Online Appendix for "Multiple Testing and the Distributional Effects of Accountability Incentives in Education"

Steven F. Lehrer[*] R. Vincent Pohl[†] Kyungchul Song[‡]

Queen's University and NBER

Mathematica

University of British Columbia

May 2020

Abstract

This is the online appendix for Lehrer, Pohl, and Song (2020). Four sections are included, which provide further details regarding (1) additional motivation for the testing procedure, (2) the asymptotic validity of the multiple testing procedure, (3) empirical results under block bootstrapping, and (4) an additional application of multiple testing to distributional treatment effects.

[*] Economics Department, email: lehrers@queensu.ca.

[†] Email: vincent.pohl@gmail.com.

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

A Additional Motivation for the Testing Procedure From an Empirical Perspective

The well-documented diverse and heterogeneous behavior in how individuals respond to a particular treatment or intervention 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. 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). 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. While an increasing number of studies account for possible treatment effect heterogeneity when evaluating programs or other interventions, most conduct statistical inference without allowing for dependence across subgroups.

As Fink, McConnell, and Vollmer (2014) point out, a majority of studies based on field experiments published in 10 specific journals estimate separate average causal parameters for different subgroups, but report traditional standard errors and p-values when testing for heterogeneous treatment effects through interaction terms or subgroup analyses. This is inappropriate because each interaction term represents a separate hypothesis beyond the original experimental design and results in a substantially increased type I error. 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, since the probability of a false positive equals 0.05 for each individual hypothesis, the probability of falsely rejecting at least one true null hypothesis may be much larger. Hence, the type I error exceeds the nominal size of the test.

A similar observation can be made for distributional treatment effects. A growing number of studies examine if treatment effects differ 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). Testing for the presence of positive (or, generally, non-zero) QTEs involves a test of multiple hypotheses, for example 99 hypotheses in the case of percentile treatment effects. Therefore, the 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). As a result, the type I error rates can exceed the desired level of the test, which leads researchers to reject "too many" individual hypotheses.[1] To the best of our knowledge, no published study estimating distributional treatment effects makes such a correction. Among articles published in five high-impact economic journals between 2008 to 2017 that estimate distributional treatment effects none corrects inference for multiple testing (Allen, Clark, and Houde, 2014; Angrist, Lang, and Oreopoulos, 2009; Bandiera et al., 2017; Banerjee et al., 2015; Behaghel, de Chaisemartin, and Gurgand, 2017; Brown et al., 2014; Crepon et al., 2015; Evans and Garthwaite, 2012; Fack and Landais, 2010; Fairlie and Robinson, 2013; McKenzie, 2017; Meyer and Sullivan, 2008; Muralidharan, Niehaus, and Sukhtankar, 2016). The absence of these corrections may reflect that econometric testing procedures for QTEs were not previously developed. Lehrer, Pohl, and Song (2020) aims to fill that gap and also provide a formal result of asymptotic validity when the propensity scores are parametrically specified.

B Mathematical Proofs

Our first result is the asymptotic linear representation of $\sqrt{n}(\hat{q}_d(\tau) - q_d(\tau))$ and its bootstrap version that is uniform over $\tau \in [\tau_L, \tau_U]$. Let us introduce some notation. Let

$$a_\tau(Y_i; q) = \tau - 1\{Y_i \leqslant q\}, \tag{1}$$

and for $u \in \mathbb{R}$,

$$\Delta(Y_i; q) = \sqrt{n} \int_0^1 \left(a_\tau(Y_i; q + n^{-1/2}us) - a_\tau(Y_i; q) \right) ds. \tag{2}$$

(Note that the right hand side in (2) does not depend on τ.)

Let $J_d(\tau_U, \tau_L) = \{q_d(\tau) : \tau \in [\tau_L, \tau_U]\}$, and assume that it is bounded in $\mathbb{R}$. (See Assumption 2.3(ii) of the main text.) For $q \in J_d(\tau_U, \tau_L)$ and $u \in \mathbb{R}$, let

$$\varphi_n(Y_i; q) = \Delta(Y_i; q)/n^{3/4} = -n^{-1/4} \int_0^1 1\{q < Y_i \leqslant q + n^{-1/2}us\} ds. \tag{3}$$

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

Let $\mathcal{B}$ be a class of bounded measurable functions $b : \mathbb{R} \times \mathcal{X} \times \{0,1\} \to \mathbb{R}$. Define

$$\mathcal{H}_n = \{\varphi_n(\cdot; q)b(\cdot) : (q,b) \in J_d(\tau_U, \tau_L) \times \mathcal{B}\}. \tag{4}$$

We let $V_i = (Y_i, X_i', D_i) \in \mathbb{R}^{d_V}$ for brevity of notation.

We introduce a pseudo-norm $\|\cdot\|_{P,2}$ on the set of measurable functions on $\mathbb{R} \times \mathcal{X} \times \{0,1\}$: $\|f\|_{P,2} = (E|f(V_i)|^2)^{1/2}$, for any measurable map f. We also denote the sup norm by $\|\cdot\|_\infty$: $\|f\|_\infty = \sup_{v \in \mathbb{R}^{d_V}} |f(v)|$. For each $\varepsilon > 0$, let $N_{[]}(\varepsilon, \mathcal{H}_n, \|\cdot\|_{P,2})$ denote the ε-bracketing number of $\mathcal{H}_n$ with respect to $\|\cdot\|_{P,2}$ (see van der Vaart and Wellner, 1996, p. 83).

Lemma B.1 *There exist constants $C_1, C_2, C_3, C_4 > 0$ such that for each $\varepsilon \in (0,1)$, there exists a set of brackets $[h_{L,j}, h_{U,j}]$ with $1 \leqslant j \leqslant N(\varepsilon)$, such that the brackets cover $\mathcal{H}_n$ and for each $k \geqslant 2$,*

$$E[|h_{L,j}(V_i) - h_{U,j}(V_i)|^k] \leqslant C_1(C_2 n^{-1/4})^{k-2}\varepsilon^2, \tag{5}$$

and

$$\log N(\varepsilon) \leqslant C_3 - C_3 \log(\varepsilon) + C_4 \log N_{[]}(C\varepsilon^2, \mathcal{B}, \|\cdot\|_2). \tag{6}$$

Proof: First, define for $\delta > 0$,

$$\begin{aligned} z_\delta(y; q, s) \quad = \quad & (1 - \min\{(y - q - n^{-1/2}us)/\delta, 1\})1\{0 < y - q - n^{-1/2}us\} \\ & +1\{y - q - n^{-1/2}us \leqslant 0\}. \end{aligned} \tag{7}$$

Define

$$\varphi_{U,\delta}(y; q) = \int_0^1 \varphi_{U,\delta}(y; q, s)ds, \text{ and } \varphi_{L,\delta}(y; q) = \int_0^1 \varphi_{L,\delta}(y; q, s)ds, \tag{8}$$

where

$$\varphi_{U,\delta}(y; q, s) \quad = \quad z_\delta(y; q, s) - z_\delta(y + \delta + n^{-1/2}us; q, s), \text{ and} \tag{9}$$

$$\varphi_{L,\delta}(y; q, s) \quad = \quad \min\left\{z_\delta(y + \delta; q, s), z_\delta(-y + 2q + \delta + n^{-1/2}us; q, s)\right\}. \tag{10}$$

Let $\varphi(y; q, s) = 1\{y - q - n^{-1/2}us \leqslant 0\} - 1\{y - q \leqslant 0\}$ and define

$$\varphi_n(y; q) = n^{-1/4} \int_0^1 \varphi(y; q, s)ds. \tag{11}$$

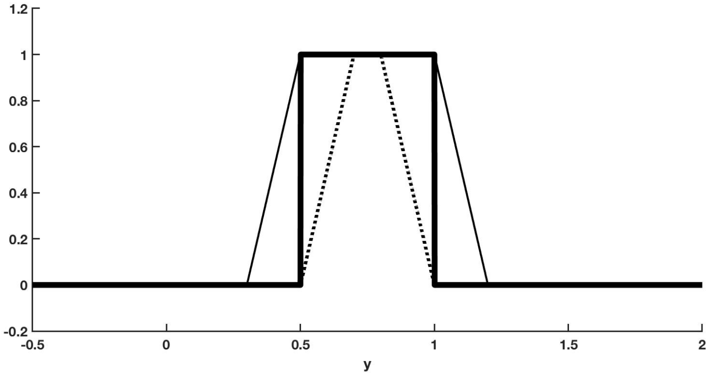


Figure 1: Illustration of $\varphi_{U,\delta}, \varphi_{L,\delta}$: The solid thick line depicts $\varphi(\cdot; q, s)$, the solid thin line $\varphi_{U,\delta}(\cdot; q, s)$ and the dotted line $\varphi_{L,\delta}(\cdot; q, s)$. Here we take $n^{-1/2}us = 0.5, q = 0.5$ and $\delta = 0.2$. The absolute slope of both maps $\varphi_{U,\delta}(\cdot; q, s)$ and $\varphi_{L,\delta}(\cdot; q, s)$ are bounded by $1/\delta$.

