Reinvestigating How Welfare Reform Influences Labor Supply: A Multiple Testing Approach[*]

Steven F. Lehrer[†]
Queen's University and NBER

Vincent Pohl[‡]
Queen's University

Kyungchul Song[§]
University of British Columbia

November 2014

Abstract

Economic theory suggests that individual responses to welfare reform depend on pre-treatment labor supply and other characteristics. Resulting changes in earnings may be positive or negative, leading to heterogenous treatment effects. In this paper, we extend the literature on treatment effect heterogeneity by introducing six nonparametric tests that are implemented using bootstrap testing of functional inequalities. The proposed tests view treatment effect heterogeneity as a multiple testing problem and make corrections for the family-wise error rate. To facilitate comparisons to the existing literature we re-examine the extent of heterogeneity in labor supply responses to the Jobs First welfare experiment across both quantiles of the earnings distribution and individual subgroups. Our results shed new light on who truly benefits from welfare reform and demonstrate the importance of correcting for multiple testing.

Keywords: multiple testing, bootstrap tests, quantile treatment effects, welfare reform, labor supply

JEL classification: C12, C21, I38, J22

[*] We wish to thank participants at the 2014 CLSRN annual conference for helpful comments and suggestions. Jacob Schwartz provided excellent research assistance. Lehrer and Song respectively wish to thank SSHRC for research support. The usual caveat applies.

[†] School of Policy Studies and Economics Department, email: lehrers@queensu.ca.

[‡] Industrial Relations Program and Economics Department, email: vincent.pohl@queensu.ca.

[§] Vancouver School of Economics, email: kysong@mail.ubc.ca.

1 Introduction

Individuals differ not only in their characteristics but also in how they respond to a particular treatment or intervention. Treatment effects may vary between subgroups defined by individual characteristics such as gender or race. In addition, individuals' response to a particular treatment may vary across quantiles of the unconditional outcome distribution. For example, welfare programs that provide work incentives may affect welfare recipients differently according to their demographic characteristics such as education or number and ages of children. In addition, welfare programs induce kinks in the recipients' budget constraint, so the treatment effect may also vary depending on their pre-treatment earnings.

This diverse and heterogenous behavior has not only changed how economists think about econometric models and policy evaluation but also has profound consequences for the scientific evaluation of public policy.[1] Although the importance of heterogeneous treatment effects is widely recognized in the causal inference literature, common practice remains to report an average causal effect parameter, even in cases where it is not possible to identify for which subset of individuals does this effect apply to.[2]

While an increasing number of studies account for possible treatment effect heterogeneity in evaluating social programs, most conduct statistical inference without allowing for dependence across subgroups. For example, Fink, McConnell, and Vollmer (2014) report that over 75 percent of studies that analyze data from field experiments published in 10 specific journals estimate separate average causal parameters for different subgroups. Fink, McConnell, and Vollmer (2014) argue that it is inappropriate in those studies to apply traditional standard errors and p-values when testing for heterogeneous treatment effects through interaction terms or subgroup analyses. After all, each interaction term represents a separate hypothesis beyond the original experimental design and results in a substantially increased type I error. Lee and Shaikh (2014) address this issue in their study of data from a randomized experiment by adopting a multiple testing procedure for subgroup treatment effects that controls the familywise error rate (FWER) in finite samples.

A similar observation can be made for distributional treatment effects. A growing number of studies examine if there are different treatment effects across quantiles of the outcome level, i.e. they estimate quantile treatment effects (QTEs) (e.g., Heckman, Smith, and Clements, 1997; Friedlander and Robins, 1997; Abadie, Angrist, and Imbens, 2002; Bitler, Gelbach, and Hoynes, 2006; Firpo, 2007). Individual test statistics at different quantiles involve their

[1] James Heckman stresses this point in his 2001 Nobel lecture, where he notes that conditional mean impacts including the average treatment effect may provide limited guidance for policy design and implementation (Heckman, 2001).

[2] In particular, a large academic debate (e.g., Deaton, 2009; Imbens, 2009; Heckman and Urzua, 2010) questions whether the local average treatment effect parameter obtained from an IV estimand has policy relevance.

sample counterparts across different quantiles, which are correlated. A naive approach of comparing individual test results to find percentile groups with positive treatment effects inevitably suffers from the issue of data mining due to the reuse of the same data as emphasized by White (2000).[3] Bitler, Gelbach, and Hoynes (2006), for instance, report point-wise confidence intervals for their estimated QTEs of the Jobs First welfare experiment.

In this paper, we develop a multiple testing procedure to analyze treatment effect heterogeneity across subgroups and outcome quantiles. Our flexible approach allows us to analyze treatment effect heterogeneity using various hypothesis testing procedures. First, investigating the existence of positive treatment effects for some subgroups or some outcome quantiles is formulated as a hypothesis testing problem. Second, the procedure enables us to identify the subgroups and outcome quantiles for which the treatment effect is estimated to be conspicuous beyond sampling variations. As the result is obtained through a formal multiple testing procedure, it properly takes into account the reuse of the same data for different demographic groups or quantile groups and controls the FWER so that it is unaffected by data mining.[4] Controlling the FWER in multiple comparisons across different quantiles is crucial for the validity of the inference procedure, as estimated treatment effects across different percentiles of the outcome distribution are highly unlikely to be independent.[5]

The multiple testing approach provides not only a basis for judging the empirical relevance of treatment effect heterogeneity. It also provides further information on the pattern of treatment effect heterogeneity across different population groups.[6] This information can offer important insights about how scarce social resources are to be distributed in an unequal society. Policymakers would have richer information to more effectively assign different treatments to individuals so as to balance competing objectives. For example, some welfare recipients may not change their labor supply when faced with work incentives because they

[3] In part as a response, statistical inference procedures developed in Abadie (2002), Rothe (2010) and Maier (2011), among others, focus on the whole distribution of potential outcomes to side-step multiple comparisons.

[4] More specifically, our procedure involves multiple inequalities of unconditional quantile functions, and draws on a bootstrap method for functional inequalities in a spirit similar to Lee, Song, and Whang (2014). To construct a multiple testing procedure that controls the FWER, we adapt the step-down method proposed by Romano and Wolf (2005a) to our context of testing multiple inequalities of conditional quantiles.

[5] Similarly as ours, Lee and Shaikh (2014) adopt a multiple testing procedure to identify subgroups of conspicuous treatment effects. However, there are several notable differences. First, they do not account for within-subgroup treatment effect heterogeneity in contrast to our approach. Second, Lee and Shaikh (2014) require the treatment to be randomly assigned unconditionally. In contrast, our approach is built on the assumption of selection on observables. Hence it accommodates non-experimental data whenever the assumption is deemed plausible.

[6] Our approach differs from Crump et al. (2008) in several aspects. First, Crump et al. (2008) focus on heterogeneity of the average treatment effect across subgroups, while our focus is on treatment effect heterogeneity across quantiles of the outcome distribution, motivated by the findings of Bitler, Gelbach, and Hoynes (2006). Second, Crump et al. (2008) use a joint test for treatment effect heterogeneity covering all the subgroups. In contrast, we use a multiple testing procedure to detect quantiles and/or subgroups for which there is a positive treatment effect. Finally, unlike Crump et al. (2008), we also investigate treatment effect heterogeneity across quantiles *within* each subgroup, so that the focus here is also on whether treatment effect heterogeneity across quantiles is mostly due to subgroup differences or not.

are constrained by other factors such as childcare needs. Moreover, policymakers can design welfare programs more effectively if they know which over range of the earnings distribution welfare recipients increase and reduce labor supply.

Our use of various formal testing procedures for treatment effect heterogeneity is not solely motivated by policy considerations but also economic theory. We demonstrate that a simple static model of labor supply predicts heterogenous responses to changes in the parameters of a welfare reform policy within and between subgroups along the intensive margin. To illustrate the tests we explore the extent of heterogeneity in labor supply responses in the Jobs First welfare experiment across quantiles of the earnings distribution. This paper builds on earlier research that examines the extent of heterogeneity in labor supply responses with this data including Bitler, Gelbach, and Hoynes (2006).[7] A follow-up paper by Bitler, Gelbach, and Hoynes (2014) presents evidence that treatment effect heterogeneity in terms of quantile treatment effects cannot be all ascribed to cross-subgroup variations in mean treatment effects with this data. In contrast to Bitler, Gelbach, and Hoynes (2014), we consider treatment effect heterogeneity both across subgroups and within subgroups by estimating QTEs for each subgroup. Importantly, we do not assume that treatment effects are constant within subgroups, but rather estimate subgroup specific QTEs. Therefore, our results shed additional light on the effects of welfare reform. Specifically, we identify both the subgroups and within subgroups the range of the earnings distribution, for which treatment effects are significantly positive.

In addition, we make an important methodological contribution to the literature that tests for treatment effect heterogeneity. While Bitler, Gelbach, and Hoynes (2014) allow for multiple tests across subgroups, we also adjust for dependencies between quantiles. Thereby, we provide a unified framework to test for treatment effect heterogeneity. Finally, we believe that these tests are important since recent work by Solon, Haider, and Wooldridge (2013) has shown that even when unconfoundedness holds (or with experimental data), researchers who estimate models that do not account for heterogeneous effects may provide inconsistent estimates of average effects.[8]

The rest of this paper is organized as follows. In Section 2, we motivate the tests that we develop by describing both the policy and data being investigated, and in Section 3 we

[7] In related work, Kline and Tartari (2013) demonstrate how economic theory imposes restrictions that can be used to develop bounds on the frequency of intensive and extensive margin responses to welfare reform. Our primary goal is not to develop tests to see if observed behavior is consistent with the quantitative predictions of a theory but rather whether qualitative differences in the pattern of QTEs between subgroups emerge.

