NBER WORKING PAPER SERIES

TARGETING POLICIES: MULTIPLE TESTING AND DISTRIBUTIONAL TREATMENT EFFECTS

Steven F. Lehrer
R. Vincent Pohl
Kyungchul Song

Working Paper 22950
http://www.nber.org/papers/w22950

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2016

This paper was previously circulated under the title "Reinvestigating How Welfare Reform Influences Labor Supply: A Multiple Testing Approach." The data used in this paper are derived from data files made available to researchers by the Manpower Demonstration Research Corporation (MDRC). The authors remain solely responsible for how the data have been used or interpreted. We are grateful to MDRC for providing the public access to the data used here. We wish to thank Jonah Gelbach, Pat Kline, Jeff Smith, and seminar and conference participants at the University of Georgia, Hunter College, the University of North Carolina Greensboro, Sciences Po Paris, AEA, CLSRN, the Econometric Society Australasian Meeting and North American Summer Meeting, and SOLE/EALE for helpful comments and suggestions. Jacob Schwartz and Thor Watson provided excellent research assistance. Lehrer and Song respectively thank SSHRC for research support. The usual caveat applies. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Targeting Policies: Multiple Testing and Distributional Treatment Eects
Steven F. Lehrer, R. Vincent Pohl, and Kyungchul Song
NBER Working Paper No. 22950
December 2016
JEL No. C12,C21,I38,J22

ABSTRACT

Economic theory often predicts that treatment responses may depend on individuals' characteristics and location on the outcome distribution. Policymakers need to account for such treatment effect heterogeneity in order to efficiently allocate resources to subgroups that can successfully be targeted by a policy. However, when interpreting treatment effects across subgroups and the outcome distribution, inference has to be adjusted for multiple hypothesis testing to avoid an overestimation of positive treatment effects. We propose six new tests for treatment effect heterogeneity that make corrections for the family-wise error rate and that identify subgroups and ranges of the outcome distribution exhibiting economically and statistically significant treatment effects. We apply these tests to individual responses to welfare reform and show that welfare recipients benefit from the reform in a smaller range of the earnings distribution than previously estimated. Our results shed new light on effectiveness of welfare reform and demonstrate the importance of correcting for multiple testing.

Steven F. Lehrer
School of Policy Studies
and Department of Economics
Queen's University
Kingston, ON K7L 3N6
CANADA
and NBER
lehrers@queensu.ca

Kyungchul Song
Vancouver School of Economics
University of British Columbia
997 - 1873 East Mall
Vancouver, BC
Canada, V6T 1Z1
kysong@mail.ubc.ca

R. Vincent Pohl
Department of Economics
Terry College of Business
University of Georgia
310 Herty Drive
Athens, GA 30602
vincent.pohl@gmail.com

A Online appendix is available at http:///post.queensu.ca/~lehrers/app_temult.pdf

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. For example, welfare programs that provide work incentives may affect welfare recipients differently according to their demographic and socio-economic characteristics such as education or number and ages of children. In addition, individuals' response to a particular treatment may vary across quantiles of the unconditional outcome distribution. After all, welfare programs induce kinks in the recipients' budget constraint, so the treatment effect may also vary depending on the location of their pre-treatment earnings.

This diverse and heterogeneous 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 this effect applies 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.[3] Lee and Shaikh (2014) address this issue in their study of data from a randomized experi-

[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.

[3] 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 confidence level. However, if these subsamples depend on each other, more than five rejections may occur. Hence, the type I error would exceed the nominal level of the test (see, e.g., Romano, Shaikh, and Wolf, 2010). The same issue arises when testing a hypothesis across the percentiles of an outcome variable, as we discuss in the text below.

ment by adopting a multiple testing procedure for subgroup treatment effects that controls the family-wise error rate (FWER) in finite samples.[4]

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 variable, i.e. they estimate quantile treatment effects (QTEs) (e.g., Heckman, Smith, and Clements, 1997; Friedlander and Robins, 1997; Abadie, 2002; Bitler, Gelbach, and Hoynes, 2006; Firpo, 2007). Individual test statistics at different quantiles involve their sample counterparts across different quantiles, which are correlated. A naive approach of comparing individual test results to find quantile groups with positive and statistically significant treatment effects inevitably suffers from the issue of data mining due to the reuse of the same data as emphasized by White (2000).[5]

In this paper, we develop a multiple testing procedure to analyze different dimensions of 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.[6] Controlling the FWER in multiple comparisons across different quantiles is crucial for the validity of the inference procedure, as estimated treatment effects across different quantiles of the outcome distribution are not independent.[7]

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.[8] This information can

[4]The FWER is defined as the probability of falsely rejecting at least one true hypothesis when performing multiple tests.

[5]In part as a response, statistical inference procedures developed in Heckman, Smith, and Clements (1997), Abadie (2002), Rothe (2010) and Maier (2011), among others, focus on the whole distribution of potential outcomes to side-step multiple comparisons.

[6]More specifically, our procedure involves multiple inequalities of unconditional quantile functions, and draws on a bootstrap method for testing for inequality restrictions. 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 unconditional quantiles.

[7]Lee and Shaikh (2014) also 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 observed variables. Hence it accommodates non-experimental data whenever the assumption is deemed plausible.

[8]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

offer important insights about how scarce social resources are to be distributed. 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 are constrained by other factors such as childcare needs. Moreover, policymakers can design welfare programs more effectively if they know over which ranges 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. To illustrate the tests we explore the extent of heterogeneity in labor supply responses in the Jobs First welfare experiment across percentiles 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).[9] A follow-up paper by Bitler, Gelbach, and Hoynes (forthcoming) 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 (forthcoming), 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 not only relax assumptions but 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 positive and statistically significant.

In addition, we make an important methodological contribution to the literature that tests for treatment effect heterogeneity. While Bitler, Gelbach, and Hoynes (forthcoming) 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 (2015) has shown that even when unconfoundedness holds (or with experimental data), researchers

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.

[9] In related work, Kline and Tartari (2016) 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.

who estimate models that do not account for heterogeneous effects may provide inconsistent estimates of average effects.[10]

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 present a simple labor supply model that predicts heterogenous treatment effects both within and across subgroups. We next describe the general testing procedures for treatment effect heterogeneity without and with subgroups 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 half of the QTEs become statistically insignificant when we account for potential dependencies. Taken together, our results shed new light on the effectiveness of welfare reform, further indicating how the composition of the labor force changes in response to public policy. The concluding Section 6 discusses the benefits of this approach and discusses directions for future research.

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.[11] 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, conducted randomized experiments to provide an evidence base for subsequent reforms as well as to receive a waiver from the federal government which allowed state governments to implement their own version of TANF.

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 who were either new welfare applicants or had applied for a continued receipt of benefits. Participants were randomly assigned to either receive a new program called Jobs First, which was the basis of the subsequent TANF program, while

[10] 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.

