k, Y_i, T_i, Z_i, \mathbf{X}_i) + o_p(1)
\end{aligned}
$$

The theorem is then proven by applying the Central Limit Theorem. ∎

Proof of Corollary 1

Proof. Under H_0, $\mathbf{m}^1 = \mathbf{m}^0$. So the test statistic

$$
\begin{aligned}
W &= n\left((\hat{\mathbf{m}}^1 - \hat{\mathbf{m}}^0) - (\mathbf{m}^1 - \mathbf{m}^0)\right)' \hat{\mathbf{V}}^{-1}\left((\hat{\mathbf{m}}^1 - \hat{\mathbf{m}}^0) - (\mathbf{m}^1 - \mathbf{m}^0)\right) \\
&\Rightarrow \chi^2(K(J-1))
\end{aligned}
$$

as $n \to \infty$. The convergence result follows from Theorem 2 and the fact the $\hat{\mathbf{V}}$ is a consistent estimator of $\mathbf{V}$. Therefore when the null is true,

$$
P(\textit{reject the null}) = P(W > c_\alpha) \to \alpha.
$$

Under the alternative, $\mathbf{m}^1 - \mathbf{m}^0 = A$, which is a $K(J-1) \times 1$ vector of constants that are not all zero. Then

$$\begin{aligned} W &= n\left((\hat{\mathbf{m}}^1 - \hat{\mathbf{m}}^0) - (\mathbf{m}^1 - \mathbf{m}^0) + A\right)' \hat{\mathbf{V}}^{-1} \left((\hat{\mathbf{m}}^1 - \hat{\mathbf{m}}^0) - (\mathbf{m}^1 - \mathbf{m}^0) + A\right) \\ &\to \infty \end{aligned}$$

as $n \to \infty$. Therefore when the null is not true,

$$P(\textit{reject the null}) = P(W > c_\alpha) \to 1.$$

∎

Proof of Corollary 2

Proof. For any fixed S, since $\hat{q}_{t|C}(\tau) \xrightarrow{p} q_{t|C}(\tau)$ for both $t = 0, 1$ and all $\tau \in (0,1)$,

$$\begin{aligned} \hat{\mathsf{R}}(y,t) &= \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\left(\left(t\hat{q}_{1|C}(\tau^s) + (1-t)\hat{q}_{0|C}(\tau^s)\right) \le y\right) \\ &\xrightarrow{p} \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}\left(\left(tq_{1|C}(\tau^s) + (1-t)q_{0|C}(\tau^s)\right) \le y\right). \end{aligned}$$

Then $\hat{\mathsf{R}}(y,t) \xrightarrow{p} \mathsf{R}(y,t)$ as $S, n \to \infty$ by the Law of Large Numbers.

To show the weak convergence result stated in the corollary, notice that

$$
\begin{aligned}
\sqrt{n}\left(\ddot{m}_j^1 - \ddot{m}_j^0\right) =& \sqrt{n}\left(\frac{1}{S}\sum_{s=1}^{S}\left(1-\hat{m}_j^1(\tau^s)\right) - \left(1-\hat{m}_j^0(\tau^s)\right)\right) \\
=& \frac{1}{S}\sum_{s=1}^{S}\sqrt{n}\left(\hat{m}_j^0(\tau^s) - \hat{m}_j^1(\tau^s)\right) \\
=& \int_0^1 \sqrt{n}\left(\hat{m}_j^0(\tau) - \hat{m}_j^1(\tau)\right) d\tau \\
& + \frac{1}{S}\sum_{s=1}^{S}\sqrt{n}\left(\hat{m}_j^0(\tau^s) - \hat{m}_j^1(\tau^s)\right) - \int_0^1 \sqrt{n}\left(\hat{m}_j^0(\tau) - \hat{m}_j^1(\tau)\right) d\tau \\
=& I + II.
\end{aligned}
$$

Since under the null hypothesis, $\bar{m}_j^1 - \bar{m}_j^0 = 0$ for all $j = 1, ..., J-1$,

$$
\begin{aligned}
I &= \int_0^1 \sqrt{n}\left(\hat{m}_j^0(\tau) - m_j^0(\tau) - \left(\hat{m}_j^1(\tau) - m_j^1(\tau)\right)\right) d\tau \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^1 \left(\phi_j^0(\tau, Y_i, T_i, Z_i, \mathbf{X}_i) - \phi_j^1(\tau, Y_i, T_i, Z_i, \mathbf{X}_i)\right) d\tau + o_p(1).
\end{aligned}
$$

The second equality follows from Theorem 2 and the Dominated Convergence Theorem.