[8] Under unconfoundedness, it is well known that matching and regression estimators may yield different estimates since they weight observations differently. Intuitively if there are heterogeneous treatment effects across groups in the sample, the OLS estimator gives a weighted average of these effects. The weights depend not only on the frequency of the subgroups, but also upon sample variances within the subgroup. This differs from the sample-weighted average which would be given by the average of each subgroup's partial effect weighted by its frequency in the sample.

present a simple labor supply model that predicts heterogenous treatment effects both within and across subgroups. We next describe the general testing procedure and how it improves the economic significance of studying heterogeneous treatment effects in Section 4. In Section 5, we present results from an empirical application of the methods to Jobs First data which yields two main findings. First, while there is clear evidence of treatment effect heterogeneity in the full sample, this is observed in most but not every subgroup. Second, we demonstrate the importance of making corrections for multiple testing since approximately a third of the QTEs become statistically insignificant when we account for potential dependencies. Taken together, our results shed new light on who truly benefits from welfare reform, further indicating how the composition of the labor force changes in response to public policy. The concluding Section 6 discusses the benefits and limitations of this approach.

2 Policy Background and Data

Following years of debate and after President Clinton vetoed two earlier welfare reform bills, the federal Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA) was passed in 1996.[9] PRWORA provided a major change in how federal cash assistance would be provided by requiring each state to replace their Aid to Families with Dependent Children (AFDC) program with a Temporary Assistance to Needy Families (TANF) program. In addition, PRWORA gave state governments more autonomy over welfare delivery. Several states including Connecticut elected to conduct randomized experiments to provide an evidence base for subsequent reforms.

Connecticut's Job First experiment was carried out by the Manpower Demonstration and Research Corporation and involved about 4,800 women residing in New Haven and Manchester in 1996 and 1997 that were either new applicants to welfare or had applied for a continued receipt of benefits. Participants were randomly assigned to either receive a new program called Jobs First that was the basis of the subsequent TANF program whose main features included 1) time limits and work requirements on welfare benefits and 2) a lower implicit marginal tax rate on earnings. Control units were assigned to the existing AFDC program.[10] Participants and their families were followed up until 2001 via surveys and administrative records from multiple sources including unemployment insurance earnings, food stamps, and AFDC/TANF benefits have been merged to the data.

[9] Haskins (2006) details the political battles underlying the passage of this act.

[10] Under Jobs First, all earnings up to the federal poverty level were disregarded while AFDC participants faced an implicit tax rate of 49 percent during the first three months of employment and 73 percent thereafter. Other differences include a $3,000 asset disregard and two years of transitional Medicaid for Jobs First and a $1,000 disregard and one year of transitional Medicaid for AFDC (see Bloom et al., 2002; Bitler, Gelbach, and Hoynes, 2006).

Table 1: Summary Statistics by Experimental Group

	Jobs First Mean (Std.dev.)	AFDC Mean (Std.dev.)	Difference *p*-value
Mother's age < 20	0.0889	0.0856	0.684
Mother's age 20 to 29	0.214	0.216	0.898
Mother's age ≥ 30	0.497	0.488	0.537
White	0.362	0.348	0.307
Black	0.368	0.371	0.836
Hispanic	0.207	0.216	0.423
Never married	0.654	0.661	0.624
Separated/divorced/living apart	0.332	0.327	0.715
No educational degree	0.350	0.334	0.242
High school degree/GED or more	0.650	0.666	0.242
Youngest child < 6	0.605	0.614	0.520
Youngest child ≥ 6	0.395	0.386	0.520
Number of children	1.649 (0.932)	1.591 (0.944)	0.037
Mean quarterly earnings pre-RA	682.7 (1304.1)	796.0 (1566.0)	0.006
Mean quarterly welfare benefits pre-RA	890.8 (806.0)	835.1 (784.8)	0.015
Mean quarterly foods stamp benefits pre-RA	352.1 (320.0)	339.4 (303.9)	0.156
Fraction of quarters employed pre-RA	0.327 (0.370)	0.357 (0.379)	0.006
Fraction of quarters welfare receipt pre-RA	0.573 (0.452)	0.544 (0.450)	0.026
Fraction of quarters food stamps receipt pre-RA	0.607 (0.438)	0.598 (0.433)	0.486
Observations	2396	2407	4803

Connecticut's Job First experiment is well-studied (Bitler, Gelbach, and Hoynes, 2006, 2014; Kline and Tartari, 2013). We use it to illustrate the methods proposed in the paper because it facilitates comparisons with the existing literature that used the same data extract. Summary statistics are reported in Table 1 where the second and third columns present characteristics of those women respectively assigned to the treated and control groups. On average, the single mothers in this sample have lower educational attainment and are much more likely to be part of a minority than the general population. About 60 percent of the sample have a child under the age of six, indicating that there may additional constraints in their labor supply decisions. The women in this sample earn less than $800 per quarter before random assignment and therefore rely heavily on welfare and food stamps. The last column in Table 1 contains p-values for the test that individuals' characteristics assigned to the Jobs First and AFDC groups do not differ in observed characteristics. For most characteristics and as shown in both Bloom et al. (2002) and Bitler, Gelbach, and Hoynes (2006), we cannot reject the null hypothesis of no difference. There are small but statistically significant differences in a few variables, and two differences are particularly surprising given the random assignment protocol. We observe that women assigned to the control group (AFDC) have significantly higher earnings and hence receive significantly lower welfare benefits before random assignment. To control for the full set of differences and ensure covariate balance we make adjustments via propensity score weighting in our analyses.

3 Economic Model Predicting Heterogeneous Treatment Effects

A simple static labor supply model motivates our investigation of treatment effect heterogeneity.[11] Individuals maximize their utility over consumption (C) and earnings (E) subject to a budget constraint:

$$\max_{C,E} U = U(C, E; X_1) \tag{1}$$

$$\text{s.t. } C = E + W(E; X_2, Z^t) \tag{2}$$

where X_1 denotes characteristics including subgroups that may affect individual preferences and $W(\cdot)$ denotes the welfare benefit function, which depends on the level of earnings, household characteristics X_2, and policy parameters Z^t with $t = \{AFDC, JF\}$. The vector Z^t includes the base grant amount, earnings disregards, and time limits, so it traces out the budget constraint faced by a welfare participant in the AFDC or Jobs First group. Following

[11] Static models are commonly used in the literature on single mothers' labor supply (Keane, 2011, p. 1070). Our discussion follows earlier work on static labor supply models including Kline and Tartari (2013). We extend this literature by considering differences across subgroups.

Saez (2010) we assume that the marginal utility of consumption is positive and marginal utility of earnings is negative.

We use the panels of Figure 1 to demonstrate how economic theory predicts treatment effect heterogeneity for all subgroups. This heterogeneity arises because there is a differential labor supply response on both the intensive and extensive margins between the AFDC and the Jobs First program due to different budget constraints and earnings distributions in the two experimental groups.[12] The solid line in the top panel of Figure 1 illustrates the budget constraint faced by Jobs First participants containing the points A_0, F_1, E and G_0. A_0 denotes the base grant amount and and the segment A_0F_1 represents the implicit tax rate of zero for participants with earnings below the federal poverty line (FPL). The dashed line represents the budget constraint faced under AFDC and contains points A_0, C and G_0. The segment A_0C represents the earnings disregard under AFDC.[13] The middle panel of Figure 1 presents hypothetical cumulative distribution functions of earnings for those in AFDC (dashed) and Jobs First (solid) groups that are the result of different welfare program parameters. QTEs are presented in the bottom panel and equal simply the horizontal distance at each quantile between the two earnings distributions in the middle panel. To provide intuition for the shape of the QTEs presented we will consider the thought experiment of moving individuals from AFDC to Job First.

First, consider an individual located at point A_0 under AFDC. Supposing she is assigned to receive Jobs First, she can now either remain at point A_0 or can move along the budget constraint to point A_1. The observed choice depends on her preferences over consumption and earnings. In particular, women with steeper indifference curves at point A_0 are less likely to change their decisions between AFDC and Jobs First.

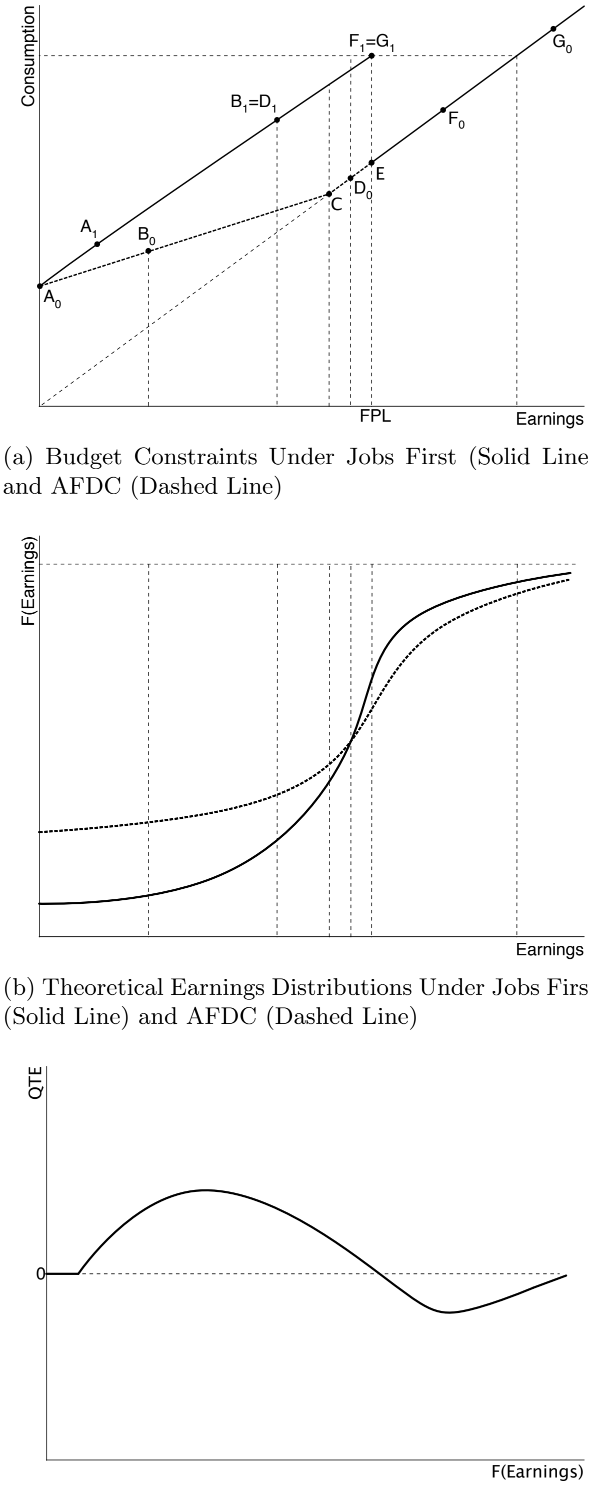
We next consider a woman on AFDC at point B_0 who works while receiving welfare. Transiting to Jobs First lowers her implicit tax rate from either 49 or 73 percent to zero boosting her wage.[14] If the substitution effect exceeds the income effect, the labor supply response moves her to point B_1 and leads to a rightward shift in the earnings distribution. Hence, for workers whose earnings lie in the segment between A_0 and C, theory predicts positive QTEs.