(Note that the definition (3) conforms with this.)

Then, we have for all $y \in \mathbb{R}$, (see Figure 1)

$$\varphi_{L,\delta}(y; q) \leqslant \varphi(y; q) \leqslant \varphi_{U,\delta}(y; q). \tag{12}$$

It is not hard to see that for all $q, q' \in \mathbb{R}$, and all $y \in \mathbb{R}$,

$$\begin{aligned} |\varphi_{U,\delta}(y; q) - \varphi_{U,\delta}(y; q')| &\leqslant |q - q'|/\delta, \text{ and} \\ |\varphi_{L,\delta}(y; q) - \varphi_{L,\delta}(y; q')| &\leqslant |q - q'|/\delta. \end{aligned} \tag{13}$$

Furthermore, for some constant $C > 0$,

$$E\left[(\varphi_{U,\delta}(Y_i; q) - \varphi_{L,\delta}(Y_i; q))^2 \,|D_i = d\right] \leqslant C\delta, \tag{14}$$

and

$$E\left[\varphi^2_{U,\delta}(Y_i; q)|D_i = d\right] \leqslant 1, \text{ and } E\left[\varphi^2_{L,\delta}(Y_i; q)|D_i = d\right] \leqslant 1. \tag{15}$$

Define

$$\mathcal{H}_{L,\delta} = \{\varphi_{L,\delta}(\cdot; q)b(\cdot)/n^{1/4} : (q, b) \in J_d(\tau_U, \tau_L) \times \mathcal{B}\}, \text{ and} \tag{16}$$

$$\mathcal{H}_{U,\delta} = \{\varphi_{U,\delta}(\cdot; q)b(\cdot)/n^{1/4} : (q, b) \in J_d(\tau_U, \tau_L) \times \mathcal{B}\}. \tag{17}$$

From (13) and using the fact that $n \geqslant 1$, $J_d(\tau_U, \tau_L)$ is bounded, and $\varphi_{L,\delta}(\cdot; q)$, $\varphi_{U,\delta}(\cdot; q)$ and $b(\cdot)$ are bounded maps, we find that

$$\begin{aligned} N_{[]}(\varepsilon, \mathcal{H}_{L,\delta}, \|\cdot\|_{P,2}) &\leqslant C(\varepsilon\delta)^{-1} \times N_{[]}(C\varepsilon, \mathcal{B}, \|\cdot\|_{P,2}), \text{ and} \\ N_{[]}(\varepsilon, \mathcal{H}_{U,\delta}, \|\cdot\|_{P,2}) &\leqslant C(\varepsilon\delta)^{-1} \times N_{[]}(C\varepsilon, \mathcal{B}, \|\cdot\|_{P,2}), \end{aligned} \tag{18}$$

for all $\varepsilon > 0$, for some constant $C > 0$. We take $\delta = \varepsilon^2$ and ε^2-brackets $[h_{L,a,j}, h_{L,b,j}]_{j=1}^N$ and $[h_{U,a,j}, h_{U,b,j}]_{j=1}^N$ such that the former set of brackets cover $\mathcal{H}_{L,\varepsilon^2}$ and the latter $\mathcal{H}_{U,\varepsilon^2}$, both with respect to $\|\cdot\|_{P,2}$, and

$$\|h_{U,a,j}\|_\infty + \|h_{U,b,j}\|_\infty + \|h_{L,a,j}\|_\infty + \|h_{L,b,j}\|_\infty \leqslant \frac{C}{n^{1/4}}, \tag{19}$$

for all $j = 1, ..., N$, for some constant $C > 0$. By (12) and (18), we lose no generality by taking brackets so that for each $h \in \mathcal{H}_n$, there exists $j \in \{1, ..., N\}$ such that[2]

$$\min\{h_{U,b,j}, h_{L,a,j}\} \leqslant h \leqslant \max\{h_{L,a,j}, h_{U,b,j}\}, \tag{20}$$

and

$$\log N \leqslant C - C\log\varepsilon + C\log N_{[]}(C\varepsilon^2, \mathcal{B}, \|\cdot\|_{P,2}), \tag{21}$$

for some $C > 0$. We set

$$h_{L,j} = \min\{h_{U,b,j}, h_{L,a,j}\}, \text{ and } h_{U,j} = \max\{h_{U,b,j}, h_{L,a,j}\}. \tag{22}$$

Therfore, by (19) and (14), for each $k \geqslant 2$,

$$\begin{aligned} E[|h_{L,j}(V_i) - h_{U,j}(V_i)|^k] &\leqslant C(Cn^{-1/4})^{k-2} E\left[(h_{L,a,j}(V_i) - h_{U,b,j}(V_i))^2\right] \\ &\leqslant 2C(Cn^{-1/4})^{k-2} E\left[(h_{L,a,j}(V_i) - h_{L,b,j}(V_i))^2\right] \\ &\quad +4C(Cn^{-1/4})^{k-2} E\left[(h_{L,b,j}(V_i) - h_{U,a,j}(V_i))^2\right] \\ &\quad +4C(Cn^{-1/4})^{k-2} E\left[(h_{U,a,j}(V_i) - h_{U,b,j}(V_i))^2\right] \\ &\leqslant C_1(C_2 n^{-1/4})^{k-2}\left(\varepsilon^4 + \varepsilon^2 + \varepsilon^4\right), \end{aligned} \tag{23}$$

for some constants $C, C_1, C_2 > 0$. The terms ε^4 are due to the choice of ε^2-brackets and the term ε^2 comes from (14) and $\delta = \varepsilon^2$. ■

[2] Since b can take negative values, the inequality (12) does not necessarily imply that $h_{L,a,j} \leqslant h \leqslant h_{U,b,j}$.

Define for each $\tau \in [\tau_L, \tau_U]$ and $b \in \mathcal{B}$,

$$U(\tau, b; \delta) = \{(\tau_1, b_1) \in [\tau_L, \tau_U] \times \mathcal{B} : |\tau - \tau_1| + \|b - b_1\|_{P,2} \leqslant \delta\}. \tag{24}$$

Recall the definitions of $a_\tau(Y_i; q)$ and $\Delta(Y_i; q)$ in (1).

Lemma B.2 *Suppose that $\mathcal{B}$ and $\mathcal{H}_n$ are as in Lemma B.1, and that for $d = 0, 1$, the density f_d of Y_{di} is bounded. Furthermore, assume that there exists $C > 0$ such that for all $\varepsilon > 0$,*

$$\log N_{[]}(\varepsilon, \mathcal{B}, \|\cdot\|_{P,2}) \leqslant C - C \log \varepsilon. \tag{25}$$

Then the following statements hold.

(i) There exist $s > 0$ and $C > 0$ such that for all $n \geqslant 1$, and for all $\delta > 0$,

$$E\left[\sup_{(\tau_1,b_1)\in U(\tau,b;\delta)} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(a_{\tau_1}(Y_i; q_d(\tau_1))b_1(V_i) - E\left[a_{\tau_1}(Y_i; q_d(\tau_1))b_1(V_i)\right]\right)\right|\right] \leqslant C\delta^s.$$

(ii) There exists $C > 0$ such that for all $n \geqslant 1$,

$$E\left[\sup_{h\in\mathcal{H}_n}\left|\sum_{i=1}^{n}\left(h(V_i) - E\left[h(V_i)\right]\right)\right|\right] \leqslant C \log n.$$

Proof: (i) Let

$$\mathcal{B}_1 = \{a_\tau(\cdot; q) : (\tau, q) \in [\tau_L, \tau_U] \times \mathbb{R}\}. \tag{26}$$

Certainly, $\mathcal{B}_1$ is a VC class. Both classes $\mathcal{B}_1$ and $\mathcal{B}$ are classes of bounded functions. Hence, if we let $\mathcal{B}_2$ be the collection of functions $f(\cdot)g(\cdot)$ as we run $f \in \mathcal{B}_1$ and $g \in \mathcal{B}$, we have for some constant $C > 0$,

$$\log N_{[]}(\varepsilon, \mathcal{B}_2, \|\cdot\|_{P,2}) \leqslant C - C \log \varepsilon + \log N_{[]}(C\varepsilon, \mathcal{B}, \|\cdot\|_{P,2}). \tag{27}$$

Using (25), we obtain the finite integral bracketing entropy bound for the left hand side of (27). The desired result of (i) follows by the maximal inequality. (For example, see (1) from van der Vaart (1996).)