Meanwhile, under the null hypothesis

$$
\begin{aligned}
E(II) = E[I] \xrightarrow{p} 0, \\
Var(II) = \frac{1}{S}Var(I) \xrightarrow{p} 0,
\end{aligned}
$$

as $S, N \to \infty$. By Chebyshev's Inequality, we have $II = o_p(1)$.

Since

$$
\int_0^1 I(\tau)d\tau = 1 - U, \text{ and } \int_0^1 m_j^z(\tau)d\tau = 1 - \bar{m}_j^z,
$$

plugging the results into the $\phi_j^z(\tau, Y_i, T_i, Z_i, \mathbf{X}_i)$ functions defined in Theorem 2 and applying the central limit theorem proves the result in Corollary 2. ∎

Proof for Corollary 3

Proof. Let $\tilde{m}_j^z(\tau_k) = E\left[\mathbf{1}\left(Y_i \leq T_i\hat{q}_{1|C}(\tau_k) + (1-T_i)\, Y_i \leq \hat{q}_{0|C}(\tau_k)\right)|Z_i = z, \mathbf{X}_i = \mathbf{x}_j\right]$. First we have

$$\begin{aligned}
\sqrt{n_j^z}\left(\hat{m}_j^z(\tau_k) - m_j^z(\tau_k)\right) &= \sqrt{n_j^z}\left(\hat{m}_j^z(\tau_k) - \tilde{m}_j^z(\tau_k)\right) + \sqrt{n_j^z}\left(\tilde{m}_j^z(\tau_k) - m_j^z(\tau_k)\right) \\
&= I + II.
\end{aligned}$$

First notice that by the consistency of $\hat{q}_{1|C}(\tau_k)$ and $\hat{q}_{0|C}(\tau_k)$ and the continuous mapping theorem, for all $z = 0, 1$, $k = 1, ..., K$, and $j = 1, ..., J-1$,

$$\begin{aligned}
&\int \left[\mathbf{1}\left(Y \leq \hat{q}_{1|C}(\tau_k)\right) - \mathbf{1}\left(Y \leq q_{1|C}(\tau_k)\right)\right]^2 dF_{Y,T|Z,\mathbf{X}}(Y, 1|z, \mathbf{x}_j) \\
&+ \int \left[\mathbf{1}\left(Y \leq \hat{q}_{0|C}(\tau_k)\right) - \mathbf{1}\left(Y \leq q_{0|C}(\tau_k)\right)\right]^2 dF_{Y,T|Z,\mathbf{X}}(Y, 0|z, \mathbf{x}_j) \\
&\xrightarrow{p} 0
\end{aligned}$$

Similar to the arguments in the proof of 2 and by Lemma 19.24 of Van der Vaart (1998), we have that for all $z = 0, 1$, $k = 1, ..., K$, and $j = 1, ..., J-1$,

$$I = \frac{1}{\sqrt{n_j^z}} \sum_{Z_i = z, \mathbf{X}_i = \mathbf{x}_j} \left(I_i(\tau_k) - m_j^z(\tau_k)\right) + o_p(1).$$

Meanwhile, for all $z = 0, 1$, $k = 1, ..., K$, and $j = 1, ..., J - 1$,

$$
\begin{aligned}
II =& \sqrt{n_j^z} E\left[\mathbf{1}\left(Y_i \le T_i \hat{q}_{1|C}(\tau_k) + (1 - T_i)\, \hat{q}_{0|C}(\tau_k)\right) | Z_i = z, \mathbf{X}_i = \mathbf{x}_j)\right] \\
&- \sqrt{n_j^z} E\left[\mathbf{1}\left(Y_i \le T_i q_{1|C}(\tau_k) + (1 - T_i)\, q_{0|C}(\tau_k)\right) | Z_i = z, \mathbf{X}_i = \mathbf{x}_j)\right] \\
=& \sqrt{n_j^z / n} f_{Y_i|T_i, Z_i, \mathbf{X}_i}(\bar{q}_1^k | 1, z, \mathbf{x}_j) P_{T|Z, \mathbf{X}_j}(z, \mathbf{x}_j) \sqrt{n}\left(\hat{q}_{1|C}(\tau_k) - q_{1|C}(\tau_k)\right) \\
&+ \sqrt{n_j^z / n} f_{Y_i|T_i, Z_i, \mathbf{X}_i}(\bar{q}_0^k | 0, z, \mathbf{x}_j)\left(1 - P_{T|Z, \mathbf{X}_j}(z, \mathbf{x}_j)\right) \sqrt{n}\left(\hat{q}_{0|C}(\tau_k) - q_{0|C}(\tau_k)\right) \\
=& o_p(1),
\end{aligned}
$$

where $\bar{q}_t^k$ is a value between $\hat{q}_{t|C}(\tau_k)$ and $q_{t|C}(\tau_k)$ for both $t = 0, 1$. The second equality follows from the Mean Value Theorem.