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

participants assigned to the control group received the original AFDC benefits. In contrast to AFDC, Jobs First imposed a time limit of 21 months on welfare receipt. In addition, participants assigned to the Jobs First group were required to attend job training programs or provide proof of job search activities to remain eligible for benefits. On the other hand, Jobs First included more generous earnings disregards. Specifically, under Jobs First, all earnings up to the federal poverty level (FPL) were disregarded, whereas AFDC participants faced an implicit tax rate of 49 percent during the first three months of employment and 73 percent thereafter.[12] 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. The Jobs First experiment is well-studied (Bitler, Gelbach, and Hoynes, 2006, forthcoming; Kline and Tartari, 2016, among others). We use it to illustrate the methods proposed in the paper because it facilitates comparisons with the existing literature that used the same data.

Summary statistics are reported in Table 1 where the second and third columns present characteristics of those women respectively assigned to the Jobs First and AFDC 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 be additional constraints on 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 standard deviation of earnings is high relative to the mean, suggesting heterogeneous responses to different welfare policies across the earnings distribution. The last column in Table 1 contains p-values for the test that individuals assigned to the Jobs First and AFDC groups do not differ in observed characteristics. For most characteristics and as shown in 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. Specifically, 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 ensure covariate balance we make adjustments via propensity score weighting in our analyses.

[12] 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 Experiment

	Jobs First Mean (Std.dev.)	AFDC Mean (Std.dev.)	Difference p-value
Mother's age < 20	0.089	0.086	0.684
Mother's age 20 to 29	0.214	0.216	0.898
Mother's age $\geqslant 30$	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 $\geqslant 6$	0.395	0.386	0.520
Number of children	1.649	1.591	0.037
	(0.932)	(0.944)	
Mean quarterly earnings pre-RA	682.7	796.0	0.006
	(1304.1)	(1566.0)	
Mean quarterly welfare benefits pre-RA	890.8	835.1	0.015
	(806.0)	(784.8)	
Mean quarterly foods stamp benefits pre-RA	352.1	339.4	0.156
	(320.0)	(303.9)	
Fraction of quarters employed pre-RA	0.327	0.357	0.006
	(0.370)	(0.379)	
Fraction of quarters welfare receipt pre-RA	0.573	0.544	0.026
	(0.452)	(0.450)	
Fraction of quarters food stamps receipt pre-RA	0.607	0.598	0.486
	(0.438)	(0.433)	
Observations	2396	2407	4803

Source: Manpower Demonstration Research Corporation's study of Connecticut's Jobs First Program. Note: p-values are obtained from two sided t-tests of the equality of means between the Jobs First and AFDC groups.

3 Economic Model Predicting Heterogeneous Treatment Effects

A simple static labor supply model motivates our investigation of treatment effect heterogeneity.[13] Individuals maximize their utility over consumption (C) and earnings (E) subject to a budget constraint:

$$\max_{C,E} U = U\left(C, E(X_1); X_2\right) \quad (1)$$

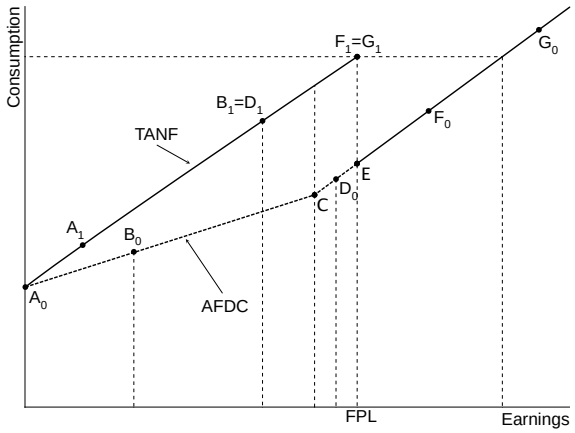
$$\text{s.t. } C = E(X_1) + W\left(E(X_1); X_3, Z^t\right) \quad (2)$$

where X_1 and X_2 denote characteristics that may affect earnings and individual preferences, respectively, and $W(\cdot)$ denotes the welfare benefit function, which depends on the level of earnings $E(\cdot)$, household characteristics X_3, 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 Saez (2010), we assume that the marginal utility of consumption is positive ($\frac{\partial U}{\partial C} > 0$) and the marginal utility of earnings is negative ($\frac{\partial U}{\partial E} < 0$).

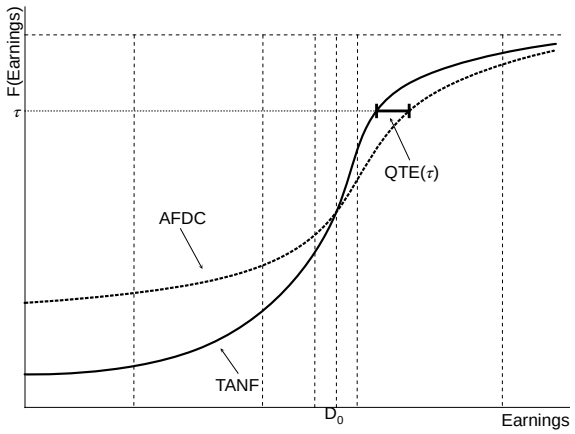
We use the panels of Figure 1 to demonstrate how economic theory predicts treatment effect heterogeneity. This heterogeneity arises because there is a differential labor supply response on both the intensive and extensive margins between the AFDC and the Jobs First program due to different budget constraints and earnings distributions in the two experimental groups.[14] The solid line in the top panel of Figure 1 illustrates the budget constraint faced by Jobs First participants and is defined by the points A_0, F_1, E and G_0. A_0 denotes the base grant amount. The segment A_0F_1 is parallel to the 45 degree line due to the implicit tax rate of 0 for welfare recipients with earnings below the FPL. The dashed line represents the budget constraint faced under AFDC and is defined by the points A_0, C and G_0, where C corresponds to the eligibility threshold, which is below the FPL. In particular, the segment A_0C represents the earnings disregard under AFDC with a positive implicit tax rate. 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 the horizontal distance at each quantile between the two earnings distributions in the middle panel.

[13] 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 (2016). We extend this literature by considering differences across subgroups.

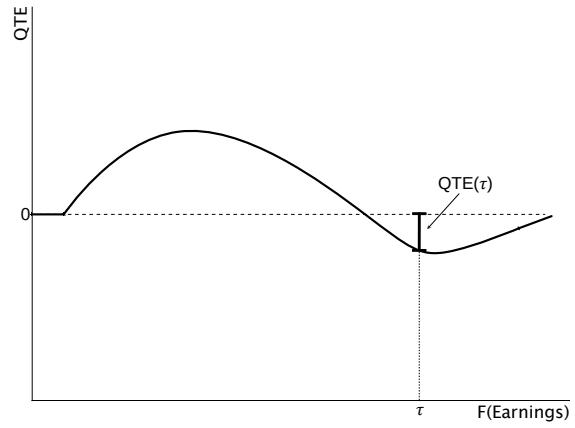
[14] We abstract from welfare stigma and the hassle associated with not working while on welfare modeled by Kline and Tartari (2016). We are interested in the distribution of earnings and how it varies by subgroup, but not in the welfare participation decision or decomposing labor supply responses into the extensive and intensive margin here.