(ii) When $u = 0$, we have $\Delta(Y_i; q_d(\tau)) = 0$, a.s. We focus on the case $u \neq 0$. Observe that since b is bounded and $|\varphi_n(Y_i; q)| \leqslant 1$, for some constant $C > 0$,

$$E\left[|\varphi_n(Y_i; q) b(V_i)|^k\right] \leqslant CE\left[|\varphi_n(Y_i; q)|^k\right] \leqslant CE\left[\varphi_n^2(Y_i; q)\right]. \tag{28}$$

Observe that

$$\begin{aligned} E\left[\varphi_n^2(Y_i; q)\right] &\leqslant \frac{1}{\sqrt{n}} \int_0^1 P\{q \leqslant Y_i \leqslant q + n^{-1/2} us\} ds & (29) \\ &= \frac{1}{\sqrt{n}} \int_0^1 P\{q \leqslant Y_{1i} \leqslant q + n^{-1/2} us | D_i = 1\} P\{D_i = 1\} ds \\ &\quad + \frac{1}{\sqrt{n}} \int_0^1 P\{q \leqslant Y_{0i} \leqslant q + n^{-1/2} us | D_i = 0\} P\{D_i = 0\} ds \\ &\leqslant \frac{1}{\sqrt{n}} \int_0^1 P\{q \leqslant Y_{1i} \leqslant q + n^{-1/2} us\} ds & (30) \\ &\quad + \frac{1}{\sqrt{n}} \int_0^1 P\{q \leqslant Y_{0i} \leqslant q + n^{-1/2} us\} ds \\ &\leqslant \frac{1}{n} \max_{d=0,1} \sup_{q \in \mathbb{R}} f_d(q) n^{-1}, \end{aligned}$$

Using this in combination with Lemma B.1, we apply Theorem 6.8 of Massart (2007) (taking $b = 1$ and $\sigma = Cn^{-1/2}$ there) to obtain that

$$\begin{aligned} E\left[\sup_{h \in \mathcal{H}_n} \left|\sum_{i=1}^n (h(V_i) - Eh(V_i))\right|\right] &\leqslant C_1 + C_1 \sqrt{n} \int_0^{C_1/\sqrt{n}} \sqrt{\log(1/u)} du + C_1 \log n \\ &\leqslant C_2 \log n, \end{aligned}$$

for some constants $C_1, C_2 > 0$ from large n on. Thus we obtain the desired result. ■

Theorem B.1 *Suppose that Assumptions 2.2 and 2.3 in the main text hold, and let*

$$\zeta_i = \psi(V_i) - E\psi(V_i), \text{ and } \zeta_i^* = \psi(V_i^*) - E[\psi(V_i^*) | \mathcal{F}_n]. \tag{31}$$

Then the following statements hold.

(i)

$$
\begin{aligned}
&\sqrt{n}(\hat{q}_d(\tau) - q_d(\tau)) && (32)\\
= \quad & -\frac{1}{\sqrt{n} f_d(q_d(\tau))} \sum_{j=1}^{n} \frac{a_\tau(Y_j; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \\
& + \frac{1}{\sqrt{n} f_d(q_d(\tau))} \sum_{j=1}^{n} E\left[\frac{a_\tau(Y_j; q_d(\tau)) g_d(X_i; \beta_0)' 1\{D_i = d\}}{p_d^2(X_i)}\right] \zeta_j + o_P(1), && (33)
\end{aligned}
$$

uniformly over $\tau \in [\tau_L, \tau_U]$.

(ii)

$$
\begin{aligned}
&\sqrt{n}(\hat{q}_d^*(\tau) - \hat{q}_d(\tau)) && (34)\\
= \quad & -\frac{1}{\sqrt{n} f_d(q_d(\tau))} \sum_{j=1}^{n} \left(\frac{a_\tau(Y_j^*; q_d(\tau)) 1\{D_j = d\}}{p_d(X_j^*)} - \frac{1}{n} \sum_{i=1}^{n} \frac{a_\tau(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \right) \\
& + \frac{1}{\sqrt{n} f_d(q_d(\tau))} \sum_{j=1}^{n} E\left[\frac{a_\tau(Y_j; q_d(\tau)) g_d(X_i; \beta_0)' 1\{D_i = d\}}{p_d^2(X_i)}\right] \zeta_j^* + o_P(1), && (35)
\end{aligned}
$$

uniformly over $\tau \in [\tau_L, \tau_U]$.

Proof: (i) Note that

$$\sqrt{n}(\hat{q}_d(\tau) - q_d(\tau)) = \arg\min_{u \in \mathbb{R}} \left(\hat{Q}_d(q_d(\tau) + n^{-1/2}u; \tau) - Q_d(q_d(\tau); \tau) \right). \quad (36)$$

Recall the definitions of $\hat{Q}_d(q;\tau)$ and $Q_d(q;\tau)$ in Section 2.2.3 of the main text. We write

$$\hat{Q}_d(q_d(\tau) + n^{-1/2}u; \tau) - Q_d(q_d(\tau); \tau) = A_n + B_n, \quad (37)$$

where

$$
\begin{aligned}
A_n &= \hat{Q}_d(q_d(\tau) + n^{-1/2}u; \tau) - Q_d(q_d(\tau) + n^{-1/2}u; \tau), \text{ and} && (38)\\
B_n &= Q_d(q_d(\tau) + n^{-1/2}u; \tau) - Q_d(q_d(\tau); \tau).
\end{aligned}
$$

We can follow the same arguments as in the proof of Theorem 3 of Kato (2009), and show that

$$B_n = -\frac{u}{\sqrt{n}} \sum_{i=1}^{n} \frac{a_\tau(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} + \frac{u^2}{2} f_d(q_d(\tau)) + o_P(1), \quad (39)$$

uniformly over $\tau \in [\tau_L, \tau_U]$. Let us focus on A_n. We write

$$A_n = A_{n,1} + \xi_{n,d}(\tau), \tag{40}$$

where

$$\begin{aligned} A_{n,1} &= \hat{Q}_d(q_d(\tau) + n^{-1/2}u; \tau) - \hat{Q}_d(q_d(\tau); \tau) & (41) \\ &\quad - \left(Q_d(q_d(\tau) + n^{-1/2}u; \tau) - Q_d(q_d(\tau); \tau)\right), & (42) \end{aligned}$$

and

$$\xi_{n,d}(\tau) = \hat{Q}_d(q_d(\tau); \tau) - Q_d(q_d(\tau); \tau). \tag{43}$$

Using Knight's identity (see (15) on page 1855 of Kato (2009)), we write $A_{n,1}$ as

$$uZ_{n,d}^{(1)}(\tau) + Z_{n,d}^{(2)}(u, \tau), \tag{44}$$

where

$$\begin{aligned} Z_{n,d}^{(1)}(\tau) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{a_\tau(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \left(\frac{p_d(X_i)}{\hat{p}_d(X_i)} - 1 \right), \text{ and} & (45) \\ Z_{n,d}^{(2)}(u, \tau) &= -\frac{u}{n} \sum_{i=1}^{n} \left(\frac{p_d(X_i)}{\hat{p}_d(X_i)} - 1 \right) \frac{\Delta(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)}. & (46) \end{aligned}$$

(The dependence of $Z_{n,d}^{(2)}(u, \tau)$ on u is due to the dependence of $\Delta(Y_i; q_d(\tau))$ on u.) We write

$$Z_{nd}^{(1)}(\tau) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{a_\tau(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \frac{p_d(X_i) - \hat{p}_d(X_i)}{p_d(X_i)} + R_n(\tau), \tag{47}$$

where

$$R_n(\tau) = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{a_\tau(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \frac{(p_d(X_i) - \hat{p}_d(X_i))^2}{p_d(X_i) \hat{p}_d(X_i)}. \tag{48}$$

By expanding $G(x; \hat{\beta})$ around β_0 and using Assumption 2.3 (i), it is not hard to see that

$$\hat{p}_d(x) - p_d(x) = g_d(x; \beta_0)'(\hat{\beta} - \beta_0) + O_P(n^{-1}), \tag{49}$$

uniformly over $x \in \mathcal{X}$. Since $|a_\tau(Y_i; q_d(\tau))/p_d(X_i)| \leqslant \varepsilon^{-1}$ for all $\tau \in [\tau_L, \tau_U]$, we find that

$$\sup_{\tau \in [\tau_L, \tau_U]} |R_n(\tau)| = O_P(n^{-1/2}). \tag{50}$$

Applying this and the expansion in (49) to the leading term on the right hand side of (47), we obtain that

$$\begin{aligned} Z_{nd}^{(1)}(\tau) &= -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{a_\tau(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \frac{g_d(X_i; \beta_0)'(\hat{\beta} - \beta_0)}{p_d(X_i)} + o_P(1) \\ &= -\left(\frac{1}{n} \sum_{i=1}^{n} \frac{a_\tau(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \frac{g_d(X_i; \beta_0)'}{p_d(X_i)} \right) \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_i + o_P(1) \right) + o_P(1) \\ &= -E\left[\frac{a_\tau(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \frac{g_d(X_i; \beta_0)'}{p_d(X_i)} \right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_i + o_P(1), \end{aligned}$$

by Lemma B.2(i). Using the same arguments, we also obtain that

$$Z_{nd}^{(2)}(\tau) = -u \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\Delta(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \frac{g_d(X_i; \beta_0)'}{p_d(X_i)} \right) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_i + o_P(1) = o_P(1).$$