The QTEs in the bottom panel shift from positive to negative around the point D_0 which corresponds to women ineligible for welfare under AFDC but eligible under Jobs First. Theory predicts moving to Jobs First would lead to a reduction in labor supply to point D_1,

[12] We abstract from welfare stigma and the hassle associated with not working while on welfare modeled by Kline and Tartari (2013). We are interested in the distribution of earnings and how it varies by subgroup, but not in the welfare participation decision or decomposing labor supply responses into the extensive and intensive margin here.

[13] The earnings disregard was 51 percent for the first four months and 27 percent thereafter under AFDC.

[14] Note that we will use changes in labor supply and earnings interchangeably here because the wage is assumed to be constant.



(a) Budget Constraints Under Jobs First (Solid Line) and AFDC (Dashed Line)

(b) Theoretical Earnings Distributions Under Jobs First (Solid Line) and AFDC (Dashed Line)

(c) Theoretical Quantile Treatment Effects

Figure 1: Theoretical Predictions of Labor Supply

if we make the standard assumption that leisure is a normal good.[15,16]

Similarly, negative QTEs arise for women located at point F_0 who face perverse incentives when transiting to Jobs First. These women would neither qualify for AFDC nor Job First at point F_0, however, the generous earnings disregard under Jobs First would incentivize women to reduce their labor supply to qualify for the new benefits leading to a movement to point F_1 which is characterized by higher consumption and lower earnings.

Last, women with even higher earnings under AFDC at point G_0 face a tougher choice when moving to Jobs First since the point $G_1 = F_1$ does not strictly dominate point G_0. For women whose marginal disutility from earnings outweighs the marginal utility from consumption, labor supply will fall from G_0 to G_1, whereas women with different preferences may choose to at point G_0 and not change their labor supply. Thus, we predict the QTEs to be negative around G_0 but at higher quantiles, the QTEs may become zero.

Taken together, theory predicts that the earnings QTEs of Job First will start at zero and be positive for a range of quantiles before becoming negative and eventually reaching zero again as illustrated in the bottom panel of Figure 1.

The above discussion concerned the general shape of treatment effect heterogeneity but did not consider subgroups. Subgroup membership denoted by X_1 and X_2 in equations (1) and (2) is hypothesized to affect preferences and directly influence the budget constraint. Since the parameters in this optimization problem vary, the QTEs could be shifted to the left or right, be compressed or stretched, or otherwise be transformed without losing their overall shape depicted in the bottom panel of Figure 1. To illustrate, consider subgroups defined by maternal education. We ignore the potential effect of education on preferences, but assume that women with more education receive higher wage offers. Therefore, we would expect a larger fraction of women with higher educational attainment to be located around the points F_0 and G_0 and correspondingly less mass around A_0, B_0, and D_0 compared to women with less education.[17] Thus, we would expect an overall shift of the QTEs to the left with fewer quantiles at the bottom of the distribution where the QTEs equal zero for higher educated women.

A similar shift would be anticipated for subgroups defined by earnings and welfare history, where we would also expect qualitative differences in the shape of the QTEs.[18] Mechanically, recent welfare recipients have little if any positive earnings in the period before the experiment

[15] To avoid clutter, we set $D_1 = B_1$ without loss of generality.

[16] Intuitively, by moving to Jobs First at point D_0 (at which welfare is not available under AFDC), women now gain the base grant and face an implicit tax rate of 0 percent. Thus, changing from AFDC to Jobs First at point D_0 only results in an income effect.

[17] The average level of education is much lower in our sample of welfare recipients than in the general population. Therefore, we split the sample into high and low education subgroups by whether individuals have a high school degree or not.

[18] These variables reflect welfare recipients' work experience and thereby affect their current earnings through the wage (Mincer, 1974).

so there will be little mass around points D_0, F_0, and G_0 relative to A_0 and B_0. Thus, we would expect more positive QTEs (i.e. moves from B_0 to B_1) for these individuals and more negative QTEs (i.e. switches from D_0 to D_1 or from F_0 to F_1) for individuals with less recent welfare participation and higher previous earnings. In addition, we would expect that women who have relied on welfare or have low earnings to respond more strongly to Jobs First in part due to the time limits imposed.

Finally, consider subgroups defined by either the age or number of children. Additional children will mechanically influence the size of benefits. Yet, under Jobs First the potential loss of welfare benefits when time limits are imposed might be higher for women with additional children. In addition, these women may need to earn more in the labor market if they leave welfare. While it is not possible to predict differences in the range of positive and negative QTEs by number of children, it is reasonable to expect larger QTEs among women with more children. Similarly, women with older children may exhibit a similar pattern of having larger QTEs. This arises since young children impose a higher opportunity cost of work for mothers relative to older children and this cost is fixed independent of receiving AFDC or Jobs First.

In summary, economic theory predicts that there will be treatment effect heterogeneity both within and between subgroups motivating the development of tools to assess its extent in general as well as in the specific context of the Jobs First study.

4 Methodology

In this section, we begin by considering general strategies to test for treatment effect heterogeneity and then introduce tests that are motivated by the discussion in the previous section. The main focus is on performing valid inference when faced with multiple hypotheses to test at the same time. Here, each individual hypothesis corresponds to the hypothesis that there exists a positive QTE for a particular quantile and subgroup. As is well known in the econometrics literature (e.g., White, 2000), we need to properly account for the fact that we test multiple hypotheses using the same data.[19] Several adjustments have been proposed in the literature, starting with the very conservative Bonferroni method (see Romano, Shaikh, and Wolf (2010) for a recent overview). A common more recent procedure is to use stepwise methods to iteratively eliminate null hypotheses that can be rejected until a set of hypotheses remains, for which the null cannot be rejected (Romano and Wolf, 2005a,b). To do so, it is necessary to update the critical value at each step, for example by using a bootstrap

[19] The problem when testing multiple hypotheses jointly is the potential over-rejection of the null hypothesis. Intuitively, if the null hypothesis of no treatment effect is true, testing it across 100 subsamples, we expect about five rejections at the 95 percent level. However, if these subsamples depend on each other, more than five rejections may occur. Hence, the type I error would be too large. The issue arises when testing a hypothesis across the percentiles of an outcome variable.

method.[20] We follow this idea and use a bootstrap based step-down method to identify the subgroup-quantile cells for which positive treatment effects are present. By combining bootstrap tests of functional inequalities with multiple testing procedures, we produce various testing procedures suitable for analyzing treatment heterogeneity.

4.1 Treatment Effect Heterogeneity Without Subgroups

Each of the tests will focus on QTEs that are calculated by subtracting control group (AFDC) quarterly earnings at quantile τ from quarterly earnings at quantile τ in the treatment group (Jobs First).[21] To control for selection on observables into treatment and control groups,[22] we follow (Bitler, Gelbach, and Hoynes, 2006) and estimate the propensity score $\hat{p}(x)$ using a series logit specification.[23] Second, we define inverse propensity score weights as

$$\hat{\omega}_{1i} = \frac{D_i}{\hat{p}(X_i)} \quad \text{and} \quad \hat{\omega}_{0i} = \frac{1-D_i}{1-\hat{p}(X_i)}$$

for treated and control individuals, respectively with D_i being the treatment indicator that equals one in the Jobs First group and zero in the AFDC group. We then obtain quantiles of the weighted quarterly earnings as follows:

$$\hat{q}_{1,\tau} = \arg\min_q \sum_{i=1}^{n} \hat{\omega}_{1i}\rho_\tau\left(Y_i - q\right) \quad \text{and}$$
$$\hat{q}_{0,\tau} = \text{argmin}_q \sum_{i=1}^{n} \hat{\omega}_{0i}\rho_\tau\left(Y_i - q\right),$$

where $\rho_\tau(x) = x \cdot (\tau - \mathbf{1}\{x \leq 0\})$ is a control function (Firpo, 2007). That is, $\hat{q}_{\tau,1}$ and $\hat{q}_{\tau,0}$ are the τ-th empirical quantiles of propensity score weighted quarterly earnings

$$\left\{\frac{Y_i D_i}{\hat{p}(X_i)}\right\}_{i=1}^{n} \quad \text{and} \quad \left\{\frac{Y_i(1-D_i)}{1-\hat{p}(X_i)}\right\}_{i=1}^{n}$$

(or $\{\hat{Y}_{1i}\}_{i=1}^{n}$ and $\{\hat{Y}_{0i}\}_{i=1}^{n}$), respectively and $Y_{1i} = Y_i D_i$ and $Y_{0i} = Y_i(1-D_i)$ are treatment and control group quarterly earnings, respectively.

[20] This procedure controls the FWER, i.e. the probability of at least one type I error, at the desired level.

[21] We follow Bitler, Gelbach, and Hoynes (2006) and pool all the data from the first seven quarters after random assignment. To infer treatment effects for specific individuals from QTEs we have to assume that there are no rank reversals in the earnings distribution between the Jobs First and AFDC groups. This assumption is likely violated and even predicted not to hold by labor supply theory (see Section 3 above). However, positive QTEs imply that the treatment has a positive effect for some interval of the earnings distribution (Bitler, Gelbach, and Hoynes, 2006).