(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 a Static Labor Supply Model

To provide intuition for the shape of the QTEs presented, we consider the thought experiment of moving individuals from AFDC to Jobs 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. The potential move from A_0 to A_1 corresponds to less mass at zero in the earnings distribution under TANF compared to AFDC in panel (b).

We next consider a woman on AFDC at point B_0 who works while receiving welfare. Transitioning to Jobs First lowers her implicit tax rate from either 49 or 73 percent to 0, therefore boosting her net wage.[15] 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 panel (c) shift from positive to negative around the point D_0, which corresponds to the earnings of women who are ineligible for welfare under AFDC but who would be eligible under Jobs First. Theory predicts moving to Jobs First would lead to a reduction in labor supply to point D_1, if we make the standard assumption that leisure is a normal good. Intuitively, by moving fro AFDC to Jobs First at point D_0 (where welfare benefits are not available under AFDC), women now gain the base grant, resulting only in an income effect.[16] Similarly, negative QTEs arise for women located at point F_0. 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 that is characterized by higher consumption and lower earnings.

Last, women with relatively high earnings under AFDC at point G_0 face a trade-off 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 remain 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 panel (c) of Figure 1.

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

[16] Note, to avoid clutter, we set $D_1 = B_1$ without loss of generality.

The above discussion concerned the general shape of treatment effect heterogeneity but did not consider subgroups. Subgroup membership denoted by X_1, X_2, and X_3 in equations (1) and (2) affects preferences and the budget constraint. Hence, the parameters in this optimization problem vary, so the resulting QTEs could be shifted to the left or right, be compressed or stretched, or otherwise be transformed without losing their overall shape depicted in panel (c) 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 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 expect an overall shift of the QTEs to the left with less mass in the lower tail of the distribution where the QTEs equal zero for higher educated women.

A similar shift is anticipated for subgroups defined by earnings and welfare history, where we also expect qualitative differences in the shape of the QTEs. Recent welfare recipients have little if any positive earnings in the period before the experiment, so there is little mass around points D_0, F_0, and G_0 relative to A_0 and B_0. Thus, we 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.

Finally, we consider subgroups defined by either the age or number of children. Additional children will mechanically influence the size of benefits because the latter increase with family size. Yet, under Jobs First the potential loss of welfare benefits when time limits are imposed might be higher for women with additional children. While it is not possible to predict differences in the range of positive and negative QTEs by the 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 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 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 experiment.

[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 either a high school degree or a GED versus no degree at all.

4 Methodology

In this section, we begin by introducing three tests for treatment effect heterogeneity in the full sample. Motivated by the discussion in the previous section, we then propose three additional tests for treatment effect heterogeneity both within and between subgroups.

4.1 Treatment Effect Heterogeneity Without Subgroups

Each test we introduce requires estimates of QTEs that in the full sample are calculated by subtracting the unconditional outcome at quantile τ for the control group from the respective outcome at quantile τ for the treatment group. To control for possible selection on observed variables into treatment and control groups, we weight the outcome variable by inverse propensity score weights (IPSW). We define the IPSW 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, where D_i is the treatment indicator, X_i is a vector of observed characteristics, and $\hat{p}(\cdot)$ is the estimated propensity score. We then obtain quantiles of the weighted outcome as follows:

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

where $\rho_\tau(x) = x \cdot (\tau - \mathbf{1}\{x \leqslant 0\})$ is the check function and n is the size of the full sample. That is, $\hat{q}_{\tau,1}$ and $\hat{q}_{\tau,0}$ are the τ-th empirical quantiles of the propensity score weighted outcome variable

$$\left\{\hat{Y}_{1i}\right\}_{i=1}^{n} = \left\{\frac{Y_i D_i}{\hat{p}(X_i)}\right\}_{i=1}^{n} \text{ and } \left\{\hat{Y}_{0i}\right\}_{i=1}^{n} = \left\{\frac{Y_i(1 - D_i)}{1 - \hat{p}(X_i)}\right\}_{i=1}^{n}$$

for treatment and control group, respectively.

Formally, the estimated QTE at τ is then defined as

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

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 unconditional outcome 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 an intervention had any detectable positive effect on the outcome of interest.[18] We consider the following null and alternative hypotheses:

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

where $\mathcal{T} \subset [0,1]$ is the finite set of quantiles considered. The alternative hypothesis states that there exists a positive treatment effect for at least one quantile. Therefore, the null hypothesis is rejected if treatment has any positive effect on some range of the outcome distribution, given reasonable power.

To develop a bootstrap test of the null hypothesis of no positive treatment effect (H.1), 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 observed QTE provides the clearest evidence against the null hypothesis (White, 2000). To implement the test, we calculate a critical value using a bootstrap method. Specifically, we first resample with replacement from the original sample B times and construct the propensity score weighted outcomes $\hat{Y}_{1i}^* = Y_i^* D_i^* / \hat{p}^*(X_i^*)$ and $\hat{Y}_{0i}^* = Y_i^*(1-D_i^*)/(1-\hat{p}^*(X_i^*))$, where $\{Y_i^*, D_i^*, X_i^*\}_{i=1}^n$ denotes each bootstrap sample and $\hat{p}^*(X_i^*)$ the estimated propensity score using the bootstrap sample. Then the bootstrap test statistic for bootstrap draw $b = \{1, \ldots, B\}$ is given by

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

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

[18] 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.

4.1.2 Testing for General Treatment Effect Heterogeneity

We now test for treatment effect homogeneity, which provides an answer to the policy-relevant question of whether individuals across quantiles differ in their response to a particular intervention. While one may obtain information from a visual inspection of QTEs across quantiles of the outcome distribution, a formal test is necessary to properly account for sampling variations.

We consider the following hypotheses:

$$\begin{aligned} H_0 &: q_\tau^\Delta = c \text{ for all } \tau \in \mathcal{T} \text{ and for some } c \in \mathbb{R} \\ H_1 &: q_\tau^\Delta \neq c \text{ for some } \tau \in \mathcal{T} \text{ and for all } c \in \mathbb{R}. \end{aligned} \tag{H.2}$$

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 treatment depending on where on the outcome distribution they are located.[19]

To test (H.2) we construct the following test statistic:

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

where $\bar{q}^\Delta = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \hat{q}_\tau^\Delta$ is the sample mean of the estimated QTEs. That is, we set the constant c in (H.2) equal to the sample mean, $\bar{q}^\Delta$, and subtract it from the estimated QTEs, so the test statistic will be small if the QTEs are very similar across τ.[20] The max 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:

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

The bootstrap test also incorporates re-centering in order to impose the null restriction. To test null hypothesis (H.2), we compare the test statistic (5) to the critical value obtained from

[19] Note that testing for treatment effect heterogeneity has appeared in Appendix E of Heckman, Smith, and Clements (1997) though in a different form. Their null hypothesis posits that the variance of the individual treatment effect $Y_{1i} - Y_{0i}$ is zero, i.e. there is no treatment effect heterogeneity across individuals. On the other hand, our null hypothesis allows for treatment effect heterogeneity across individuals. Our null hypothesis rather states that the QTEs are constant across quantiles. After all, randomness of the individual treatment effect appears to be a less interesting hypothesis to test, given that it is well accepted that individuals have heterogeneous responses to policy interventions including experiments such as PROGRESA (see, e.g., Djebbari and Smith, 2008).