We define $b_0(V_i) = 1\{D_i = d\} g_{d,k}(X_i; \beta_0)/p_d^2(X_i)$, where $g_{d,k}(X_i; \beta_0)$ is the k-th entry of $g_d(X_i; \beta_0)$, and take $\mathcal{B} = \{b_0\}$, i.e., the singleton of b_0 in the definition of $\mathcal{H}_n$ in (4). We bound

$$\begin{aligned} &\sup_{\tau \in [\tau_L, \tau_U]} \left| \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \frac{g_{d,k}(X_i; \beta_0)}{p_d(X_i)} - E\left[\frac{\Delta(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \frac{g_{d,k}(X_i; \beta_0)}{p_d(X_i)} \right] \right| \\ \leqslant\ & n^{-1/4} \sup_{h \in \mathcal{H}_n} \left| \sum_{i=1}^{n} (h(V_i) - Eh(V_i)) \right|. \end{aligned}$$

By Lemma B.2(ii), we find that uniformly over $\tau \in [\tau_L, \tau_U]$,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \frac{g_{d,k}(X_i; \beta_0)}{p_d(X_i)} &= E\left[\frac{\Delta(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \frac{g_{d,k}(X_i; \beta_0)}{p_d(X_i)} \right] \\ &\quad + O(n^{-1/4} \log n). \end{aligned}$$

Therefore, we conclude that

$$A_n = -E\left[\frac{a_\tau(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \frac{g_d(X_i; \beta_0)'}{p_d(X_i)} \right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_i + \xi_{n,d}(\tau) + o_P(1). \tag{51}$$

Combining this with (39), and applying Theorem 2 of Kato (2009), we obtain the desired result of (i).

(ii) The proof of the bootstrap version is similar to that of (i). First, we write

$$\sqrt{n}(\hat{q}_d^*(\tau) - \hat{q}_d(\tau)) = \arg\min_{u \in \mathbb{R}} \hat{Q}_d^*(\hat{q}_d(\tau) + n^{-1/2}u; \tau) - Q_d^*(\hat{q}_d(\tau); \tau), \tag{52}$$

where

$$\begin{aligned} \hat{Q}_d^*(q; \tau) &= \sum_{i=1}^n \frac{1\{D_i^* = d\}}{\hat{p}_d^*(X_i^*)} \rho_\tau(Y_i^* - q), \text{ and} \quad (53) \\ Q_d^*(q; \tau) &= \sum_{i=1}^n \frac{1\{D_i^* = d\}}{\hat{p}_d(X_i^*)} \rho_\tau(Y_i^* - q). \quad (54) \end{aligned}$$

Similarly as before, we write

$$\hat{Q}_d^*(\hat{q}_d(\tau) + n^{-1/2}u; \tau) - Q_d^*(\hat{q}_d(\tau); \tau) = A_n^* + B_n^*, \tag{55}$$

where

$$\begin{aligned} A_n^* &\equiv \hat{Q}_d^*(\hat{q}_d(\tau) + n^{-1/2}u; \tau) - Q_d^*(\hat{q}_d(\tau) + n^{-1/2}u; \tau), \text{ and} \quad (56) \\ B_n^* &\equiv Q_d^*(\hat{q}_d(\tau) + n^{-1/2}u; \tau) - Q_d^*(\hat{q}_d(\tau); \tau). \end{aligned}$$

We write

$$A_n^* = A_{n,1}^* + \xi_{n,d}^*(\tau), \tag{57}$$

where

$$\xi_{n,d}^*(\tau) = \hat{Q}_d^*(\hat{q}_d(\tau); \tau) - Q_d^*(\hat{q}_d(\tau); \tau), \tag{58}$$

and

$$\begin{aligned} A_{n,1}^* &= \hat{Q}_d^*(\hat{q}_d(\tau) + n^{-1/2}u; \tau) - \hat{Q}_d^*(\hat{q}_d(\tau); \tau) \quad (59) \\ &\quad -(Q_d^*(\hat{q}_d(\tau) + n^{-1/2}u; \tau) - Q_d^*(\hat{q}_d(\tau); \tau)). \end{aligned}$$

Following the similar arguments as before, we obtain that

$$A_{n,1}^* = -uE\left[\frac{a_\tau(Y_i^*;\hat{q}_d(\tau))1\{D_i^*=d\}}{\hat{p}_d(X_i^*)}\frac{g_d(X_i^*;\hat{\beta})'}{\hat{p}_d(X_i^*)}|\mathcal{F}_n\right]\frac{1}{\sqrt{n}}\sum_{i=1}^n \zeta_i^* + o_P(1). \tag{60}$$

Note that from (i) of this theorem and Assumption 2.2(i),

$$\sup_{\tau\in[\tau_L,\tau_U]}|\hat{q}_d(\tau)-q_d(\tau)| = o_P(1), \text{ and } \hat{\beta}=\beta_0+o_P(1). \tag{61}$$

Hence using Lemma B.2(i), we obtain that

$$\begin{aligned} & E\left[\frac{a_\tau(Y_i^*;\hat{q}_d(\tau))1\{D_i^*=d\}}{\hat{p}_d(X_i^*)}\frac{g_d(X_i^*;\hat{\beta})'}{\hat{p}_d(X_i^*)}|\mathcal{F}_n\right] \\ = & \frac{1}{n}\sum_{i=1}^n \frac{a_\tau(Y_i;\hat{q}_d(\tau))1\{D_i=d\}}{\hat{p}_d(X_i)}\frac{g_d(X_i;\hat{\beta})'}{\hat{p}_d(X_i)} \\ = & \frac{1}{n}\sum_{i=1}^n \frac{a_\tau(Y_i;q_d(\tau))1\{D_i=d\}}{p_d(X_i)}\frac{g_d(X_i;\beta_0)'}{p_d(X_i)} + o_P(1) \\ = & E\left[\frac{a_\tau(Y_i;q_d(\tau))1\{D_i=d\}}{p_d(X_i)}\frac{g_d(X_i;\beta_0)'}{p_d(X_i)}\right] + o_P(1). \end{aligned} \tag{62}$$

Let us turn to B_n^* defined in (56). Using Knight's identity, we write B_n^* as

$$uZ_{n,d}^{*(1)}(\tau) + Z_{n,d}^{*(2)}(u,\tau), \tag{63}$$

where

$$Z_{n,d}^{*(1)}(\tau) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{a_\tau(Y_i^*;\hat{q}_d(\tau))1\{D_i^*=d\}}{\hat{p}_d(X_i^*)}, \text{ and} \tag{64}$$

$$Z_{n,d}^{*(2)}(u,\tau) = -\frac{u}{n}\sum_{i=1}^n \frac{\hat{\Delta}(Y_i^*;\hat{q}_d(\tau))1\{D_i^*=d\}}{\hat{p}_d(X_i^*)}, \tag{65}$$

with

$$\hat{\Delta}(Y_i^*;\hat{q}_d(\tau)) = \sqrt{n}\int_0^1 \left(a_\tau(Y_i^*;\hat{q}_d(\tau)+n^{-1/2}us) - a_\tau(Y_i^*;\hat{q}_d(\tau))\right)ds. \tag{66}$$

We write

$$Z_{n,d}^{*(1)}(\tau) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{a_\tau(Y_i^*; q_d(\tau))1\{D_i^* = d\}}{G_d(X_i^*; \beta_0)} - \frac{1}{n}\sum_{i=1}^{n}\frac{a_\tau(Y_i; q_d(\tau))1\{D_i = d\}}{G_d(X_i; \beta_0)}\right) + R_{n,1}(\tau),$$

where

$$R_{n,1}(\tau) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\eta_\tau(V_i^*; \hat{q}_d(\tau), \hat{\beta}) - \frac{1}{n}\sum_{i=1}^{n}\eta_\tau(V_i; \hat{q}_d(\tau), \hat{\beta})\right), \tag{67}$$

with

$$\eta_\tau(V_i^*; \hat{q}_d(\tau), \hat{\beta}) = \frac{a_\tau(Y_i^*; \hat{q}_d(\tau))1\{D_i^* = d\}}{G_d(X_i^*; \hat{\beta})} - \frac{a_\tau(Y_i^*; q_d(\tau))1\{D_i^* = d\}}{G_d(X_i^*; \beta_0)}. \tag{68}$$

Using Lemma B.2(i) again, we can show that $R_{n,1}(\tau) = o_P(1)$ uniformly over $\tau \in [-\tau_L, \tau_U]$.