Now, let $\lambda_{n,j} = n_j^0 / \left(n_j^1 + n_j^0\right)$. Then

$$
\begin{aligned}
& \sqrt{\frac{n_j^1 n_j^0}{n_j^1 + n_j^0}} \left\{\hat{m}_j^1(\tau_k) - \hat{m}_j^0(\tau_k) - \left(m_j^1(\tau_k) - m_j^0(\tau_k)\right)\right\} \\
=& \sqrt{\lambda_{n,j}} \frac{1}{\sqrt{n_j^1}} \sum_{Z_i = 1, \mathbf{X}_i = \mathbf{x}_j} \left(I_i(\tau_k) - m_j^1(\tau_k)\right) - \sqrt{1 - \lambda_{n,j}} \frac{1}{\sqrt{n_j^0}} \sum_{Z_i = 0, \mathbf{X}_i = \mathbf{x}_j} \left(I_i(\tau_k) - m_j^0(\tau_k)\right) \\
&+ o_p(1).
\end{aligned}
$$

For any j, compiling all K quantiles together gives

$$
\begin{aligned}
& \sqrt{\frac{n_j^1 n_j^0}{n_j^1 + n_j^0}} \left\{\hat{\mathsf{m}}_j^1 - \hat{\mathsf{m}}_j^0 - \left(\mathsf{m}_j^1 - \mathsf{m}_j^0\right)\right\} \\
& \Rightarrow \sqrt{1 - \pi(\mathbf{x}_j)} N_1(0, \mathbf{V}_1) + \sqrt{\pi(\mathbf{x}_j)} N_0(0, \mathbf{V}_0),
\end{aligned}
$$

where N_1 and N_0 are two independent normally distributed random variable because the data are i.i.d. and the variance-covariance matrix $\mathbf{V}_1$ and $\mathbf{V}_2$ have (k, k')-th element equal to $m_j^1(\tau_k) - m_j^1(\tau_k)m_j^1(\tau_{k'})$ and $m_j^0(\tau_k) - m_j^0(\tau_k)m_j^0(\tau_{k'})$ respectively. Then we know that

$$\begin{aligned}&\sqrt{\frac{n_j^1 n_j^0}{n_j^1 + n_j^0}} \left\{\hat{\mathsf{m}}_j^1 - \hat{\mathsf{m}}_j^0 - \left(\mathsf{m}_j^1 - \mathsf{m}_j^0\right)\right\} \\ &\Rightarrow N(0, \mathbf{V}_j),\end{aligned}$$

where the (k, k')-th element of $\mathbf{V}_j$ equal to

$$(1 - \pi(\mathbf{x}_j)) \left(m_j^1(\tau_k \wedge \tau_{k'}) - m_j^1(\tau_k)m_j^1(\tau_{k'})\right) + \pi(\mathbf{x}_j) \left(m_j^0(\tau_k \wedge \tau_{k'}) - m_j^0(\tau_k)m_j^0(\tau_{k'})\right).$$

∎

Proof of Corollary 4

Proof. The critical value satisfies equation (3) by the arguments in Example 7 of Chernozhukov, Lee, and Rosen (2011). ∎

References

Abadie, A. (2002): "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97(457), 284–292.

——— (2003): "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, 113, 231–263.

Abadie, A., J. Angrist, and G. Imbens (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70, 91–117.

Achilles, C., H. P. Bain, F. Bellott, J. Boyd-Zaharias, J. Finn, J. Folger, J. Johnston, and E. Word (2008): "Tennessee's Student Teacher Achievement Ratio (STAR) project," *http://hdl.handle.net/1902.1/10766, Harvard Dataverse, V1.*

ANGRIST, J., G. IMBENS, AND D. RUBIN (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91(434), 444–455.

ARMSTRONG, T., AND S. SHEN (2015): "Inference on Optimal Treatment Assignment," *Unpublished manuscript.*

BITLER, M. P., J. B. GELBACH, AND H. W. HOYNES (2006): "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments," *American Economic Review*, 96, 988–1012.