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

the bootstrap test statistics (6), which is equal to the $(1-\alpha)$-th quantile of the bootstrap test statistics $T^*_{n,b}, b = 1, \ldots, B$.

4.1.3 Testing for Which Quantiles the Treatment Effect Is Positive

The next test employs a multiple testing approach to identify the ranges of the outcome distribution that exhibit positive treatment effects. This is important since rejecting null hypothesis (H.1) only informs us that individuals in some range of the outcome distribution exhibit positive QTEs, and rejecting (H.2) just provides evidence that the treatment effect is not constant across the outcome distribution. The identified 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. We follow recent developments in the multiple testing literature (see, e.g., Romano and Wolf, 2005a,b) and use a bootstrap based step-down method to identify the quantiles for which positive treatment effects are present.[21] To do so, it is necessary to update the critical value at each step, for example by using a bootstrap method. By combining bootstrap tests of inequality restrictions with multiple testing procedures, we produce a testing procedure suitable for analyzing treatment heterogeneity that controls the FWER at the desired level.

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

$$\begin{aligned} H_{0,\tau} &: q^{\Delta}_{\tau} \leqslant 0 \\ H_{1,\tau} &: q^{\Delta}_{\tau} > 0. \end{aligned} \tag{H.3}$$

Then the goal is to find a set of individual hypotheses, for which the null is false, in a way that controls the FWER.[22]

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 the smallest $\hat{c}_1$ such that

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

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

[21] Other adjustments have been proposed in the literature, starting with the very conservative Bonferroni method (see Romano, Shaikh, and Wolf, 2010, for a recent overview).

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

we retain those quantiles that do not exceed the critical value $\hat{c}_1$, i.e. we define

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

so $\mathcal{T}_2$ is a subset of $\mathcal{T}_1$. Now, we construct $T_{n,b}^*(\mathcal{T}_2) = \max_{\tau\in\mathcal{T}_2}\left\{\hat{q}_\tau^{\Delta *} - \hat{q}_\tau^\Delta\right\}$, find $\hat{c}_2$ such that

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

and define

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

This procedure is repeated until at step k, we obtain $\mathcal{T}_k = \left\{\tau \in \mathcal{T}_{k-1} : \hat{q}_\tau^\Delta \leqslant \hat{c}_{k-1}\right\}$ such that no further element of $\mathcal{T}_k$ is eliminated (i.e. $\mathcal{T}_k = \mathcal{T}_{k-1}$). Then the resulting set $\mathcal{T}_k$ is the subset of $\mathcal{T}$ such that there is no empirical support for positive and statistically significant treatment effects at quantiles $\tau \in \mathcal{T}_k$. Conversely, this procedure provides evidence for positive treatment effects at quantiles in the set $\mathcal{T}\backslash\mathcal{T}_k$. From the result of Romano and Shaikh (2010), it is not hard to show that this multiple testing procedure asymptotically controls the FWER at α.[23]

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 subgroups defined by observed variables that exhibit differential treatment effects, 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.[24]

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

[23] See the Online Appendix for a sketch of proofs for the validity of all testing procedures introduced in the paper.

[24] Our interest is not in optimal treatment assignment in the spirit of Manski (2004), Dehejia (2005), and others. Armstrong and Shen (2015) recently extended optimal treatment assignment to additionally consider multiple testing procedures for treatment effects that control for the FWER. In contrast, we do not assume that the researcher ex ante has full knowledge of the distribution of outcomes in the population. Our interest is rather to propose a multiple testing framework for identifying subpopulations with positive responses to the outcome variable.

for the equations

$$\tau = P\left\{Y_{1i} \leqslant q_{\tau,1}(z) | Z_i = z\right\} \quad \text{and}$$
$$\tau = P\left\{Y_{0i} \leqslant q_{\tau,0}(z) | Z_i = z\right\},$$

where Y_{1i} and Y_{0i} are the potential outcomes under treatment and control, respectively, and 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 (i.e. each category j corresponds to one observed variable that can take on multiple values). Hence $q_{\tau,1}(z)$ and $q_{\tau,0}(z)$ are the quantiles of the outcome variable 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 estimated propensity score $\hat{p}(X_i)$ except that Z_i is replaced by z.[25] We define the empirical quantiles of the outcome variable for subgroup z in the treatment and control group as

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

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

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

This test extends the test of hypothesis (H.3) to a setting with subgroups. That is, we identify the quantile-subgroup cells that have statistically significantly positive treatment effects. We consider the following individual hypotheses: for each $\tau \in \mathcal{T}$ and $z \in \mathcal{Z}$,

$$H_{0,\tau,z} : q_\tau^\Delta(z) \leqslant 0$$
$$H_{1,\tau,z} : q_\tau^\Delta(z) > 0. \tag{H.4}$$

[25] Following Smith and Todd (2005), the propensity score $\hat{p}(x)$ is estimated using data from the full sample.

Hence, we test a total number of $|\mathcal{T}| \times |\mathcal{Z}|$ hypotheses. We denote the set of quantile-subgroup cells by $\mathcal{W} = \mathcal{T} \times \mathcal{Z}$.

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

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

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 $Z_i^* = z$. 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. That is, (τ, z) cells that are eliminated throughout this procedure constitute the subgroup-quantile groups with evidence for positive treatment effects.

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

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

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

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

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

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

We then define

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

The process is repeated until we obtain $\mathcal{W}_k = \left\{ (\tau, z) \in \mathcal{W}_{k-1} : \hat{q}_\tau^{\Delta}(z) \leqslant \hat{c}_{k-1} \right\}$ 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; specifically they are given by quantile-subgroup pairs $(\tau, z) \in \mathcal{W} \backslash \mathcal{W}_k$.

4.2.2 Testing for Subgroup-Specific Treatment Effect Heterogeneity Across Quantiles

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 exhibit heterogeneity of treatment effects across different quantiles $\tau \in (0,1)$:

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

The null hypothesis states that the heterogeneity in treatment effects disappears when we condition on Z_i. In other words, it posits that the QTEs are constant across quantiles within all subgroups z. However, the null hypothesis still allows for treatment effect heterogeneity across different subgroups. Rejection of the null hypothesis suggests the presence of QTE heterogeneity across quantiles even after we control for Z_i. Bitler, Gelbach, and Hoynes (forthcoming) ask exactly this question. In contrast to their paper, however, we do not constrain the treatment effect to be constant within subgroups.[26]

Hypothesis (H.5) explicitly tests the validity of the above assumption by using the following test statistic:

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

where $\bar{q}^\Delta(z) = \frac{1}{|\mathcal{T}|} \sum_{\tau \in \mathcal{T}} \hat{q}_\tau^\Delta(z)$ is the sample mean of the estimated 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 estimated QTEs from their sample 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 maximum value over each subgroup's test statistic. Intuitively, we search for evidence that there exists a subgroup that exhibits treatment effect heterogeneity, so we restrict our attention to the subgroups that have the largest degree of heterogeneity.