We turn to $Z_{n,d}^{*(2)}(u, \tau)$ which we write as

$$-\frac{u}{n}\sum_{i=1}^{n}\left(\tilde{\eta}(V_i^*; \hat{q}_d(\tau), \hat{\beta}) - \frac{1}{n}\sum_{i=1}^{n}\tilde{\eta}(V_i; \hat{q}_d(\tau), \hat{\beta})\right) \tag{69}$$

$$-\frac{u}{n}\sum_{i=1}^{n}\left(\tilde{\eta}(V_i; \hat{q}_d(\tau), \hat{\beta}) - \int \tilde{\eta}(v; \hat{q}_d(\tau), \hat{\beta})dF_V(v)\right)$$

$$-u\left(\int \tilde{\eta}(v; \hat{q}_d(\tau), \hat{\beta})dF_V(v) - \int \tilde{\eta}(v; q_d(\tau), \beta_0)dF_V(v)\right)$$

$$-uE\left[\frac{\Delta(Y_i; q_d(\tau))1\{D_i = d\}}{p_d(X_i)}\right], \tag{70}$$

where F_V is the CDF of V_i, and

$$\tilde{\eta}(V_i^*; \hat{q}_d(\tau), \hat{\beta}) = \frac{\Delta(Y_i^*; \hat{q}_d(\tau))1\{D_i^* = d\}}{\hat{p}_d(X_i^*)}. \tag{71}$$

We show that the first two terms in (69) are $o_P(1)$ uniformly over $\tau \in [\tau_L, \tau_U]$. We will deal with the first term in (69). By Assumptions 2.1(ii) and 2.3(i) in the main text, we can find $\varepsilon > 0$ such that $G_d(x; \beta) > 0$ for all $x \in \mathcal{X}$ and all $\beta \in B(\beta_0; \varepsilon)$, where $B(\beta_0; \varepsilon) = \{\beta \in \Theta : \|\beta - \beta_0\| \leqslant \varepsilon\}$. Define

$$b_\beta(V_i) = \frac{1\{D_i = d\}}{G_d(X_i; \beta)}, \tag{72}$$

and let

$$\mathcal{B} = \{b_\beta : \beta \in B(\beta_0; \varepsilon)\}, \tag{73}$$

and define $\mathcal{H}_n$ in (4) using this $\mathcal{B}$. Since the set $B(\beta_0; \varepsilon)$ is bounded in $\mathbb{R}^{d_\beta}$, by Assumptions 2.1(ii) and 2.3(i) in the main text, we find that the bracketing condition in (25) is satisfied for this set $\mathcal{B}$. Furthermore, by Assumption 2.2(i) in the main text, we have $\hat{\beta} \in B(\beta_0; \varepsilon)$ with probability approaching one. Now, observe that

$$\sup_{\tau \in [\tau_L, \tau_U]} \left| \frac{1}{n} \sum_{i=1}^n \left(\frac{\Delta(Y_i^*; \hat{q}_d(\tau)) 1\{D_i^* = d\}}{\hat{p}_d(X_i^*)} - E\left[\frac{\Delta(Y_i^*; \hat{q}_d(\tau)) 1\{D_i^* = d\}}{\hat{p}_d(X_i^*)} | \mathcal{F}_n \right] \right) \right| \tag{74}$$

$$\leqslant \; n^{-1/4} \sup_{h \in \mathcal{H}_n} \left| \sum_{i=1}^n (h(V_i^*) - E[h(V_i^*) | \mathcal{F}_n]) \right| = O_P(n^{-1/4} \log n),$$

by Lemma B.2(ii). Thus the first term in (69) is $o_P(1)$ uniformly over $\tau \in [\tau_L, \tau_U]$. The second term can be dealt with in the same way.

Let us turn to the third term in (69). This term is also $o_P(1)$ because $\hat{q}_d(\tau) = q_d(\tau) + o_P(1)$ and $\hat{\beta} = \beta_0 + o_P(1)$. Following precisely the same argument in the proof of Theorem 3 in Kato (2009) used to deal with B_n in (38), we can show that

$$-uE\left[\frac{\Delta(Y_i; q_d(\tau)) 1\{D_i = d\}}{p_d(X_i)} \right] = \frac{u^2}{2} f_d(q_d(\tau)) + o(1), \tag{75}$$

uniformly over $\tau \in [\tau_L, \tau_U]$. Hence we conclude that

$$\hat{Q}_d^*(\hat{q}_d(\tau) + n^{-1/2}u; \tau) - Q_d^*(\hat{q}_d(\tau); \tau) \tag{76}$$

$$= -u \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\frac{a_\tau(Y_i^*; q_d(\tau)) 1\{D_i^* = d\}}{G_d(X_i^*; \beta_0)} - \frac{1}{n} \sum_{i=1}^n \frac{a_\tau(Y_i; q_d(\tau)) 1\{D_i = d\}}{G_d(X_i; \beta_0)} \right) \tag{77}$$

$$+ \frac{u^2}{2} f_d(q_d(\tau)) + o_P(1). \tag{78}$$

Now the desired result follows from Theorem 2 of Kato (2009). ■

Let us define

$$q^\Delta(\tau) = q_1(\tau) - q_0(\tau), \text{ and } \hat{q}^\Delta(\tau) = \hat{q}_1(\tau) - \hat{q}_0(\tau). \tag{79}$$

Similarly, we define a bootstrap version $\hat{q}_1^*(\tau) - \hat{q}_0^*(\tau)$. The following theorem gives the weak convergence of the process $\{\sqrt{n}(\hat{q}^\Delta(\tau) - q^\Delta(\tau)) : \tau \in [\tau_L, \tau_U]\}$. Let $\ell^\infty([\tau_L, \tau_U])$ be the collection of bounded and measurable functions on $[\tau_L, \tau_U]$. Let BL_1 be the bounded Lipschitz functionals on $\ell^\infty([\tau_L, \tau_U])$ with Lipschitz constant 1, i.e.,

$$\mathrm{BL}_1 = \{h \in \ell^\infty([\tau_L, \tau_U]) : |h(\tau_1) - h(\tau_2)| \leqslant |\tau_1 - \tau_2|, \tau_1, \tau_2 \in [\tau_L, \tau_U]\}\,. \tag{80}$$

For a sequence of stochastic processes $\mathbb{G}_n$ and a process $\mathbb{G}$ on $[\tau_L, \tau_U]$, we write

$$\mathbb{G}_n \rightsquigarrow \mathbb{G}, \text{ in } \ell^\infty[\tau_L, \tau_U]), \tag{81}$$

as $n \to \infty$, if

$$\sup_{h \in \mathrm{BL}_1} |E^*\left[h(\mathbb{G}_n)\right] - E[h(\mathbb{G})]| \to 0, \tag{82}$$

as $n \to \infty$, where E^* denotes the outer expectation. Let $\mathbb{G}_n^*$ be a stochastic process on $[\tau_L, \tau_U]$ such that for each $\tau \in [\tau_L, \tau_U]$, $\mathbb{G}_n^*(\tau)$ is a measurable map of the bootstrap sample (Y_i^*, X_i^*). Then if for any $\varepsilon > 0$,

$$P^*\left\{\sup_{h \in \mathrm{BL}_1} |E\left[h(\mathbb{G}_n^*)|\mathcal{F}_n\right] - E[h(\mathbb{G})]| > \varepsilon\right\} \to 0, \tag{83}$$

as $n \to \infty$, for some we write $\mathbb{G}_n^* \rightsquigarrow_* \mathbb{G}$ in $\ell^\infty([\tau_L, \tau_U])$. Here P^* denotes the outer probability.

Theorem B.2 *Suppose that Assumptions 2.2 and 2.3 in the main text hold. Then the following statements hold.*

(i)

$$\sqrt{n}(\hat{q}^\Delta - q^\Delta) \rightsquigarrow \mathbb{G}, \; \textit{in } \ell^\infty([\tau_L, \tau_U]). \tag{84}$$

(ii)

$$\sqrt{n}(\hat{q}^{\Delta *} - q^{\Delta *}) \rightsquigarrow_* \mathbb{G}, \; \textit{in } \ell^\infty([\tau_L, \tau_U]). \tag{85}$$

Proof: Define

$$\mathcal{G} = \{\xi(\cdot; q, \tau) : (q, \tau) \in J_d([\tau_U, \tau_L]) \times [\tau_L, \tau_U]\}, \tag{86}$$

where

$$\xi(V_j; q, \tau) = -\frac{a_\tau(Y_j; q) 1\{D_i = d\}}{f_d(q) p_d(X_i)} \tag{87}$$

$$+E\left[\frac{a_\tau(Y_j; q) g_d(X_i; \beta_0)' 1\{D_i = d\}}{f_d(q) p_d^2(X_i)}\right] \zeta_j + o_P(1). \tag{88}$$

For (i), it suffices to show that $\mathcal{G}$ is P-Donsker. This also implies (ii) by Theorem 2.2 of Giné (1997, p. 104). Define

$$\nu_n(\xi) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\xi(V_i) - E\xi(V_i)). \tag{89}$$