[22] Our tests allow for selection on observables and do not assume that treatment is randomly assigned as in Lee and Shaikh (2014).

[23] We use a nonparametric approach since the tests are also nonparametric. That said, the vast majority of our results are robust to using a parametric logit estimator.

Formally, the QTE at τ is then defined as

$$q_\tau^\Delta = \hat{q}_{1,\tau} - \hat{q}_{0,\tau}.$$

Intuitively and as shown in panels (b) and (c) of Figure 1, the QTE is equal to the horizontal difference between the graphs of the earnings distributions of treatment and control group at quantile τ.

4.1.1 Testing for the Presence of Positive Quantile Treatment Effects

The first test is designed to determine whether the welfare program had any positive effect on earnings in some range of the earnings distribution at all.[24] We test the following null and alternative hypotheses:

$$\begin{aligned} H_0 &: q_\tau^\Delta \leq 0 \text{ for all } \tau \in \mathcal{T} \\ H_1 &: q_\tau^\Delta > 0 \text{ for some } \tau \in \mathcal{T}, \end{aligned} \tag{H.1}$$

where $\mathcal{T}$ is the set of quantiles (in this case percentiles).[25] The alternative hypothesis states that there exists a positive treatment effect for at least one quantile. Therefore, we expect that the null hypothesis will be rejected if Jobs First had any positive effect on labor supply in same range of the earnings distribution.

The hypothesis (H.1) illustrates why we need to account for dependencies between quantiles and therefore develop a multiple testing procedure. We do not test if each QTE is larger than zero individually by calculating separate point-wise confidence intervals for each quantile. That approach would not take into account that the QTEs at any two or more quantiles are derived from the same underlying outcome variable and are therefore dependent.

We consider a test statistic of the following form:

$$T_n = \max_{\tau \in \mathcal{T}} \hat{q}_\tau^\Delta. \tag{3}$$

Intuitively, since the null hypothesis states that all QTEs are weakly negative, the largest (positive) observed QTE provides the clearest evidence against the null hypothesis (White, 2000). To test the null hypothesis (H.1), we calculate a critical value using a bootstrap method. Specifically, we first resample with replacement from the original sample B times and construct $\hat{Y}_{1i}^* = Y_i^* D_i^* / \hat{p}^*(X_i^*)$ and $\hat{Y}_{0i}^* = Y_i^*(1-D_i^*)/(1-\hat{p}^*(X_i^*))$, where $\{Y_i^*, D_i^*, X_i^*\}_{i=1}^n$

[24] The idea for this test has policy appeal since, given limited resources, policymakers first need to know if individuals react to a specific policy intervention at all. In contrast, the average treatment effect may conceal positive QTEs if they are entirely offset by negative QTEs in a different range of the outcome distribution.

[25] In contrast to Crump et al. (2008) who use two-sided tests, our approach focuses on one-sided tests concentrating our power on the alternative hypotheses of the positive treatment effects. So our tests have a different power behavior from those of Crump et al. (2008).

denotes each bootstrap sample and $\hat{p}^*(X_i^*)$ the logit estimator of the propensity score using the bootstrap sample. Then the bootstrap test statistic for bootstrap draw $b = \{1, \ldots, B\}$ is given by

$$T^*_{n,b} = \max_{\tau \in \mathcal{T}} \left\{ \hat{q}^{\Delta *}_{\tau} - \hat{q}^{\Delta}_{\tau} \right\}, \tag{4}$$

where $\hat{q}^{\Delta *}_{\tau} = \hat{q}^*_{\tau,1} - \hat{q}^*_{\tau,0}$ and $\hat{q}^*_{\tau,1}$ and $\hat{q}^*_{\tau,0}$ are the τ-the empirical quantiles of $\{\hat{Y}^*_{1i}\}^n_{i=1}$ and $\{\hat{Y}^*_{0i}\}^n_{i=1}$, respectively. By subtracting $\hat{q}^{\Delta}_{\tau}$ we re-center the bootstrap test statistic in order to impose the null restriction. We then compare the test statistic (3) to the bootstrap critical value, which is equal to the $(1-\alpha)$th quantile of the B bootstrap test statistics (4), where α is the level of the test.[26] We reject null hypothesis (H.1) if the test statistic exceeds the critical value. Rejection of the null hypothesis indicates evidence for positive treatment effects for some range of the earnings distribution.

4.1.2 Testing For Which Quantiles the Treatment Effect Is Positive

While rejecting the null hypothesis (H.1) informs us that there are significant positive QTEs for some quantiles, we may also want to identify the range of quantiles over which there is evidence for positive QTEs. Such a range can be of considerable interest for policymakers when they wish to define a target group for their policies in a way that carries empirical support. Given limited resources the target group should consist of those people for which the treatment program has turned out most successful.

Thus we first define individual hypothesis testing problem as follows: for each τ in a range $\mathcal{T} \subset [0, 1]$,

$$\begin{aligned} H_{0,\tau} &: q^{\Delta}_{\tau} \leq 0 \\ H_{1,\tau} &: q^{\Delta}_{\tau} > 0. \end{aligned} \tag{H.2}$$

Then the goal is to find a set of individual hypotheses, for which the null is false, in a way that controls the FWER.[27]

To implement this approach, we follow Romano and Wolf (2005a) and Romano and Shaikh (2010) by conducting stepwise elimination of quantiles using the bootstrap. More specifically, setting $\mathcal{T}_1 = \mathcal{T}$, we find $\hat{c}_1$ such that

$$\frac{1}{B} \sum_{b=1}^{B} \mathbf{1} \left\{ T^*_{n,b}(\mathcal{T}_1) > \hat{c}_1 \right\} \leq \alpha,$$

where $T^*_{n,b}(\mathcal{T}_1) = \sup_{\tau \in \mathcal{T}_1} \left| \hat{q}^{\Delta *}_{\tau} - \hat{q}^{\Delta}_{\tau} \right|$ denotes the bootstrap one-sided test statistic using the

[26] In our empirical application we set $\alpha = 0.05$ and $B = 999$.

[27] The FWER here is the probability that we mistakenly declare a positive QTE for some $\tau \in \mathcal{T}$.

b-th bootstrap sample and α is the level of the test. That is, at $\hat{c}_1$, the fraction of test statistics across the B bootstrap samples that exceed that critical value is at most α. Then, we define

$$\mathcal{T}_2 = \left\{\tau \in \mathcal{T}_1 : \hat{q}_\tau^\Delta \leq \hat{c}_1\right\}.$$

Note that $\mathcal{T}_2$ is a subset of $\mathcal{T}_1$ because it contains only the quantiles for which the treatment effect is less than the critical value $\hat{c}_1$. Now, we construct $T^*_{n,b}(\mathcal{T}_2) = \sup_{\tau \in \mathcal{T}_2} \left|\hat{q}_\tau^{\Delta *} - \hat{q}_\tau^\Delta\right|$, find $\hat{c}_2$ such that

$$\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\left\{T^*_{n,b}(\mathcal{T}_2) > \hat{c}_2\right\} \leq \alpha,$$

and define

$$\mathcal{T}_3 = \left\{\tau \in \mathcal{T}_2 : \hat{q}_\tau^\Delta \leq \hat{c}_2\right\}.$$

This procedure is repeated until at step k, we obtain $\mathcal{T}_k = \left\{\tau \in \mathcal{T}_{k-1} : \hat{q}_\tau^\Delta \leq \ \hat{c}_k\right\}$ such that no further element of $\mathcal{T}_k$ is eliminated. Then the resulting set $\mathcal{T}_k$ is the subset of $\mathcal{T}$ such that there is no empirical support for positive treatment effect at quantiles $\tau \in \mathcal{T}_k$. From the result of Romano and Shaikh (2012), it is not hard to show that this multiple testing procedure controls the FWER at α.

4.1.3 Testing For General Treatment Effect Heterogeneity

Suppose that we are interested in checking whether the data support heterogeneity of QTEs across quantiles. While one may obtain information from a visual inspection of QTEs across quantiles, a formal test is necessary to properly account for sampling variation. For this, we consider the hypotheses of the following form:

$$\begin{aligned} &H_0 : q_\tau^\Delta = c \text{ for all } \tau \in \mathcal{T} \text{ and for some } c \in \mathbb{R} \\ &H_1 : q_\tau^\Delta \neq c \text{ for some } \tau \in \mathcal{T} \text{ and for all } c \in \mathbb{R}. \end{aligned} \qquad \text{(H.3)}$$

The alternative hypothesis indicates heterogeneity of QTE across quantiles. When the null hypothesis is rejected, it suggests evidence for differential reactions by individuals to the welfare program depending on where in the earnings distribution they are located.[28,29]

[28] This test is important because it answers the policy-relevant question: do individuals differ in their response to a particular policy, in our case to welfare reform (Heckman, Smith, and Clements, 1997). In the present context, the results can guide policymakers in adjusting welfare rules, for example by introducing more (or different) conditions for welfare receipt. While we address treatment effect heterogeneity in the presence of subgroups below, this test is nevertheless important because it gives us a first and simple answer to the question if treatment effects vary at all.

[29] Crump et al. (2008) consider differences in average treatment effects across subpopulations defined by observable characteristics. In contrast to their approach involving moment inequalities, we test functional inequalities defined by quantiles of the outcome distribution.

To test (H.3) we construct the following test statistic:

$$T_n = \sup_{\tau \in \mathcal{T}} \left| \hat{q}_\tau^\Delta - \bar{q}^\Delta \right|, \tag{5}$$

where

$$\bar{q}^\Delta = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \hat{q}_{\tau,1} - \hat{q}_{\tau,0}.$$

That is, we set the constant c in (H.3) equal to the mean QTE, $\bar{q}^\Delta$, and subtract it from the estimated quantile treatment effects, so that the test statistic will be small if the QTE are very similar.[30] The sup appears in equation (5) to detect the existence of quantiles at which the deviation of the QTE from its mean occurs.