To construct a bootstrap critical value, we consider the following bootstrap test statistic that is an analog of (6) with subgroups:

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

The test statistic in equation (8) is compared to the bootstrap critical value, which equals the $(1-\alpha)$-th quantile of the bootstrap test statistics (9) as described above. This test of hy-

[26] In the test in Section 4.2.3 below we additionally adjust for multiple testing while also relaxing the assumption of constant subgroup-specific treatment effects.

pothesis (H.5) provides a simple and flexible way to see if most treatment effect heterogeneity across quantiles is in fact due to treatment effect heterogeneity across subgroups.

4.2.3 Testing for Which Subgroups Treatment Effects Are Heterogenous

The test described in the previous section tests whether treatment effect heterogeneity across quantiles exists even after controlling for subgroups. If we reject null hypothesis (H.5), we may also be interested in identifying the subgroups that exhibit QTE heterogeneity. The final test is able to do so by exploring subgroup by subgroup, if there is treatment effect heterogeneity within subgroups. For each $z \in \mathcal{Z}$, we test

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

The null hypothesis (H.6) posits that the QTEs are constant within subgroup. 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.[27]

We consider the following test statistic:

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

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 bootstrap test statistic is defined as

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

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

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

[27] 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.

and define

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

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 additional subgroup is eliminated (resulting in the set of subgroups $\mathcal{Z}_k$). Hence, there is evidence for treatment effect heterogeneity for subgroups $z \in \mathcal{Z} \backslash \mathcal{Z}_k$.

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. Following Bitler, Gelbach, and Hoynes (2006) we use quarterly earnings pooled over the seven quarters after random assignment as our outcome variable and estimate QTEs for percentiles 1 to 97.[28] To balance covariates between the Jobs First and AFDC groups, we estimate the propensity score $\hat{p}(x)$ using a series logit specification.[29] For the results that follow, we set the level of each test to $\alpha = 0.05$. The test results for the whole sample are based on bootstraps with $B = 9999$ replications while we use $B = 999$ for the subgroup-specific tests.

Figure 2 shows our estimated QTEs for the full sample along with pointwise 90 percent confidence intervals.[30] Similar to Bitler, Gelbach, and Hoynes (2006) we find pointwise significant treatment effects extending from the 48th to the 80th percentile.[31] Above the 86th percentile the point estimates for treatment effects become negative but the pointwise confidence intervals mostly include zero. Hence, the shape of the estimated QTEs aligns with the theoretical prediction in Section 3.

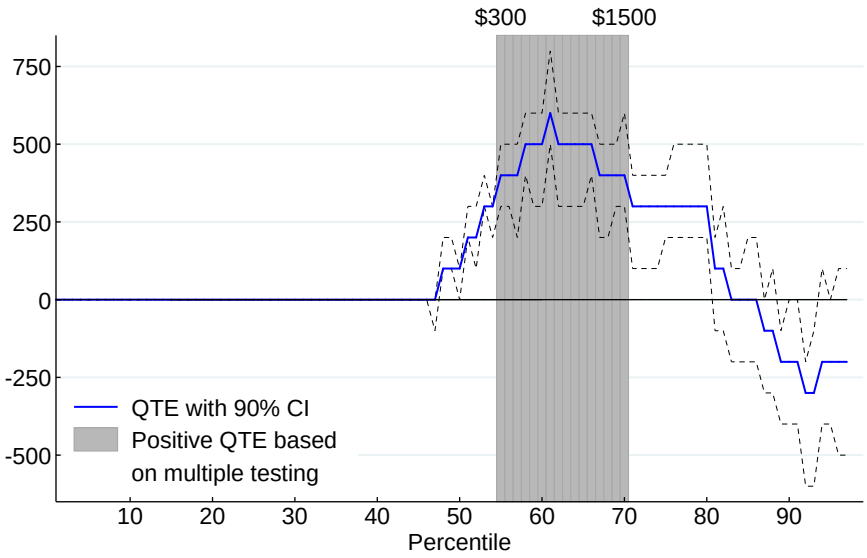
Table 2 summarizes the test result for hypotheses (H.1) and (H.2) proposed in Section 4.1. First, we test the null hypothesis of no positive treatment effect at any percentile. As shown in Figure 2, the largest QTE (which occurs at the 61st percentile) equals 600, so this value

[28] Hence, we have a total of $7 \times 4803 = 33621$ observations. 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). However, positive QTEs imply that the treatment has a positive effect for some interval of the earnings distribution (Bitler, Gelbach, and Hoynes, 2006).

[29] 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 to calculate the weights via the propensity score.

[30] We show 90 percent CI because they corresponds to a one-sided test with a level of five percent, and we implement one-sided tests that hold the FWER at that level.

[31] Our results look slightly different from the QTEs shown in Bitler, Gelbach, and Hoynes (2006, Figure 3) because we use Firpo's (2007) check function approach as described in Section 4.1 instead of estimating empirical cumulative distribution functions of Jobs First and AFDC earnings. The QTEs are in multiples of 100 because the quarterly earnings data are rounded to the closest \$100, which does not affect the validity of the results (Gelbach, 2005).



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5% (see Section 4.1.3). Dollar amounts are quarterly earnings at smallest and largest quantiles with positive and statistically significant QTE.

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

becomes the test statistic in the first row of Table 2. Comparing this test statistics to the bootstrap critical value of 300 indicates that we can reject the null hypothesis. The associated p-value equals 0.0003. Thus, there is clear evidence that the Jobs First experiment had the desired effect of increasing earnings for at least some individuals. Next, we present results from the test of no treatment effect heterogeneity across quantiles (H.2). The test statistic, which is calculated as the largest deviation from the mean estimated QTE ($\bar{q}^{\Delta} = 100$), equals 500. With a bootstrap critical value of 294.85, we also reject this null hypothesis at a p-value of 0.0017. This result implies that treatment effects are heterogenous across quantiles, thereby indicating that individuals vary in their response to welfare reform.[32]

Having rejected the null hypothesis of no treatment effect heterogeneity, we now identify the range of the earnings distribution where positive treatment effects are located, i.e. we test hypothesis (H.3). As described in Section 4.1.3, this test accounts for potential dependencies across quantiles of the same outcome variable and the number of individual hypotheses ($|\mathcal{T}| = 97$). 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 $\alpha = 0.05$. 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

[32]In the Online Appendix, we compare these test results with an alternative procedure based on Abadie (2002), which yields the same conclusions.