The convergence of finite dimensional distributions of $\{\nu_n(\xi) : \xi \in \mathcal{G}\}$ follow by the usual central limit theorem. In order to show that $\mathcal{G}$ is P-Donsker, it suffices to show that $\mathcal{G}$ is totally bounded with respect to a pseudo-norm ρ and $\{\nu_n(\xi) : \xi \in \mathcal{G}\}$ is asymptotically equicontinuous with respect to ρ. We take the norm ρ to be $\|\cdot\|_{P,2}$. The total boundedness follows by the same arguments as in the proof of (i) of Lemma B.2. It remains to show asymptotic equicontinuity of the process ν_n. For this, we write

$$\sqrt{n}(\hat{q}^\Delta(\tau) - q^\Delta(\tau)) = -A_{n,1}(q_d(\tau), \tau) + A_{n,2}(q_d(\tau), \tau) + o_P(1), \tag{90}$$

where

$$A_{n,1}(q, \tau) = \frac{1}{\sqrt{n} f_d(q)} \sum_{j=1}^{n} \frac{a_\tau(Y_j; q) 1\{D_j = d\}}{p_d(X_j)}, \text{ and} \tag{91}$$

$$A_{n,2}(q, \tau) = \frac{1}{\sqrt{n} f_d(q)} \sum_{j=1}^{n} E\left[\frac{a_\tau(Y_j; q) g_d(X_i; \beta_0)' 1\{D_i = d\}}{p_d^2(X_i)}\right] \zeta_j. \tag{92}$$

Stochastic equicontinuity of $\{A_{n,1}(q, \tau) : (q, \tau) \in J_d([\tau_L, \tau_U]) \times [\tau_L, \tau_U]\}$ obviously follows from Lemma B.2(i) and the Lipshitz continuity of $1/f_d(q)$ in $q \in J_d(\tau_U, \tau_L)$. It is not hard to show similarly that $A_{n,2}$ is stochastically equicontinuous as well. ■

We are prepared to prove Theorem 2.1 in the main text.

Proof of Theorem 2.1 in the main text: Define for any $W' \subset W$,

$$\hat{c}_{1-\alpha}(W') = \inf\left\{c \in \mathbb{R} : \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\left\{T_b^*(W') \leqslant c\right\} \geqslant 1 - \alpha\right\}.$$

In light of Theorem 2.1 of Romano and Shaikh (2010) and the fact that the functional $\sup_{w \in W'} \Gamma(\cdot, S_w)$ is increasing in W', it suffices to show that

$$\limsup_{n \to \infty} P \left\{ \sup_{w \in W_P} \Gamma(\sqrt{n}(\hat{q}^{\Delta} - q^{\Delta}); S_w) \geqslant \hat{c}_{1-\alpha}(W_P) \right\} \leqslant \alpha.$$

However, this follows immediately from Theorem B.2 and the Continuous Mapping Theorem, as $\sup_{w \in W_P} \Gamma(\cdot, S_w)$ is a continuous functional. ■

C Empirical Results Under Block Bootstrapping

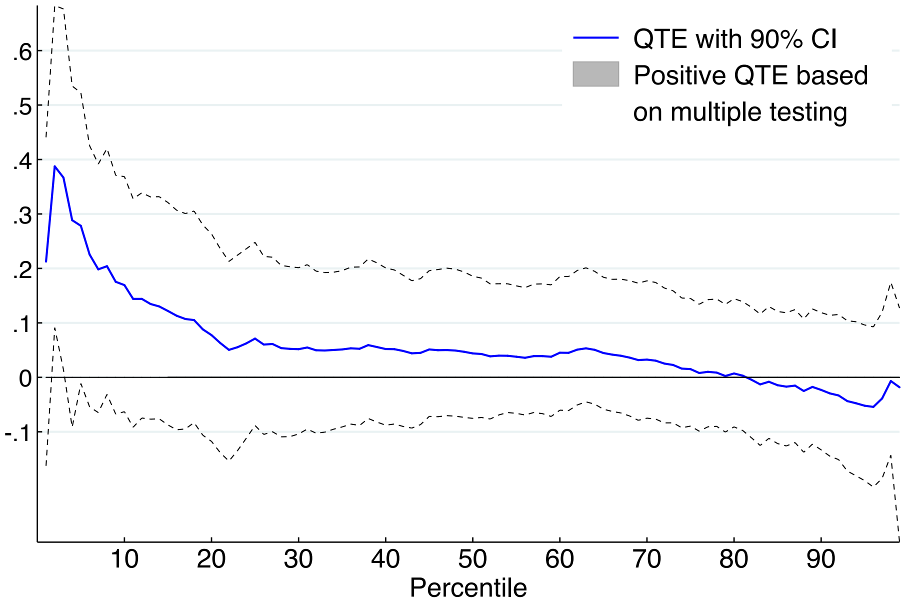
In our empirical application, we follow Andrabi, Das, and Khwaja (2017) and resample individuals in our bootstrapping procedure. Here, we report pointwise confidence intervals and multiple testing results for QTEs for the full sample under block bootstrapping. The blocks used in this analysis correspond to villages. The results in Figure 2 show that only one percentile (the second) has a pointwise statistically significant QTE. When applying our multiple testing procedure, we do not find any signifiant QTEs. In this empirical application, the within village correlation structure is such that estimating QTE leads to completely uninformative results.

D Additional Application: Connecticut's Jobs First Welfare Experiment

In this section, we present results from an application of our multiple testing procedure to a welfare experiment carried out in Connecticut in the mid-1990s. Bitler, Gelbach, and Hoynes (2006, 2017) analyze the distributional effects of this welfare reform. In the following, we describe the policy background and data, provide a simple model of labor supply that predicts heterogeneous treatment effects, and discuss the multiple testing results. This section and analyses are motivated in footnote 19 of the main text where we illustrate the broader applicability of not only the tests that we develop, but how the hypotheses themselves are motivated by economic theory.

D.1 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)



Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5% (see hypothesis (H.3) in Table 1).

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

was passed in 1996.[3] 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 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

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

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.[4] 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, 2017; 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, and as also noted by Bitler, Gelbach, and Hoynes (2006), 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.

[4] 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 (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

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.

D.2 Economic Model Predicting Heterogeneous Treatment Effects

A simple static labor supply model motivates our investigation of treatment effect heterogeneity.[5] Individuals maximize their utility over consumption (C) and earnings (E) subject to a budget constraint:

$$\max_{C,E} U = U\left(C, E(X_1); X_2\right) \tag{93}$$

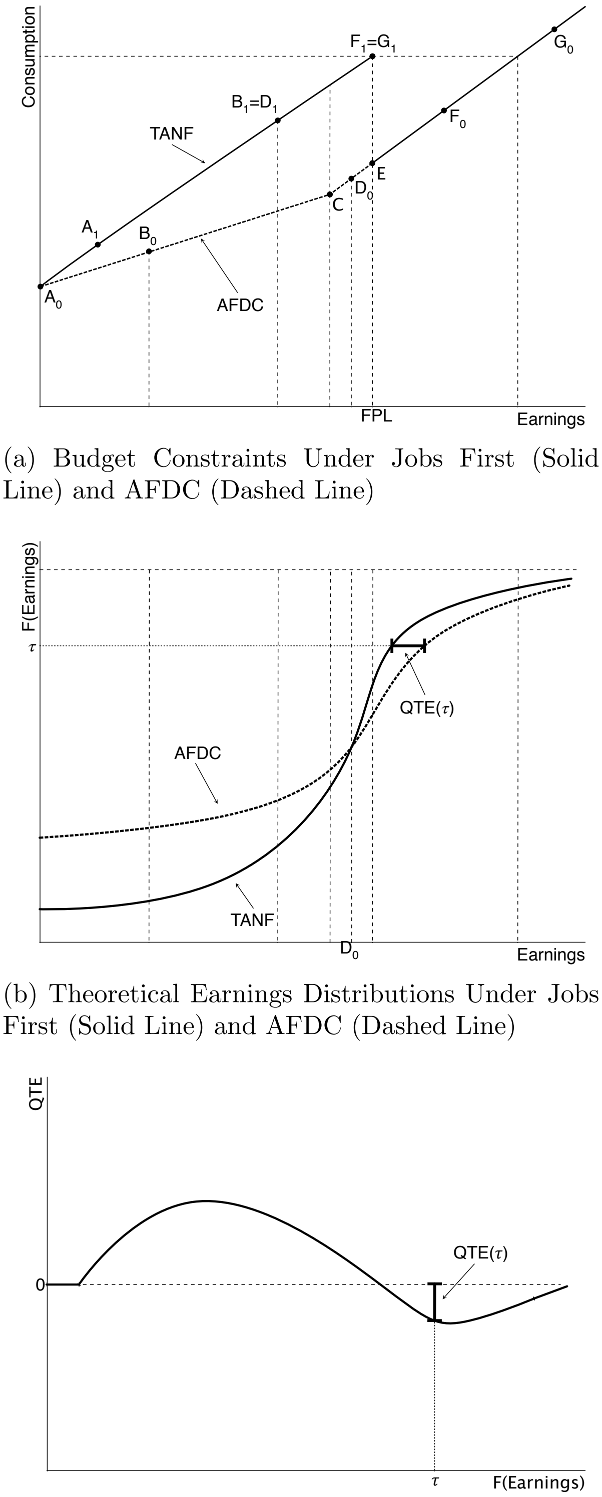
$$\text{s.t. } C = E(X_1) + W\left(E(X_1); X_3, Z^t\right) \tag{94}$$

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 3 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.[6] The solid line in the top panel of Figure 3 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 3 presents hypothetical cumulative distribution functions of earnings for those in AFDC (dashed) and Jobs First (solid) groups that are the result of

[5] 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.