We then follow the same bootstrap approach as in Section 4.1.1 above and calculate the following bootstrap test statistic:

$$T_{n,b}^* = \sup_{\tau \in \mathcal{T}} \left| \hat{q}_\tau^{\Delta *} - \hat{q}_\tau^\Delta - \left(\bar{q}^{\Delta *} - \bar{q}^\Delta \right) \right|. \tag{6}$$

The bootstrap test also incorporates re-centering in order to impose the null restriction. To test null hypothesis (H.3), we compare the test statistic (5) to the critical value obtained from the bootstrap test statistics (6), which is equal to the $(1-\alpha)$th quantile of the bootstrap statistics $T_{n,b}^*, b = 1, \ldots, B$. In the results section below we report the test statistic and critical value along with the p-value.

4.2 Incorporating Subgroups

The preceding three tests did not consider treatment effect heterogeneity across different subgroups. Tests involving subgroups can be useful when policymakers want to identify demographic subgroups that achieve differential gains, or when they are interested in the extent of heterogeneity within subgroups. For example, given limited resources, policymakers may be reluctant to extend programs to groups where a significant fraction does not receive gains. Finally, and consistent with the arguments in Lee and Shaikh (2014) and Fink, McConnell, and Vollmer (2014), it is important to develop tools for statistical inference in this setting that account for dependence both within and across subgroups.

Following Bitler, Gelbach, and Hoynes (2014), we consider subgroups defined by proxies for standard demographics, wage opportunities, fixed costs of work, preferences for income versus leisure, and employment and welfare histories. We assume that the subgroup vector Z_i is a subvector of X_i, so we write $X_i = (X_{1i}, Z_i)$, where X_{1i} indicates the vector that is not included in Z_i. As before $q_{\tau,1}(z)$ and $q_{\tau,0}(z)$ are the quantiles of the earnings distribution

[30] Since the null hypothesis involves an equality, we take the absolute value of the difference between QTE and mean QTE.

of treatment and control group, respectively, but now defined separately by subgroup z. Formally, define $q_{\tau,1}(z)$ and $q_{\tau,0}(z)$ to be the solution for the equations

$$\begin{aligned} &P\left\{Y_{1i} \le q_{\tau,1}(z) | Z_i = z\right\} \quad \text{and} \\ &P\left\{Y_{0i} \le q_{\tau,0}(z) | Z_i = z\right\}, \end{aligned}$$

where Z_i is the subgroup vector taking values from a finite set $\mathcal{Z} = \times_{j=1}^{J} \mathcal{Z}_j$, where $\mathcal{Z}_j$ is the set of values from the j-th category (e.g., education or welfare history). Hence $q_{\tau,1}(z)$ and $q_{\tau,0}(z)$ are the quantiles of quarterly earnings in the treatment and control groups conditional on subgroup z. Then the subgroup QTE is defined by

$$q_{\tau}^{\Delta}(z) = q_{\tau,1}(z) - q_{\tau,0}(z).$$

To account for covariates in the analyses, we continue to use inverse propensity score weighting with the weights given by

$$\hat{\omega}_{1i}(z) = \frac{D_i}{\hat{p}(X_{1i}, z)} \quad \text{and} \quad \hat{\omega}_{0i}(z) = \frac{1 - D_i}{1 - \hat{p}(X_{1i}, z)},$$

where $\hat{p}(X_{1i}, z)$ denotes the propensity score $\hat{p}(X_i)$ except that Z_i is replaced by z.[31] We define the empirical quantiles of the quarterly earnings for subgroup z in the treatment $\hat{q}_{1,\tau}(z)$ and control group $\hat{q}_{0,\tau}(z)$ as

$$\begin{aligned} \hat{q}_{1,\tau}(z) &= \arg\min_{q} \frac{1}{\sum_{i=1}^{n} \mathbf{1}\left\{Z_i = z\right\}} \sum_{i=1}^{n} \hat{\omega}_{1i}(z) \rho_{\tau}\left(Y_i - q\right) \mathbf{1}\left\{Z_i = z\right\} \quad \text{and} \\ \hat{q}_{0,\tau}(z) &= \arg\min_{q} \frac{1}{\sum_{i=1}^{n} \mathbf{1}\left\{Z_i = z\right\}} \sum_{i=1}^{n} \hat{\omega}_{0i}(z) \rho_{\tau}\left(Y_i - q\right) \mathbf{1}\left\{Z_i = z\right\}, \end{aligned}$$

and for the next set of tests the earnings quantiles are calculated separately for each subgroup.

4.2.1 Testing For Which Quantiles and Subgroups the Treatment Effect Is Positive

The first test considers the test of hypothesis (H.2) with subgroups. That is, we identify the quantile-subgroup cells that have significantly positive treatment effects.[32] We consider the

[31] The propensity score $\hat{p}(x)$ continues to be estimated using a series logit specification using the whole sample (Smith and Todd, 2005).

[32] As discussed in Section 3, labor supply theory predicts that individuals with different observable characteristics may react differently to the same welfare rules. In particular, characteristics such age and number of children or maternal education may determine for which range of the earnings distribution we observe an increase or decrease in labor supply.

following individual hypotheses: for each $\tau \in \mathcal{T}$ and $z \in \mathcal{Z}$,

$$\begin{aligned} &H_{0,\tau,z} : q_\tau^\Delta(z) \leq 0 \\ &H_{1,\tau,z} : q_\tau^\Delta(z) > 0. \end{aligned} \tag{H.4}$$

The test is constructed as follows. First, by resampling with replacement from the original sample, we construct $\hat{Y}_{1i}^* = Y_i^* D_i^* / \hat{p}^*(X_i^*)$ and $\hat{Y}_{0i}^* = Y_i^*(1 - D_i^*)/(1 - \hat{p}^*(X_i^*))$. Then we take our bootstrap one-sided test statistic to be

$$T_{n,b}^*(\mathcal{W}) = \sup_{(\tau,z)\in\mathcal{W}} \left\{ \hat{q}_\tau^{\Delta*}(z) - \hat{q}_\tau^\Delta(z) \right\}, \tag{7}$$

where $\hat{q}_\tau^{\Delta*}(z) = \hat{q}_{\tau,1}^*(z) - \hat{q}_{\tau,0}^*(z)$, $\hat{q}_{\tau,1}^*(z)$ and $\hat{q}_{\tau,0}^*(z)$ are the empirical quantiles of $\{\hat{Y}_{1i}^*\}_{i=1}^n$ and $\{\hat{Y}_{0i}^*\}_{i=1}^n$, respectively, at quantile τ within the samples with $Z_i^* = z$, and $\mathcal{W} = \mathcal{T} \times \mathcal{Z}$ is the set of subgroup-quantile cells. To perform multiple testing, we proceed by eliminating subgroup-quantile cells stepwise. At each step, we retain those (τ, z) cells for which no evidence for positive treatment effect can be found.

Specifically, we take $\mathcal{W}_1 = \mathcal{T} \times \mathcal{Z}_j$, and find minimum $\hat{c}_1$ such that

$$\frac{1}{B} \sum_{b=1}^{B} \left\{ T_{n,b}^*(\mathcal{W}_1) > \hat{c}_1 \right\} \leq \alpha,$$

where $T_{n,b}^*(\mathcal{W}_1)$ is defined in equation (7) and α is the desired FWER. We define

$$\mathcal{W}_2 = \left\{ (\tau, z) \in \mathcal{W}_1 : \hat{q}_\tau^\Delta(z) \leq \hat{c}_1 \right\}$$

and construct $T_{n,b}^*(\mathcal{W}_2) = \sup_{(\tau,z)\in\mathcal{W}_2} \left\{ \hat{q}_\tau^{\Delta*}(z) - \hat{q}_\tau^\Delta(z) \right\}$ to find minimum $\hat{c}_2$ such that

$$\frac{1}{B} \sum_{b=1}^{B} \left\{ T_{n,b}^*(\mathcal{W}_2) > \hat{c}_2 \right\} \leq \alpha.$$

We then define

$$\mathcal{W}_3 = \left\{ (\tau, z) \in \mathcal{W}_2 : \hat{q}_\tau^\Delta(z) \leq \hat{c}_2 \right\}.$$

The process is repeated until we obtain $\mathcal{W}_k = \left\{ (\tau, z) \in \mathcal{W}_k : \hat{q}_\tau^\Delta(z) \leq \hat{c}_{k-1} \right\}$ for some k such that no further element of $\mathcal{W}_k$ is eliminated. Then the resulting set $\mathcal{W}_k$ is the subset of $\mathcal{W}$ such that there is no empirical support that the treatment effect at quantile-subgroup pair $(\tau, z) \in \mathcal{W}_k$ is positive. This procedure will yield all the combinations of subgroups and quantiles where positive treatment effects are present; they are given by quantile-subgroup pairs $(\tau, z) \in \mathcal{W} \backslash \mathcal{W}_k$.

4.2.2 Testing Treatment Effect Heterogeneity Across Quantiles and Between Subgroups

Here we focus on the question of whether differences across subgroups can explain the observed heterogeneity of QTEs in the full sample. More specifically, we search for evidence that all subgroups exihibit heterogeneity of treatment effects across different quantiles $\tau \in (0,1)$:

$$\begin{aligned} &H_0 : q_\tau^\Delta(z) = c_z \text{ for all } \tau \in \mathcal{T}, \text{ for some } c_z \in \mathbb{R}, \text{ and some } z \in \mathcal{Z} \\ &H_1 : q_\tau^\Delta(z) \neq c_z \text{ for some } \tau \in \mathcal{T}, \text{ for all } c_z \in \mathbb{R}, \text{ and all } z \in \mathcal{Z}. \end{aligned} \tag{H.5}$$

The null hypothesis states that the heterogeneity in treatment effects disappears when we condition on z for some $z \in \mathcal{Z}$. In other words, it posits that the QTEs are constant across quantiles within each subgroup. However, there can be treatment effect heterogeneity across subgroups under the null. For example, under the null hypothesis, individuals with a high school degree may have a different mean treatment effect than high school drop-outs, but within each education category, the treatment effects are constant across the earnings distribution. The alternative hypothesis indicates heterogeneity of QTEs across different τs, even after controlling for z.[33] That is, individuals with the same level of education have different treatment effects according to their location on the earnings distribution. Note that we only reject the null hypothesis if our test detects treatment effect heterogeneity for all subgroups. Last, it is important to note that Bitler, Gelbach, and Hoynes (2014) ask exactly this question. In contrast to their paper, we 1) do not only account for multiple testing across subgroups but also across quantiles, and 2) do not constrain the treatment effect to be constant within subgroups.