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

	Test statistic	Critical value	p-value
Test of (H.1)	600	300	0.0003
Test of (H.2)	500	294.85	0.0017

Notes: This table shows test results for hypotheses (H.1) and (H.2), i.e. we test that there is no positive treatment effect for all quantiles and that the treatment effect is the same for all quantiles, respectively.

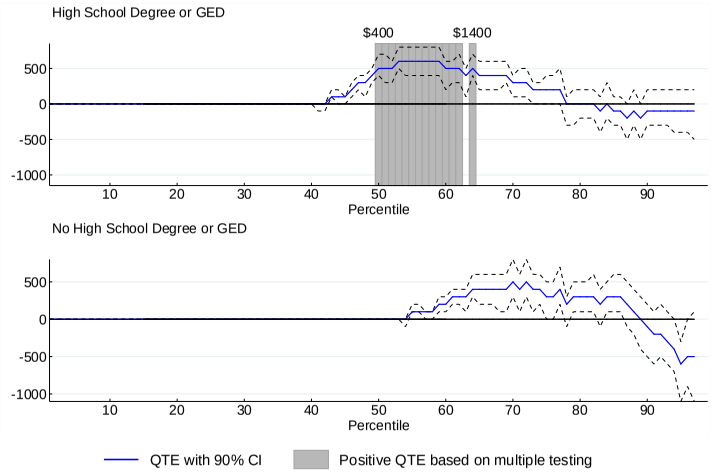
48th and 54th and the 71st and 80th percentiles of the earnings distribution do not exhibit significant QTEs once we adjust for multiple testing. The smallest and largest quantiles at which QTEs are significantly positive correspond to quarterly earnings of \$300 and \$1,500, respectively. Hence, we can conclude that the benefits of this particular welfare reform are more confined than one would otherwise find based on traditional statistical inference that ignores potential dependencies and testing at multiple percentiles. 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.[33]

Next, we present results incorporating subgroups using the tests described in Section 4.2. As discussed in Section 3, labor supply theory predicts that individuals with different observed characteristics may react differently to the same welfare rules. In particular, characteristics such as age and number of children, and prior earnings and welfare receipt may determine for which range of the earnings distribution we observe an increase or decrease in labor supply. Following Bitler, Gelbach, and Hoynes (forthcoming) and informed by the model presented in Section 3, 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.[34]

Figures 3 and 4 present QTEs conditional on demographic observables and individuals' labor market and welfare histories. Shaded areas denote significant QTEs based on our multiple testing procedure of testing hypothesis (H.4). These figures provide an easy and intuitive way to check which subgroups benefit from the welfare reform (heterogeneity across subgroups). In addition, we can inspect the figure for each subgroup to determine the range of the earnings distribution in which individuals exhibit positive subgroup-specific QTEs (heterogeneity within subgroup).

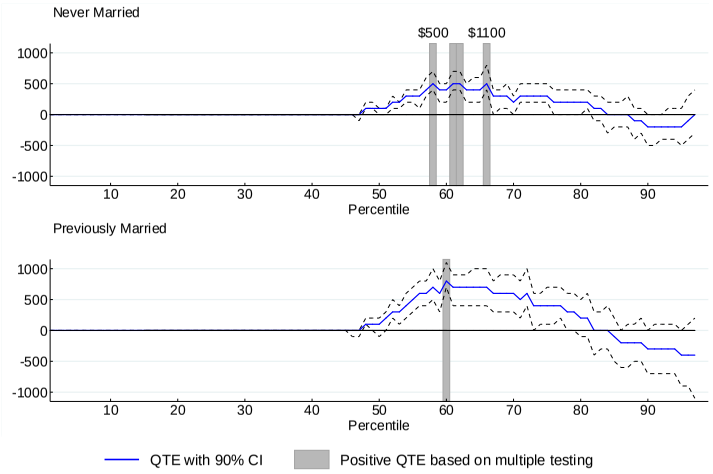
[33]We provide results using alternative testing procedures based on Bonferroni (1936), Holm (1979), and Chernozhukov, Fernandez-Val, and Melly (2013) in the Online Appendix.

[34]Note that in our application the number of hypotheses being tested is quite small particularly relative to genomic studies from genome wide association studies. If the number of hypotheses were large it is well known that FWER controlling procedures typically have low power, and in response Gu and Shen (2016) propose an optimal false discovery rate controlling method.



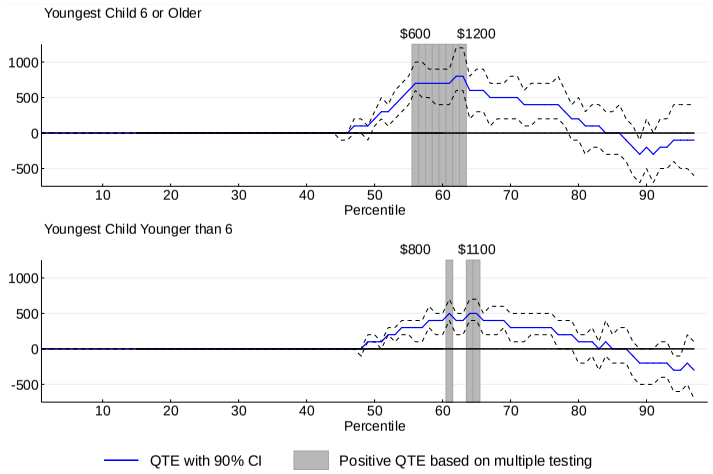
Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 4.2.1). Dollar amounts are quarterly earnings at smallest and largest quantiles with positive and statistically significant QTE.

(a) by Education



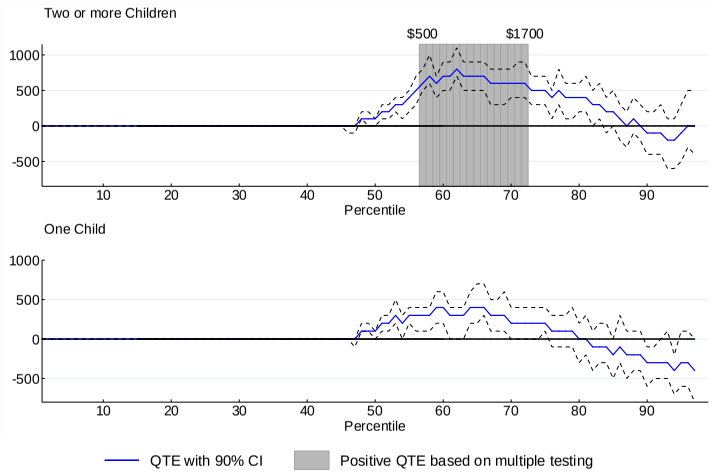
Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 4.2.1). Dollar amounts are quarterly earnings at smallest and largest quantiles with positive and statistically significant QTE.

(b) by Marital Status



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 4.2.1). Dollar amounts are quarterly earnings at smallest and largest quantiles with positive and statistically significant QTE.