[6] 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 3: Theoretical Predictions of a Static Labor Supply Model

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.

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.[7] 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.[8] 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

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

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

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 3.

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 (93) and (94) 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 3. 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.[9] 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

[9] 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.

development of tools to assess its extent in general, as well as in the specific context of the Jobs First experiment.

D.3 Results

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.[10] To balance covariates between the Jobs First and AFDC groups, we estimate the propensity score $\hat{p}(x)$ using a series logit specification.[11] 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 4 shows our estimated QTEs for the full sample along with pointwise 90 percent confidence intervals.[12] Similar to Bitler, Gelbach, and Hoynes (2006) we find pointwise significant treatment effects extending from the 48th to the 80th percentile.[13] 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 D.2.

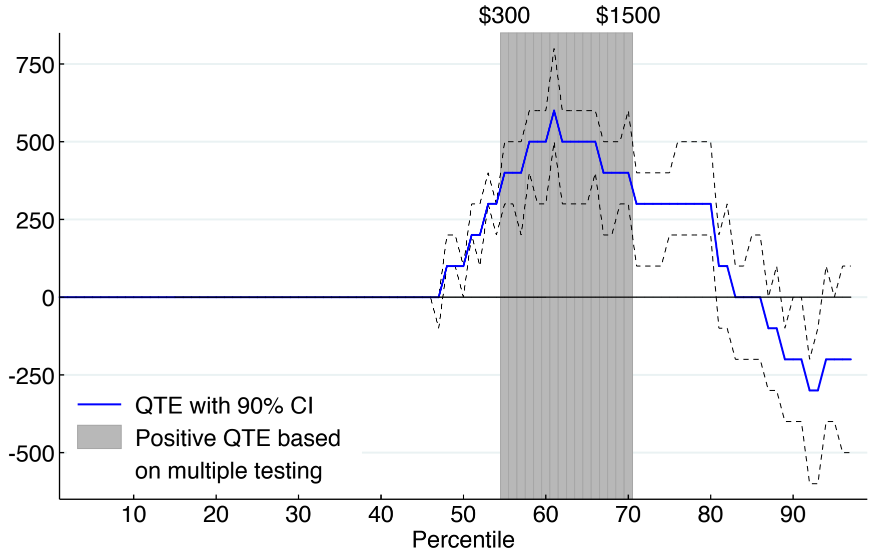
Table 2 summarizes the test result for hypotheses (H.1) and (H.2) in the main text proposed in Section 3.1 in the main text. First, we test the null hypothesis of no positive treatment effect at any percentile. As shown in Figure 4, the largest QTE (which occurs at the 61st percentile) equals 600, so this value 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

[10] 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 D.2). However, positive QTEs imply that the treatment has a positive effect for some interval of the earnings distribution (Bitler, Gelbach, and Hoynes, 2006).

[11] 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. We include the following covariates into the logit specification: age, race, education, marital status, number of children, and employment and welfare histories.

[12] 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.

[13] 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 3.1 of the main text 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 4: Quantile Treatment Effects and Multiple Testing Results, No Subgroups

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, i.e. hypothesis (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.

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

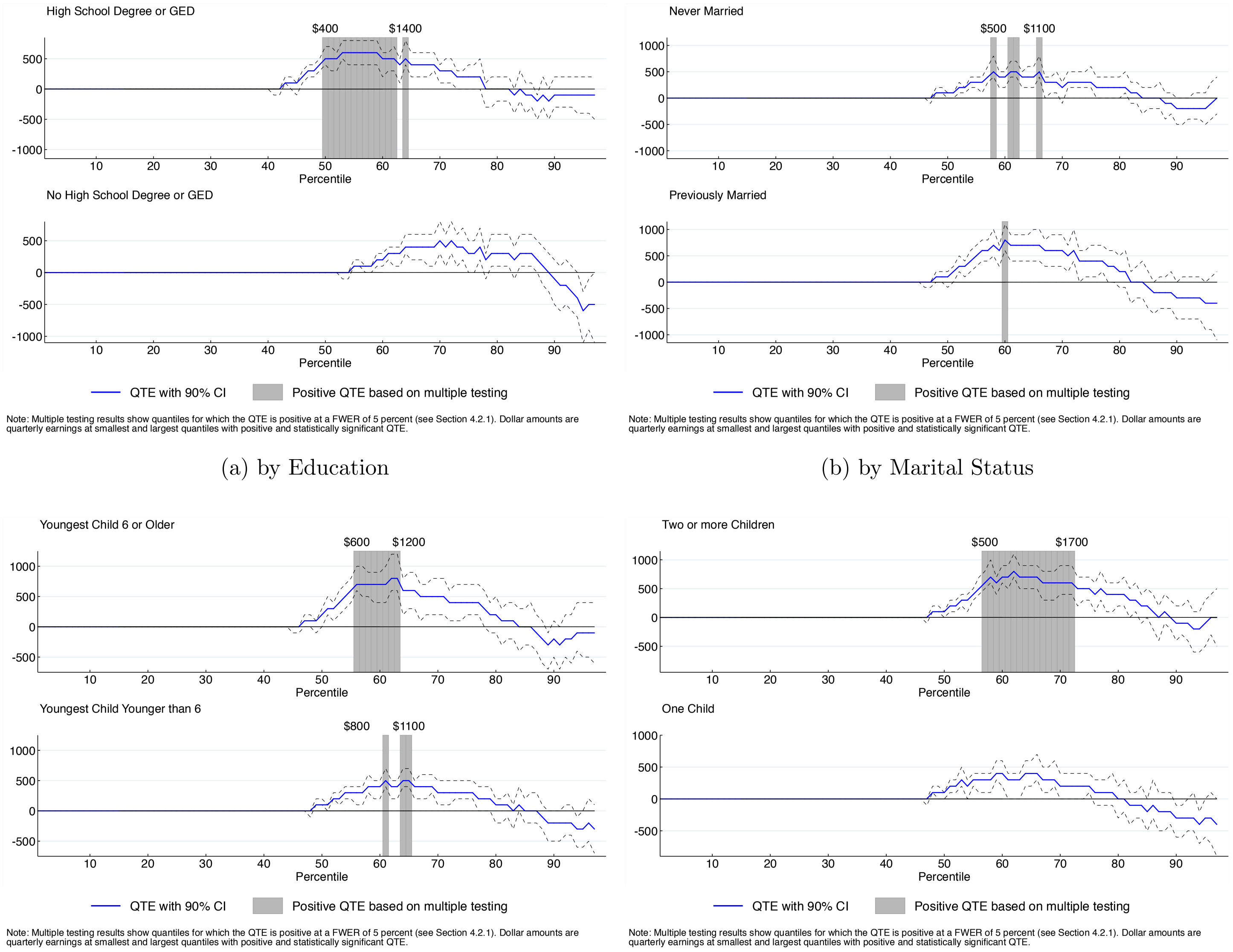
Notes: This table shows test results for hypotheses (H.1) and (H.2) in the main text, i.e. we test that there is no positive treatment effect for all quantiles and that the treatment effect is the same for all quantiles, respectively.

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) in the main text. As described in Section 3.1, 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 4 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 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 D.2, we find that there is a more limited range of individuals who increase their labor supply when assigned to the Jobs First group.

Next, we present results incorporating subgroups using the tests described in Section 3.2 in the main text. As discussed in Section D.2, 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 (2017) and informed by the model presented in Section D.2, 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.

Figures 5 and 6 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) in the main text. 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).