The following test statistic is used to test the hypothesis (H.5):

$$T_n = \min_{z \in \mathcal{Z}} \sup_{\tau \in \mathcal{T}} \left| \hat{q}_\tau^\Delta(z) - \bar{q}^\Delta(z) \right|, \tag{8}$$

where $\bar{q}^\Delta(z) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} (\hat{q}_{\tau,1}(z) - \hat{q}_{\tau,0}(z))$ is the mean of the QTEs for each subgroup. As with test statistic (5), we impose the null hypothesis by subtracting $\bar{q}^\Delta(z)$. For each subgroup, the highest deviation of the QTEs from their mean provides the clearest evidence against the null hypothesis. Then to obtain a test statistic that covers all subgroups $z \in \mathcal{Z}$, we take the minimum value over each subgroup's test statistic. Intuitively, we search for evidence that all subgroups exhibit treatment effect heterogeneity, so we restrict our attention to the subgroups that have the least amount of heterogeneity.

To construct a bootstrap critical value, we consider the following bootstrap test statistic

[33] A Kolmogov-Smirnov test would reject the equality of the distributions in this case. The advantage of our test is that when rejecting the null hypothesis, additional insights on how the null hypothesis is violated are provided (Romano and Wolf, 2005a).

that is an analogue of (6) with subgroups:

$$T^*_{n,b} = \min_{z \in \mathcal{Z}} \sup_{\tau \in \mathcal{T}} \left| \hat{q}^{\Delta *}_{\tau}(z) - \hat{q}^{\Delta}_{\tau}(z) - \left(\bar{q}^{\Delta *}(z) - \bar{q}^{\Delta}(z) \right) \right| . \tag{9}$$

To test hypothesis (H.5) we compare the test statistic (8) to the bootstrap critical value, which equals the $(1-\alpha)$th quantile of the bootstrap test statistics (9) for $b = 1, \dots, B$, as described above. In the results section below, we report the test statistic and critical value for each set of subgroups (education, number of children, etc.) along with exact p-values. Hence, our test of hypothesis (H.5) provides a simple yet flexible way to check for treatment effect heterogeneity across subgroups while allowing for distributional treatment effects within subgroups.

4.2.3 Multiple Testing Approach to Treatment Effect Heterogeneity Across Quantiles

We next test for treatment effect heterogeneity within subgroups, but separately for each subgroup. For each $z \in \mathcal{Z}$, we test

$$\begin{aligned} &H_{0,z} : q^{\Delta}_{\tau,1}(z) = c_z \text{ for all } \tau \in \mathcal{T} \text{ for some constant } c_z \in \mathbb{R} \\ &H_{1,z} : q^{\Delta}_{\tau,1}(z) \neq c_z \text{ for some } \tau \in \mathcal{T} \text{ for all constant } c_z \in \mathbb{R}. \end{aligned} \tag{H.6}$$

The null hypothesis (H.6) posits that the QTEs are constant within subgroup. However, in contrast to (H.5) we do not test for constant QTEs for all subgroups, but rather separately for each subgroup. Thus, this test can identify the subgroups that exhibit heterogeneity of QTE while accounting for dependencies both between quantiles and for each $z \in \mathcal{Z}$. This test differs from the test of hypothesis (H.5) above since we do not condition on z and test if treatment effect heterogeneity disappears, but we rather test for treatment effect heterogeneity separately for each z.[34]

We consider the following test statistic:

$$T_n(z) = \sup_{\tau \in \mathcal{T}} \left| \hat{q}^{\Delta}_{\tau}(z) - \bar{q}^{\Delta}(z) \right| , \tag{10}$$

which is equal to the test statistic (5) with QTEs calculated by subgroup. As before, we follow Romano and Wolf (2005a) and eliminate the subgroups, for which we cannot reject the null hypothesis (H.6) in a step-down procedure. Then the remaining subgroups (if any) are the ones for which we reject the null hypothesis of no treatment effect heterogeneity. The

[34] This test nevertheless differs from testing for treatment effect heterogeneity (hypothesis (H.3) for the entire sample) separately for each subgroup since we use the Romano and Wolf (2005a) approach to identify the subgroup(s) that exhibit heterogeneity in QTEs.

bootstrap test statistic is defined as

$$T^*_{n,b}(\mathcal{Z}_1) = \max_{z \in \mathcal{Z}_1} \sup_{\tau \in \mathcal{T}} \left| \hat{q}^{\Delta *}_{\tau}(z) - \hat{q}^{\Delta}_{\tau}(z) - \left\{ \bar{q}^{\Delta *}(z) - \bar{q}^{\Delta}(z) \right\} \right|, \tag{11}$$

where we first take $\mathcal{Z}_1 = \mathcal{Z}$. We then find bootstrap critical values $\hat{c}_{z,1}$ for each subgroup z such that

$$\frac{1}{B} \sum_{b=1}^{B} \mathbf{1} \left\{ T^*_{n,b}(\mathcal{Z}_1) > \hat{c}_{z,1} \right\} \leq \alpha$$

and define

$$\mathcal{Z}_2 = \left\{ z : T_n(z) \leq \hat{c}_{z,1} \right\},$$

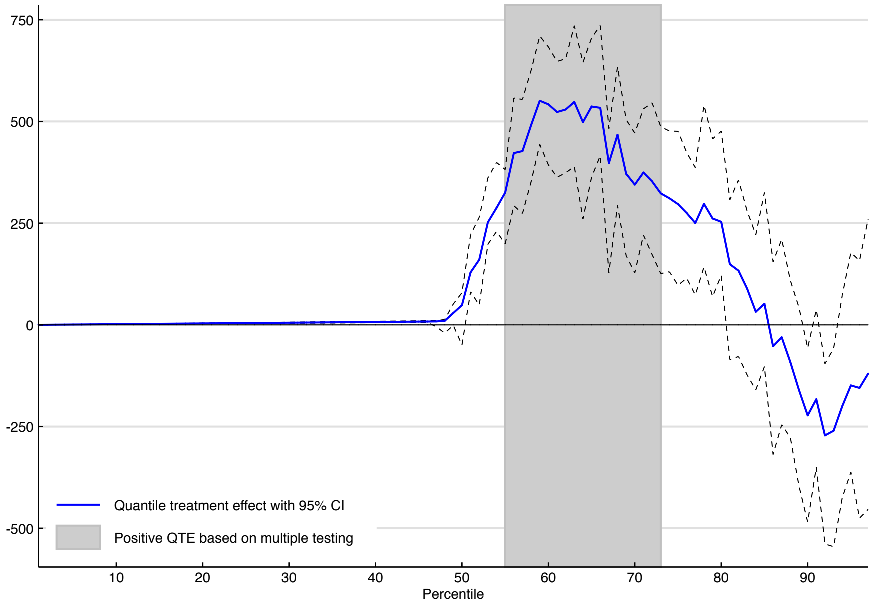
i.e. $\mathcal{Z}_2$ is the set of subgroups, for which the test statistic (10) does not exceed the critical value. Hence $z \in \mathcal{Z}_2$ are subgroups that do not exhibit significant treatment effect heterogeneity. We then repeat these steps with $\mathcal{Z}_2$, find a critical values $\hat{c}_{z,1}$ analogously, and so on, until no more subgroup is eliminated (resulting in the set of subgroups $\mathcal{Z}_k$). Hence, there is evidence for treatment effect heterogeneity for subgroups $z \in \mathcal{Z} \backslash \mathcal{Z}_k$. In the results section below, we report test statistics and p-values for each subgroup z for different subgroup categories.

5 Empirical Application

In this section, we use data from the Jobs First experiment to conduct the battery of tests presented in the preceding section. Figure 2 shows our estimated QTEs for the full sample along with point-wise 95 percent confidence intervals. Similar to Bitler, Gelbach, and Hoynes (2006) we find point-wise significant treatment effects for about the 50th to 80th percentiles.[35] Above the 80th percentile the point estimates for treatment effects become negative but the point-wise confidence intervals mostly include zero. Hence, the shape of the estimated QTEs align with the theoretical prediction in Section 3.

To shed light on the range of the earnings distribution where positive treatment effects are located, we test hypothesis (H.2) that accounts for potential dependencies across quantiles of the same outcome variable. The shaded area in Figure 2 corresponds to the set $\mathcal{T} \backslash \mathcal{T}_k$, i.e. the percentiles where the treatment effect remains significant using a FWER of five percent. Examining the plot we observe that the set of significantly positive QTEs supports the distributional effects predicted by labor supply theory. However, we find that individuals located between the 50th and 55th and the 70th and 80th percentiles of the earnings distribution do not exhibit significant QTEs once we adjust for multiple testing. Hence, we can conclude

[35] Our results look slightly different from the QTEs shown in Bitler, Gelbach, and Hoynes (2006, Figure 3) because we use Firpo's (2007) control function approach as described in Section 4.1.



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 4.1.2).

Figure 2: Quantile Treatment Effects and Multiple Testing Results, No Subgroups

that the benefits of this particular welfare reform are more confined than one would otherwise find based on standard statistical inference. Given the predictions derived in Section 3, we find that there is a more limited range of individuals who increase their labor supply when assigned to the Jobs First group.

Table 2 summarizes the results from the full set of tests proposed in Section 4.1. In the first two columns, testing the hypothesis (H.1) indicates that we can reject the null hypothesis with a *p*-value of 0.002. Thus, there is clear evidence that the welfare experiment had the desired effect of increasing earnings for at least some individuals. In the next two columns we present results from the test of the presence of treatment effect heterogeneity across quantiles (H.3) and also reject this null hypothesis with a *p*-value of 0.003. This result implies that treatment effects are heterogenous across quantiles indicating that individuals vary in their response to welfare reform.