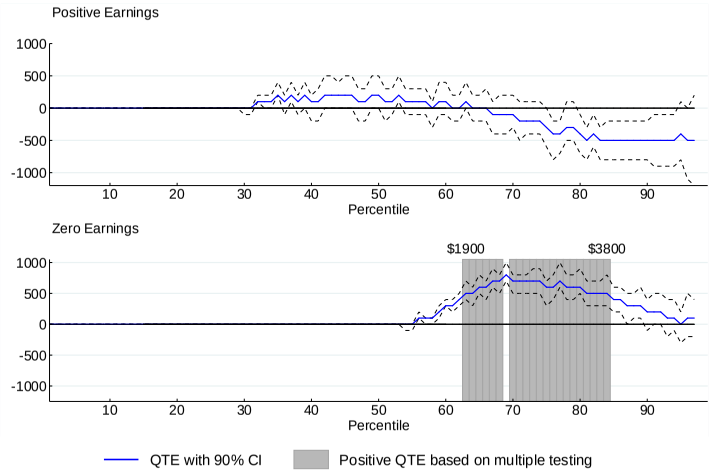
(c) by Age of Youngest Child



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 4.2.1). Dollar amounts are quarterly earnings at smallest and largest quantiles with positive and statistically significant QTE.

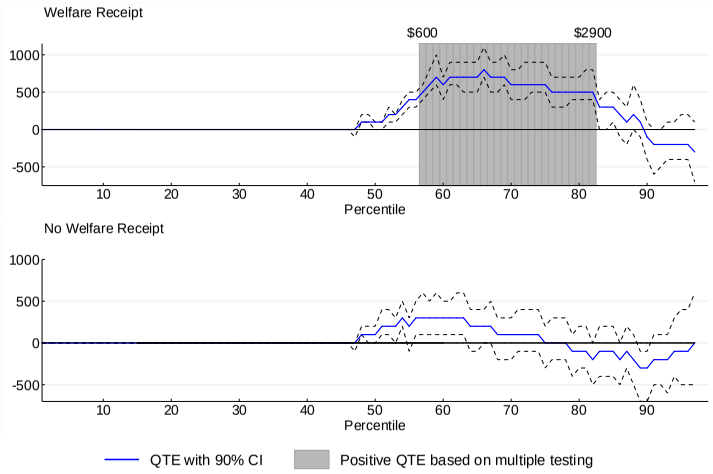
(d) by Number of Children

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



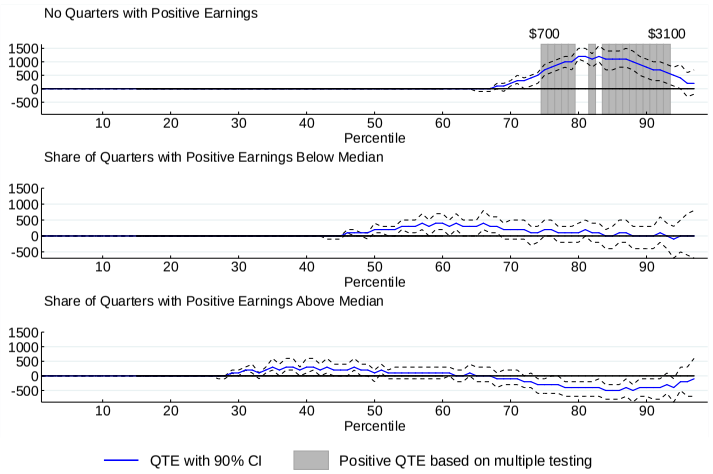
Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 4.2.1). Dollar amounts are quarterly earnings at smallest and largest quantiles with positive and statistically significant QTE.

(a) by Earnings in Quarter 7 Pre-Random Assignment



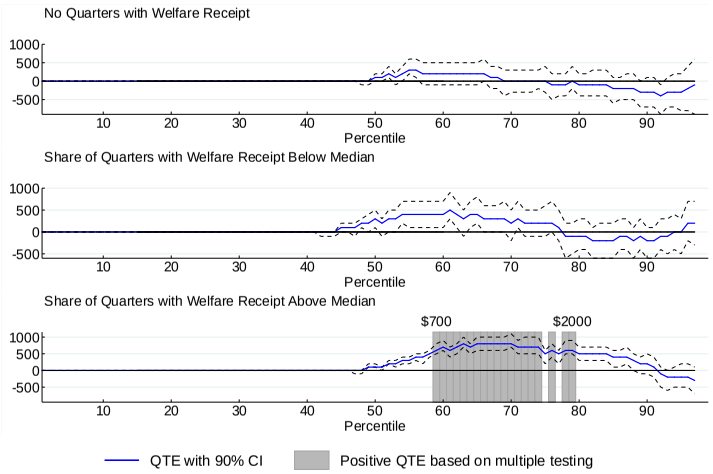
Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 4.2.1). Dollar amounts are quarterly earnings at smallest and largest quantiles with positive and statistically significant QTE.

(b) by Welfare Receipt in Quarter 7 Pre-Random Assignment



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 4.2.1). Dollar amounts are quarterly earnings at smallest and largest quantiles with positive and statistically significant QTE.

(c) by Share of Quarters with Positive Earnings



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 4.2.1). Dollar amounts are quarterly earnings at smallest and largest quantiles with positive and statistically significant QTE.

(d) by Share of Quarters with Welfare Receipt

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

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 or GED, who were never 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 who receive higher wage offers may benefit more from the generous earnings disregards under Jobs First and therefore increase their labor supply more. At the same time, women without a high school degree or GED are more likely to lower their labor supply in order to become eligible for Jobs First benefits, leading to negative QTEs in the upper range of the earnings distribution. Single mothers with young children are more restricted in their time allocation, so they are less likely to change their labor supply in response to different welfare rules. 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 should be targeted with a welfare reform such as Jobs First.

We now move to individual characteristics that reflect outcomes before random assignment, in particular past earnings and welfare receipt.[35] Bitler, Gelbach, and Hoynes (forthcoming) find that subgroup-specific constant treatment effects conditional on 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 that 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 earnings distribution, roughly between the 60th and 80th percentile when considering pre-random assignment welfare receipt and between the 75th and 95th percentile for single mothers who had no positive earnings in the seven quarters before the experiment. These percentiles correspond to quarterly earnings up to $2,000 to $3,800 depending on the subgroup category.

The results for subgroups defined based on prior earnings and welfare receipt are consistent with a static labor supply model. 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 to the right on the budget constraint and may increase

[35] Heckman and Smith (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.