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 5 show that women with a high school



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

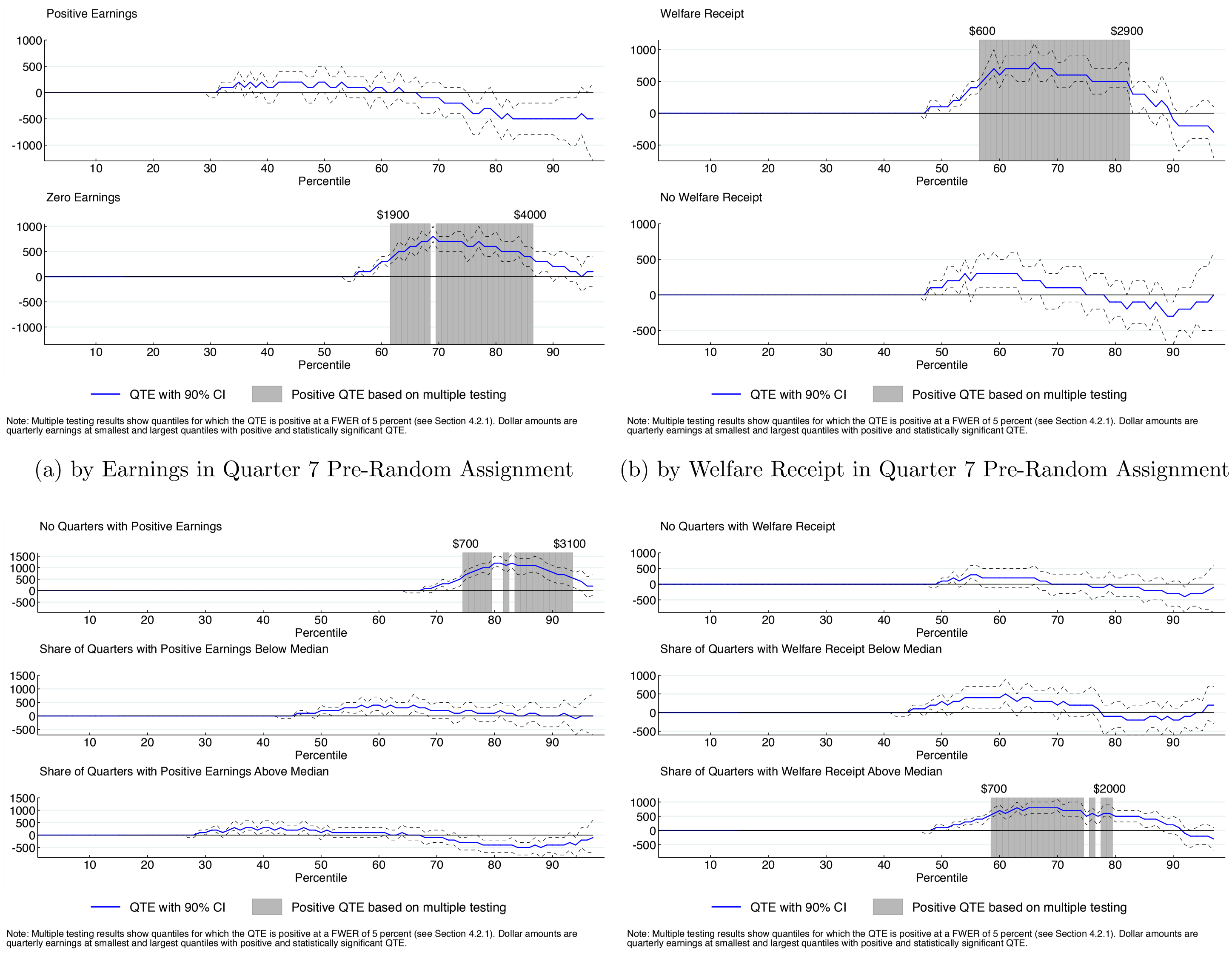
Note: Multiple testing results show quantiles for which the QTE is positive at a FWER of 5 percent (see Section 4.2.1). Dollar amounts are quarterly earnings at smallest and largest quantiles with positive and statistically significant QTE.

(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 5: 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 6: Quantile Treatment Effects and Multiple Testing Results, Earnings and Welfare History Subgroups

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 D.2. 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.[14] Bitler, Gelbach, and Hoynes (2017) 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 6 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 4, 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

[14] 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	567.11	0.093093
Marital status	672.16	587.58	0.028028
Age of youngest child	653.61	525.98	0.021021
Number of children	623.71	497.22	0.02002
Earnings in quarter 7 pre-treatment	616.49	705.26	0.089089
Welfare receipt in quarter 7 pre-treatment	617.53	508.81	0.022022
Share of quarters with positive earnings	979.38	721.19	0.007007
Share of quarters on welfare	589.69	766.65	0.13113

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

increase 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 6. 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) in the main text 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 (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 (2017), our test relaxes the strong assumption of treatment effect homogeneity within subgroups that is implicit in their test.

Table 4: Testing Which Subgroups Exhibit Treatment Effect Heterogeneity

Subgroup category	Test statistic	p-value
Education		
High School Degree or GED	477.32	0.995
No High School Degree or GED	424.74	0.995
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	0.005
Number of Children		
Two or more Children	623.71	0
One Child	358.76	0.045
Earnings in Quarter 7 Pre-Treatment		
Positive Earnings	278.35	0
Zero Earnings	616.49	0.095
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
Below Median	311.34	0.675
Above Median	337.11	0.675
Share of Quarters on Welfare		
No Quarters with Welfare Receipt	306.19	0.87
Below Median	422.68	0.42
Above Median	589.69	0.01

Notes: This table shows test results for hypothesis (H.6) in the main text, 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.

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

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.

Allen, Jason, Robert Clark, and Jean-François Houde. 2014. "The Effect of Mergers in Search Markets: Evidence from the Canadian Mortgage Industry." *American Economic Review* 104 (10):3365–3396.

Andrabi, Tahir, Jishnu Das, and Asim Ijaz Khwaja. 2017. "Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets." *American Economic Review* 107 (6):1535–63.

Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. 2009. "Incentives and Services for College Achievement: Evidence from a Randomized Trial." *American Economic Journal: Applied Economics* 1 (1):136–163.

Bandiera, Oriana, Robin Burgess, Narayan Das, Selim Gulesci, Imran Rasul, and Munshi Sulaiman. 2017. "Labor Markets and Poverty in Village Economies." *Quarterly Journal of Economics* 132 (2):811–870.

Banerjee, Abhijit V., Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. 2015. “The Miracle of Microfinance? Evidence from a Randomized Evaluation.” *American Economic Journal: Applied Economics* 7 (1):22–53.

Behaghel, Luc, Clément de Chaisemartin, and Marc Gurgand. 2017. “Ready for Boarding? The Effects of a Boarding School for Disadvantaged Students.” *American Economic Journal: Applied Economics* 9 (1):140–164.

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.

———. 2017. “Can Variation in Subgroups’ Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment.” *Review of Economics and Statistics* 99 (4):683–697.

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.”

Brown, Jason, Mark Duggan, Ilyana Kuziemko, and William Woolston. 2014. “How Does Risk Selection Respond to Risk Adjustment? New Evidence from the Medicare Advantage Program.” *American Economic Review* 104 (10):3335–3364.

Crepon, Bruno, Florencia Devoto, Esther Duflo, and William Pariente. 2015. “Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco.” *American Economic Journal: Applied Economics* 7 (1):123–150.

Evans, William and Craig Garthwaite. 2012. “Estimating Heterogeneity in the Benefits of Medical Treatment Intensity.” *Review of Economics and Statistics* 94 (3):635–649.

Fack, Gabrielle and Camille Landais. 2010. “Are Tax Incentives for Charitable Giving Efficient? Evidence from France.” *American Economic Journal: Economic Policy* 2 (2):117–141.

Fairlie, Robert and Jonathan Robinson. 2013. “Experimental Evidence on the Effects of Home Computers on Academic Achievement among Schoolchildren.” *American Economic Journal: Applied Economics* 5 (3):211–240.

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." mimeo.

Giné, Evarist. 1997. "Lecture Notes on Some Aspects of the Bootstrap." In *Ecole de Éte de Calcul de Probabilités de Saint-Flour. Lecture Notes in Mathematics*, vol. 1665.

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.

Kato, Kengo. 2009. "Asymptotics for Argmin Processes: Convexity Arguments." *Journal of Multivariate Analysis* 100 (8):1816–1829.

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.

Lehrer, Steven F., R. Vincent Pohl, and Kyungchul Song. 2020. "Multiple Testing and the Distributional Effects of Accountability Incentives in Education."

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

Massart, Pascal. 2007. *Concentration Inequalities and Model Selection.* Berlin, Heidelberg: Springer-Verlag.

McKenzie, David. 2017. "Identifying and Spurring High-Growth Entrepreneurship: Experimental Evidence from a Business Plan Competition." *American Economic Review* 107 (8):2278–2307.

Meyer, Bruce D and James X Sullivan. 2008. "Changes in the Consumption, Income, and Well-Being of Single Mother Headed Families." *American Economic Review* 98 (5):2221–2241.

Muralidharan, Karthik, Paul Niehaus, and Sandip Sukhtankar. 2016. "Building State Capacity: Evidence from Biometric Smartcards in India." *American Economic Review* 106 (10):2895–2929.

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

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.

van der Vaart, Aad W. 1996. "New Donsker Classes." *Annals of Statistics* 24:2128–2140.

van der Vaart, Aad W. and Jon A. Wellner. 1996. *Weak Convergence and Empirical Processes.* New York: Springer-Verlag.

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