——— (2008): "Distributional Impacts of the Self-Sufficiency Project," *Journal of Public Economics*, 92, 748–765.

BITLER, M. P., AND H. W. HOYNES (2014): "Experimental Evidence on Distributional Effects of Head Start," Unpublished manuscript.

CHERNOZHUKOV, V., AND C. HANSEN (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

——— (2006): "Instrumental Quantile Regression Inference for Structural and Treatment Effect Models," *Journal of Econometrics*, 132, 491–525.

——— (2008): "Instrumental Variable Quantile Regression: A Robust Inference Approach," *Journal of Econometrics*, 142, 379–398.

CHERNOZHUKOV, V., G. IMBENS, AND W. NEWEY (2007): "Instrumental Variable Estimation of Nonseparable Models," *Journal of Econometrics*, 139, 4–14.

CHERNOZHUKOV, V., S. LEE, AND A. ROSEN (2011): "Intersection Bounds: Estimation and Inference," *Working Paper.*

CHESHER, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 14051441.

——— (2005): "Nonparametric Identification under Discrete Variation," *Econometrica*, 73, 15251550.

Chetty, R., J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan (2010): "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR," Discussion paper, National Bureau of Economic Research.

Dammert, A. C. (2008): "Child labor and schooling response to changes in coca production in rural Peru," *Journal of development Economics*, 86(1), 164–180.

de Chaisemartin, C. (2014): "Tolerating defiance? Local average treatment effects without monotonicity.," *Unpublished manuscript*.

Djebbari, H., and J. A. Smith (2008): "Heterogeneous impacts in PROGRESA," *Journal of Econometrics*, 145, 64–80.

Doksum, K. (1974): "Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case," *Annals of Statistics*, 2, 267–277.

Donald, S. G., and Y. Hsu (2014): "Estimation and inference for distribution functions and quantile functions in treatment effect models," *Journal of Econometrics*, 178, 383–397.

Eren, O., and S. Ozbeklik (2014): "Who Benefits from Job Corps? A Distributional Analysis of An Active Lablor Market Program," *Journal of Applied Econometrics*, 29, 586–611.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability*. CRC Press.

Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75, 259–276.

Firpo, S., N. Fortin, and T. Lemieux (2007): "Unconditional Quantile Regressions," *Econometrica*, 77, 953–973.

Frandsen, B. R., and L. J. Lefgren (2015): "A Test for Rank Similarity and Partial Identification of the Distribution of Treatment Effects," *Unpublished manuscript.*

Frolich, M., and B. Melly (2007): "Unconditional Quantile Treatment Effects Under Endogeneity," *Working Paper.*

——— (2013): "Unconditional Quantile Treatment Effects Under Endogeneity," *Journal of Business & Economic Statistics*, 31, 346–357.

Heckman, J., J. Smith, and N. Clements (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487–535.

Horowitz, J., and S. Lee (2007): "Nonparametric Instrumental Variables Estimation of a Quantile Regression Model," *Econometrica*, 75, 1191–1208.

Imbens, G., and W. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512.

Koenker, R., and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Krueger, A. B. (1999): "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics*, 114(2), 497–532.

Krueger, A. B., and D. M. Whitmore (2001): "The Effect of Attending a Small Class in the Early Grades on College–Test Taking and Middle School Test Results: Evidence from Project STAR," *The Economic Journal*, 111(468), 1–28.

Lehmann, E. (1974): *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.

Powell, D. (2014): "Estimation of Quantile Treatment Effects in the Presence of Covariates," *Unpublished manuscript*.

Powell, J. (1986): "Censored Regression Quantiles," *Journal of Econometrics*, 32, 143–155.

Rothe, C. (2010): "Identification of Unconditional Partial Effects in Nonseparable Models," *Economics Letters*, 109, 171–174.

Roy, A. D. (1951): "Some thoughts on the distribution of earnings," *Oxford Economic Papers (New Series)*, 3, 135–146.

Rubin, D. (1990): "Formal Modes of Statistical Inference for Causal Effects," *Journal of Statistical Planning and Inference*, 25, 279–292.

Van der Vaart, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press.

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1), 331–347.

Whitmore, D. (2005): "Resource and Peer Impacts on Girls' Academic Achievement: Evidence from a Randomized Experiment," *American Economic Review*, 95(2), 199–203.

Wüthrich, K. (2014): "A Comparison of two Quantile Models with Endogeneity," *Unpublished manuscript*.