Table 3: Testing for Treatment Effect Heterogeneity Between Subgroups

Subgroup category	Test statistic	Critical value	p-value
Education	477.32	560.93	0.1021
Marital status	672.16	595.98	0.0250
Age of youngest child	653.61	521.19	0.0140
Number of children	623.71	467.58	0.0090
Earnings in quarter 7 pre-treatment	616.49	673.51	0.0631
Welfare receipt in quarter 7 pre-treatment	617.53	539.28	0.0230
Share of quarters with positive earnings	979.38	713.04	0.0100
Share of quarters on welfare	589.69	806.19	0.1431

Notes: This table shows test results for hypothesis (H.5), i.e. these tests show for which subgroups categories we can reject treatment effects that are homogenous within subgroups for all subgroups.

their labor supply only a little. Those with high earnings may even reduce their labor supply to become eligible for Jobs First. These predictions are clearly borne out by the results in Figure 4. For example, among women with an above-median share of pre-random assignment quarters with positive earnings, the range of negative QTEs is largest, because many of them reduce their labor supply in response to the Jobs First rules. Overall, our multiple testing results have clear policy implications as they show that a substantial share of the most disadvantaged women benefit from this reform.

We now formally test for treatment effect heterogeneity between and within subgroups. Table 3 presents the results for hypothesis (H.5) for the same subgroups as above. This null hypothesis posits that there are no differences across subgroups that can explain the observed heterogeneity of QTEs in the full sample. We can reject the (H.5) for all but two sets of subgroups at a level of five percent. The p-value is largest for subgroups defined by education and the share of quarter on welfare before random assignment. Hence, for these two subgroup categories, we cannot reject the null hypothesis that the treatment effect is constant across earnings percentiles for all subgroups. Overall, however, we 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 (forthcoming), our test relaxes the strong assumption of treatment effect homogeneity within subgroups that is implicit in their test.

The tests of hypothesis (H.6) shown in Table 4 additionally account for potential dependencies within and across subgroups. These test results provide additional insight beyond

Table 4: Testing Which Subgroups Exhibit Treatment Effect Heterogeneity

Subgroup category	Test statistic	p-value
Education		
High School Degree or GED	477.32	0
No High School Degree or GED	424.74	0
Marital Status		
Never Married	411.34	0
Previously Married	672.16	0
Age of Youngest Child		
Youngest Child 6 or Older	653.61	0
Youngest Child Younger than 6	418.56	0.01
Number of Children		
Two or more Children	623.71	0
One Child	358.76	0.035
Earnings in Quarter 7 Pre-Treatment		
Positive Earnings	278.35	0
Zero Earnings	616.49	0.08
Welfare Receipt in Quarter 7 Pre-Treatment		
Welfare Receipt	617.53	0
No Welfare Receipt	274.23	0.005
Share of Quarters with Positive Earnings		
No Quarters with Positive Earnings	979.38	0
Share of Quarters with Positive Earnings Below Median	311.34	0.71
Share of Quarters with Positive Earnings Above Median	337.11	0.71
Share of Quarters on Welfare		
No Quarters with Welfare Receipt	306.19	0.875
Share of Quarters with Welfare Receipt Below Median	422.68	0.38
Share of Quarters with Welfare Receipt Above Median	589.69	0.01

Notes: This table shows test results for hypothesis (H.6), i.e. these tests show for which subgroups in each subgroup category we can reject homogenous treatment effects. p-values 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.

testing (H.5) because they identify the individual subgroups that exhibit treatment effect heterogeneity. In these results, a p-value below 0.05 indicates that the corresponding subgroup exhibits a statistically significant amount of treatment effect heterogeneity across the earning distribution. The only subgroup categories for which we do not find evidence of treatment effect heterogeneity are share of quarters with positive earnings and welfare receipt, respectively. These results confirm the findings in Figure 4. In particular, they indicate that individuals with little past welfare receipt of positive past earnings generally do not increase their labor supply, so we also do not find any heterogeneity in the QTEs for these subgroups. Overall, however, 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 the present context, the results can guide policymakers in adjusting welfare rules, for example by introducing more (or different) conditions for welfare receipt. 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. Further, our tests generalize the idea of tests considered in Bitler, Gelbach, and Hoynes (forthcoming) by not restricting treatment effects to be constant across quantiles within a subgroup when determining if the distributional heterogeneity across the full sample is characterized by subgroups.

Using data from the Jobs First experiment, we not only present evidence of considerable 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.

We would like to emphasize that our multiple testing approach is generally applicable in various other ways beyond what this paper demonstrated. First, the tests can be applied to situations with multiple treatments (e.g., List, Shaikh, and Xu, 2016) or situations where there is selection on unobservables that explore if there is heterogeneity in marginal treatment effects (e.g., Heckman and Vytlacil, 2005; Brinch, Mogstad, and Wiswall, forthcoming). Second, instead of using inverse propensity score weighting, we may directly use the conditional distribution functions or conditional quantile functions to identify the treatment effects as proposed by Chernozhukov, Fernandez-Val, and Melly (2013). Certainly we can extend their

proposal to multiple testing procedures for testing for treatment effect heterogeneity across thresholds in the distribution function or quantiles in the quantile function with or without subgroups.

References

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

Armstrong, Timothy B. and Shu Shen. 2015. "Inference on Optimal Treatment Assignments."

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.

———. forthcoming. "Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment." *Review of Economics and Statistics* .

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."

Bonferroni, Carlo Emilio. 1936. *Teoria Statistica Delle Classi E Calcolo Delle Probabilit.* Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.

Brinch, Christian N., Magne Mogstad, and Matthew Wiswall. forthcoming. "Beyond LATE with a Discrete Instrument." *Journal of Political Economy* .

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.

Dehejia, Rajeev H. 2005. "Program Evaluation as a Decision Problem." *Journal of Econometrics* 125 (1-2):141–173.

Djebbari, Habiba and Jeffrey Smith. 2008. "Heterogeneous Impacts in PROGRESA." *Journal of Econometrics* 145 (1-2):64–80.

Fink, Gunther, 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.

Gelbach, Jonah B. 2005. "Inference for Sample Quantiles with Discrete Data."

Gu, Jiaying and Shu Shen. 2016. "Oracle and Adaptive False Discovery Rate Controlling Method for One-Sided Testing: Theory and Application in Treatment Effect Evaluation."

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

Heckman, James J. 2001. "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture." *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." *Review of Economic Studies* 64 (4):487–535.

Heckman, James J. and Jeffrey A. Smith. 1998. "Evaluating the Welfare State." NBER Working Paper 6542.

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.

Heckman, James J. and Edward Vytlacil. 2005. "Structural Equations, Treatment Effects, and Econometric Policy evaluation1." *Econometrica* 73 (3):669–738.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2):65–70.

Imbens, Guido W. 2009. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." NBER Working Paper 14896.

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

Kline, Patrick and Melissa Tartari. 2016. "Bounding the Labor Supply Responses to a Randomized Welfare Experiment: A Revealed Preference Approach." *American Economic Review* 106 (4):972–1014.

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.

List, John A., Azeem M. Shaikh, and Yang Xu. 2016. "Multiple Hypothesis Testing in Experimental Economics." NBER Working Paper 21875.

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

Manski, Charles F. 2004. "Statistical Treatment Rules for Heterogeneous Populations." *Econometrica* 72 (4):1221–1246.

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

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. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50 (2):301–316.

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