Table 2: Testing for Presence of QTEs and QTE Heterogeneity Without Subgroups

	Test of (H.1)		Test of (H.3)	
	Test statistic	*p*-value	Test statistic	*p*-value
No subgroups	550.9236	0.002	445.9536	0.003

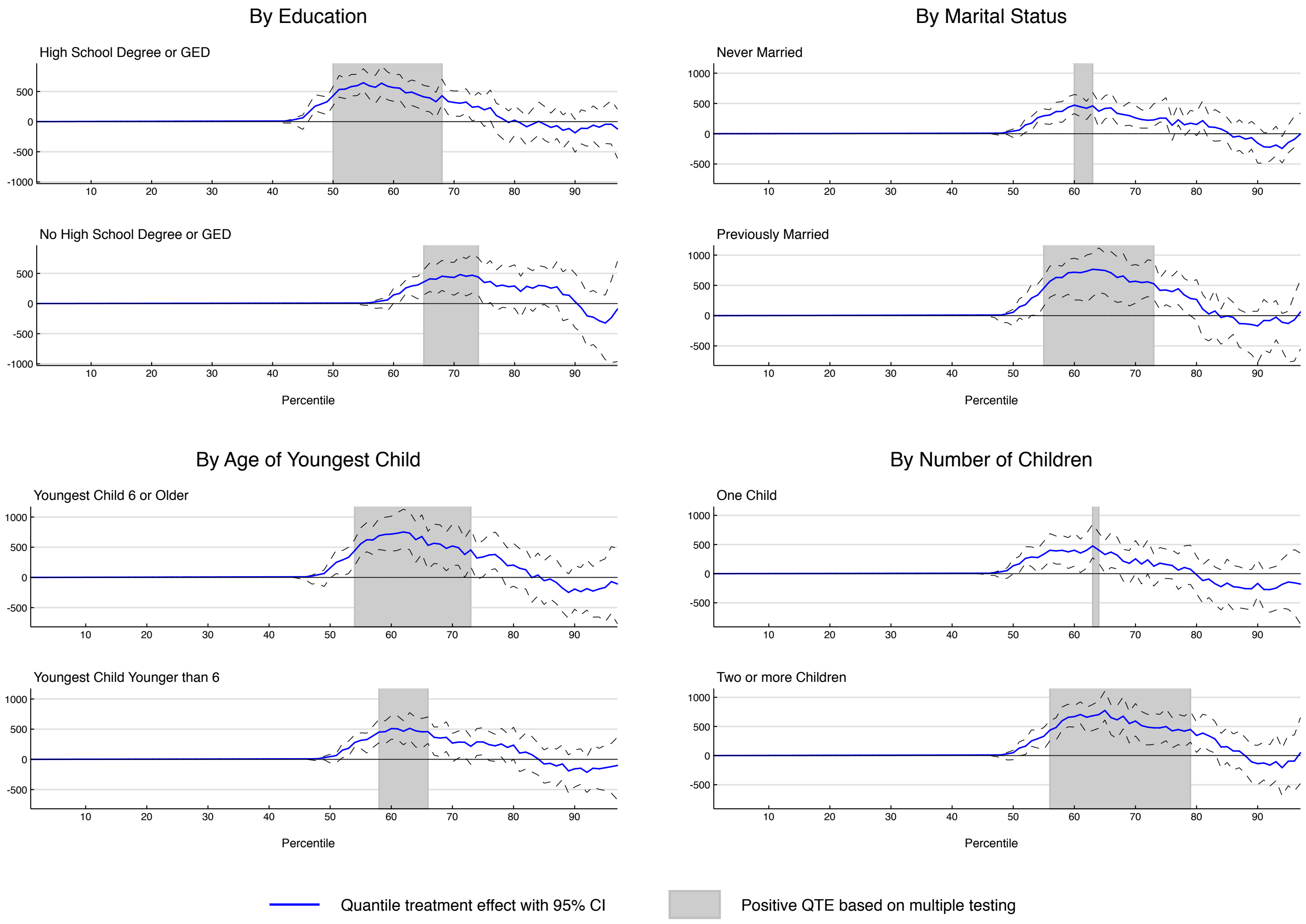
Incorporating subgroups, Figures 3 and 4 present QTEs conditional on demographic observables and individuals' labor market and welfare histories. We present results from tests of hypothesis (H.4) and shaded areas denote significant QTEs based on our multiple testing procedure. These figures provide an easy and intuitive way to check which subgroups and – within subgroups – individuals over which earnings range benefit from the welfare reform.

First, we split the sample by observable characteristics that may determine single mothers' wage offers (education) and labor supply (marital status, age and number of children). The multiple testing results illustrated in Figure 3 show that women with a high school degree, who were previously married, those with older and with two or more children, respectively, have higher earnings under Jobs First than AFDC over a wider range of the earnings distribution. These results confirm the theoretical predictions from Section 3. Better educated women may benefit more from the generous earnings disregards under Jobs First, but some of them may decrease their labor supply to take advantage of the reform. Both predictions are visible in the shape of the jointly significant QTEs. Single mothers with young children are more restricted in their time allocation. The wider range of significant QTEs among mothers with two or more children may be due to the welfare rules that make benefits a function of family size. These results are important because they can show policymakers which subgroups can be targeted with a welfare reform such as Jobs First.

We now move to individual characteristics that are not demographics but reflect choices before random assignment, in particular past earnings and welfare receipt.[36] Bitler, Gelbach, and Hoynes (2014) find that subgroup-specific constant treatment effects by previous earnings and welfare receipt come closest in explaining the observed QTEs in the entire sample. Figure 4 shows the QTEs and multiple testing results for subgroups defined by earnings and welfare receipt before the experiment. The only subgroups which exhibit jointly significant QTEs are those with either no earnings or with the highest levels of welfare receipt before random assignment. Compared to the results for the whole sample in Figure 2, women in these subgroups benefit from the reform in higher ranges of the earning distribution, roughly between the 60th and 80th percentile.

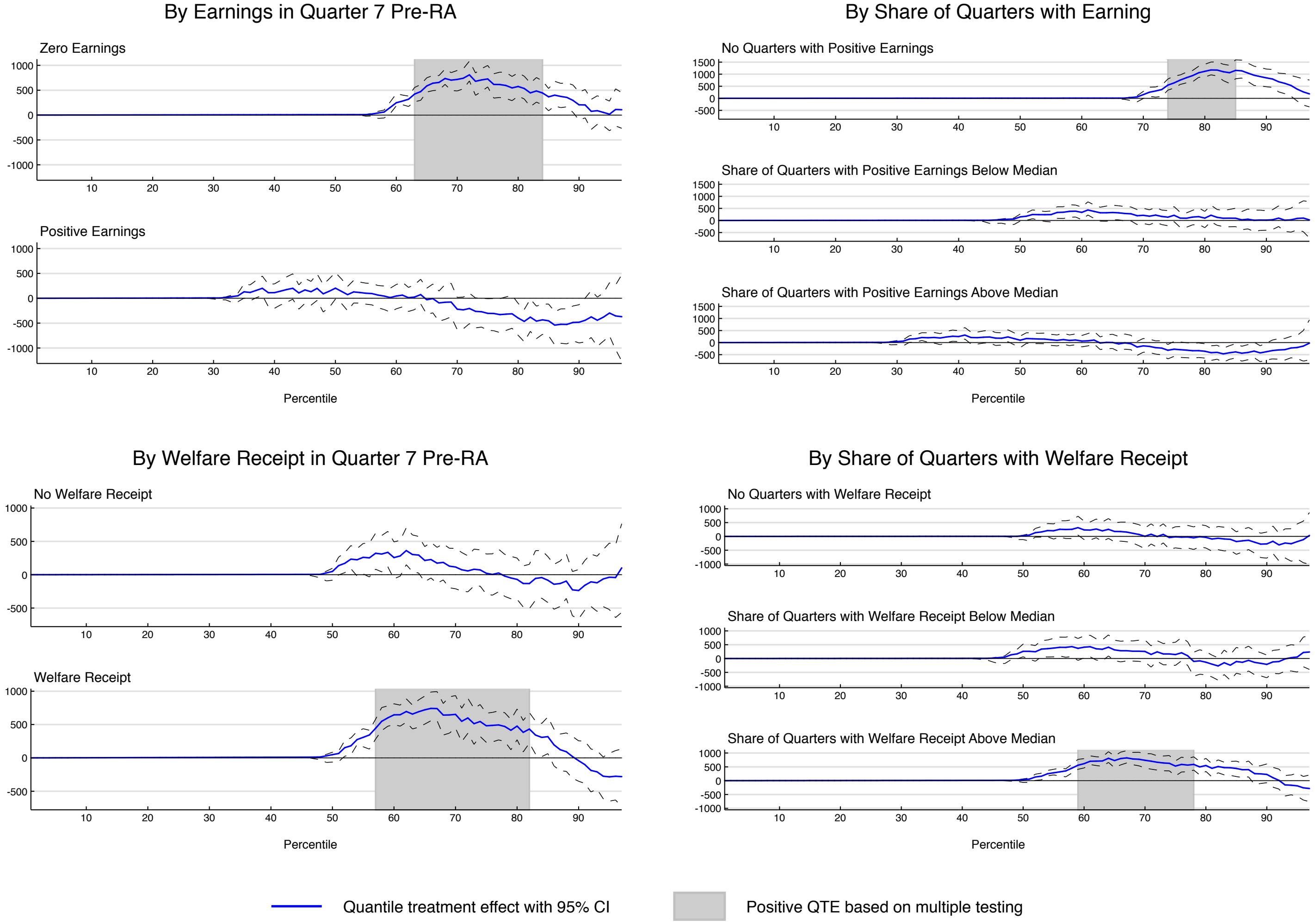
The results for both subgroups confirm labor supply theory. Welfare recipients who were not employed before participating in the Jobs First experiment, but instead relied on welfare, benefit the most from this policy. They move from non-employment to a point on the budget constraint where they have positive earnings and may take advantage of the generous earnings disregards under the new welfare rules. On the other hand, individuals who had positive earnings before the experiment are located further left on the budget constraint and may increase their labor supply only a little. Those with high earnings may reduce their labor supply to become eligible for Jobs First. Our results have clear policy implications as

[36] Note that Heckman et al. (1998) provide evidence that groups based on pre-treatment earnings are a better predictor of treatment effect heterogeneity than groups based on standard demographic variables.



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 4.2.1).

Figure 3: Quantile Treatment Effects and Multiple Testing Results, Demographic Subgroups


By Earnings in Quarter 7 Pre-RA

Zero Earnings

Positive Earnings

By Share of Quarters with Earning

No Quarters with Positive Earnings

Share of Quarters with Positive Earnings Below Median

Share of Quarters with Positive Earnings Above Median

By Welfare Receipt in Quarter 7 Pre-RA

No Welfare Receipt

Welfare Receipt

By Share of Quarters with Welfare Receipt

No Quarters with Welfare Receipt

Share of Quarters with Welfare Receipt Below Median

Share of Quarters with Welfare Receipt Above Median

1500
1000
500
0
-500
-1000
10
20
30
40
50
60
70
80
90
Percentile
Quantile treatment effect with 95% CI
Positive QTE based on multiple testing

Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 4.2.1).

Figure 4: Quantile Treatment Effects and Multiple Testing Results, Earnings and Welfare History Subgroups

they show that a substantial share of the most disadvantaged women benefit from this reform as indicated by our multiple testing results.

Table 3 presents the results for hypotheses (H.5) and (H.6) for the same subgroups as above. We can reject the null hypothesis (H.5) for all but one set of subgroups at a FWER of five percent. This null hypothesis posits that there are no differences across subgroups that can explain the observed heterogeneity of QTEs in the full sample. Hence, we can conclude that differences across subgroups do not explain the observed distributional treatment effects in the whole sample. While this result may appear similar to Bitler, Gelbach, and Hoynes (2014), our test relaxes the strong assumption of treatment effect homogeneity within subgroups that is implicit in their test. Tests of hypothesis (H.6) present nearly identical results but additionally account for potential dependencies within and across subgroups. We find that the only subgroup categories for which we do not find evidence of treatment effect heterogeneity are earnings in pre-treatment quarter seven and pre-treatment welfare receipt. These test results provide additional insight beyond testing (H.5) because they identify the individual subgroups that exhibit treatment effect heterogeneity. Overall, our results clearly suggest a substantial amount of treatment effect heterogeneity between subgroups and across the earnings distribution within subgroups.

6 Conclusion

In this paper we develop six general tests for treatment effect heterogeneity in settings with selection on observables. These tests allow researchers to provide policymakers with guidance on complex patterns of treatment effect heterogeneity both within and across subgroups. In contrast to much of the existing literature, these tests make corrections for multiple testing and therefore provide valid inference under dependence between subgroups and quantiles. Our tests complement Chernozhukov, Fernandez-Val, and Melly (2013) who show how to construct confidence sets to test functional hypotheses such as no-effect, positive effect, or stochastic dominance by shedding additional insights for which quantiles and subgroups the treatment effect is positive. In addition, our tests generalize the idea of tests considered in Bitler, Gelbach, and Hoynes (2014) by not restricting treatment effects to be constant across quantiles within a subgroup when trying to determine if the distributional heterogeneity across the full sample is characterized by subgroups.

Using data from the Jobs First experiment we not only present considerable evidence of treatment effect heterogeneity for most subgroups but show in which subgroups and which earnings quantiles within subgroups the benefits of welfare reform are highest. In addition, our empirical analysis emphasizes the importance of correcting for multiple testing. Testing across different subgroups is policy relevant, and while Crump et al. (2008) provide an approach to select which subpopulations to study, our tests go further by considering treatment effect heterogeneity conditional on observable characteristics.

Table 3: Testing for QTE Heterogeneity for Some Subgroups and Between-Subgroup QTE Heterogeneity

	Test of (H.5)		Test of (H.6)	
	Test statistic	p-value	Test statistic	p-value
Education	393.7325	0.01		
High school/GED			488.7018	0
No high school/GED			374.4823	0.025
Marital status	389.1758	0.002		
Never married			359.6945	0.005
Previously married			574.7888	0.005
Age of youngest child	423.2306	0.006		
6 or older			568.9984	0
Younger than 6			391.3454	0
Number of children	425.693	0.001		
1 child			388.0727	0.01
2 or more children			577.1826	0
Earnings in 7th quarter pre-treatment	243.2878	0.0581		
Zero earnings			629.6048	0.01
Positive earnings below median			234.6643	0.6
Positive earnings above median			224.6454	0.6
Share of quarters with positive earnings pre-treatment	340.212	0.003		
No quarters with positive earnings			1021.6	0
Share below median			332.8	0.02
Share above median			309.3	0.02
Mother on welfare in 7th quarter pre-treatment	325.9401	0.01		
Not on welfare			306.0184	0.035
On welfare			535.1389	0
Share of quarters on welfare pre-treatment	306.6839	0.01		
No quarters on welfare			278.0271	0.2
Share below median			343.2006	0.2
Share above median			594.9873	0

Note: p-values for the test of (H.6) are calculated using a grid with step size 0.005. Hence an entry of zero indicates that the corresponding p-value is below 0.005.

References

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

Abadie, Alberto, Joshua Angrist, and Guido Imbens. 2002. “Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings.” *Econometrica* 70 (1):91–117.

Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes. 2006. “What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments.” *American Economic Review* 96 (4):988–1012.

———. 2014. “Can Variation in Subgroups’ Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment.” Tech. rep.

Bloom, Dan, Susan Scrivener, Charles Michalopoulos, Pamela Morris, Richard Hendra, Diana Adams-Ciardullo, Johanna Walter, and Wanda Vargas. 2002. “Jobs First. Final Report on Connecticut’s Welfare Reform Initiative.” Tech. rep.

Chernozhukov, Victor, Ivan Fernandez-Val, and Blaise Melly. 2013. “Inference on Counterfactual Distributions.” *Econometrica* 81 (6):2205–2268.

Crump, Richard K, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. 2008. “Nonparametric Tests for Treatment Effect Heterogeneity.” *Review of Economics and Statistics* 90 (3):389–405.

Deaton, Angus S. 2009. “Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economi Development.” *NBER Working Paper* 14690.

Fink, Günther, Margaret McConnell, and Sebastian Vollmer. 2014. “Testing for heterogeneous treatment effects in experimental data: false discovery risks and correction procedures.” *Journal of Development Effectiveness* 6 (1):44–57.

Firpo, Sergio. 2007. “Efficient semiparametric estimation of quantile treatment effects.” *Econometrica* 75 (1):259–276.

Friedlander, Daniel and Philip K Robins. 1997. “The Distributional Impacts of Social Programs.” *Evaluation Review* 21 (5):531–553.

Haskins, Ron. 2006. *Work Over Welfare: The Inside Story of the 1996 Welfare Reform Law* . Brookings Institution Press.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66 (5):1017–1098.

Heckman, James J. 2001. "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture." *The Journal of Political Economy* 109 (4):673–748.

Heckman, James J, Jeffrey Smith, and Nancy Clements. 1997. "Making The Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts." *The Review of Economic Studies* 64 (4):487–535.

Heckman, James J and Sergio Urzua. 2010. "Comparing IV with structural models: What simple IV can and cannot identify." *Journal of Econometrics* 156 (1):27–37.

Imbens, Guido W. 2009. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." :1–32.

Keane, Michael. 2011. "Labor Supply and Taxes: A Survey." *Journal of Economic Literature* 49 (4):961–1075.

Kline, Patrick and Melissa Tartari. 2013. "Bounding the Labor Supply Responses to a Randomized Welfare Experiment: A Revealed Preference Approach." *working paper* :1–62.

Lee, Sokbae, Kyungchul Song, and Yoon-Jae Whang. 2014. "Testing for a General Class of Functional Inequalities." *working paper* :1–147.

Lee, Soohyung and Azeem M Shaikh. 2014. "Multiple Testing and Heterogeneous Treatment Effects: Re-Evaluating the Effect of PROGRESA on School Enrollment." *Journal of Applied Econometrics* 29 (4):612–626.

Maier, Michael. 2011. "Tests For Distributional Treatment Effects Under Unconfoundedness." *Economics Letters* 110 (1):49–51.

Mincer, Jacob A. 1974. *Schooling, Experience, and Earnings* . National Bureau of Economic Research.

Romano, Joseph P and Azeem M Shaikh. 2010. "Inference for the Identified Set in Partially Identified Econometric Models." *Econometrica* 78 (1):169–211.

———. 2012. "On the uniform asymptotic validity of subsampling and the bootstrap." *The Annals of Statistics* 40 (6):2798–2822.

Romano, Joseph P, Azeem M Shaikh, and Michael Wolf. 2010. "Hypothesis Testing in Econometrics." *Annual Review of Economics* 2 (1):75–104.

Romano, Joseph P and Michael Wolf. 2005a. “Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing.” *Journal of the American Statistical Association* 100 (469):94–108.

———. 2005b. “Stepwise Multiple Testing as Formalized Data Snooping.” *Econometrica* 73 (4):1237–1282.

Rothe, Christoph. 2010. “Nonparametric Estimation of Distributional Policy Effects.” *Journal of Econometrics* 155 (1):56–70.

Saez, Emmanuel. 2010. “Do Taxpayers Bunch at Kink Points?” *American Economic Journal: Economic Policy* 2 (3):180–212.

Smith, Jeffrey A and Petra E Todd. 2005. “Does matching overcome LaLonde’s critique of nonexperimental estimators?” *Journal of Econometrics* 125 (1-2):305–353.

Solon, Gary, Steven J Haider, and Jeffrey M. Wooldridge. 2013. “What Are We Weighting For?” *NBER Working Paper* :1–29.

White, Halbert. 2000. “A Reality Check for Data Snooping.” *Econometrica* 68 (5):1097–1126.