

Appendix: Can Variation in Subgroups' Average Treatment Effects
Explain Treatment Effect Heterogeneity? Evidence from a Social
Experiment

Marianne P. Bitler

University of California, Irvine and NBER

Jonah B. Gelbach

University of Arizona

Hilary W. Hoynes

University of California, Davis and NBER

This version: May 2010

A Appendix

This appendix contains details and notation for the testing for the adequacy of mean treatment effects. In Sections B and C, we provide a detailed discussion of the exact null hypotheses examined in our main paper. There, we also explain how we estimate the relevant Kolmogorov-Smirnov-like test statistic for testing the adequacy of the constant treatment effects within subgroup model for explaining the heterogeneity in the full-sample data, and lay out the bootstrap procedure used for inference with the test statistic.

In Section D, the ideas behind the basic constant treatment effect within subgroup (CTEWS) model are illustrated by simulation. Two data generating processes are used; one for which the participation-adjusted CTEWS model holds and one for which none of the CTEWS models hold. For these two processes, we show how the synthetic QTE created without adjusting for participation never well approximate the actual QTE. We also show, however, that the participation-adjusted synthetic QTE are a very close match to the actual QTE for the first process (as expected since the first data generating process is one to which the participation-adjusted CTEWS model applies). Finally, we show that with heterogeneity in the treatment effects within subgroup, even the participation-adjusted synthetic QTE look nothing like the actual data QTE.

B Average Treatment Effects and Quantile Treatment Effects: Notation

Recall the notation introduced for our basic model of causal effects. $D_i = 1$ if observation i receives the treatment, and 0 otherwise. Let $Y_{it}(d)$ be i 's counterfactual value of the outcome Y in period t if person i has $D_i = d$. Ideally, we want to uncover characteristics of the joint distribution of $(Y(0), Y(1))$.

Recall that the treatment effect for individual i in period t is defined as follows: $\delta_{it} \equiv Y_{it}(1) - Y_{it}(0)$. We use δ_t to denote the average over the population of δ_{it} for period t , and we use δ to represent the average over the population of δ_{it} for all periods. By virtue of random assignment, we can estimate the average effect of the policy for period t consistently using the difference in sample means: $\bar{\delta}_t \equiv \bar{Y}_t(1) - \bar{Y}_t(0)$; similarly, $\bar{\delta} \equiv \bar{Y}(1) - \bar{Y}(0)$ is a consistent estimate of $E[\delta_{it}]$, where this expectation is taken over all i and t .

In the program evaluation literature, heterogeneity is most commonly introduced by estimating

mean treatment effects for subgroups of the population.¹ Subgroups used in practice might consist of those women, in both the treatment and control group, who share a certain race, education, or history of welfare use.

Define $Y_{it}^g(d)$ as the counterfactual outcome value in period t for person i who is a member of subgroup g when she has treatment status d . Again, under random assignment, we can estimate the mean treatment effect for each subgroup g and period t by differencing subgroup means between the treatment and control groups: $\bar{\delta}_t(g) \equiv \bar{Y}_t^g(1) - \bar{Y}_t^g(0)$. Accounting for imbalance in treatment assignment simply requires calculating weighted means using the inverse propensity scores as weights.

The mean of Y is just one identified feature of the joint treatment and control group distributions. More generally, the marginal distributions $F_0(y)$ and $F_1(y)$ are always identified, where $F_d(y) \equiv \Pr[Y_i(d) \leq y]$ for a randomly drawn i . Quantile treatment effects (QTE) are simple features of these marginal distributions.² For treatment d , the q^{th} quantile of distribution F_d is defined as $y_{qd} \equiv \inf_y \{y : F_d(y_{qd}) \geq q\}$. The quantile treatment effect for quantile q is then $\Delta_q = y_{q1} - y_{q0}$. As noted in the main text, we can account for inverse propensity score weighting by defining the empirical *cdf* as $\hat{F}_d(y) \equiv \sum_{i=1}^n 1(D_i = d) \cdot 1(Y_i(d) \leq y) \cdot \hat{\omega}_i / \sum_i 1(D_i = d) \cdot \hat{\omega}_i$, where $\hat{\omega}_i$ is the estimated inverse propensity score weight, and then proceeding as before. The QTE for quantile q may be estimated consistently using the difference across treatment status in the two outcome quantiles. As with mean treatment effects, we can estimate QTE for subgroups of the population by calculating quantiles within these subgroups and proceeding as above.

C Testing for Adequacy of Mean Treatment Effects

Our goal here is to evaluate a common, parsimonious treatment effects model to which many observers seem to subscribe. As suggested above, this model involves treatment effects that are constant within subgroups of women such as those with common educational attainment, age

¹There is an extensive literature on treatment effect heterogeneity (surveyed in Angrist (2004)), including Imbens & Angrist (1994), Heckman & Vytlacil (2001), Heckman & Vytlacil (1999), Hotz, Imbens & Klerman (2006), Abadie (2002), Crump, Hotz, Imbens & Mitnik (2009), Crump, Hotz, Imbens & Mitnik (2008). Papers that use quantile treatment effects or instrumental variable quantile treatment effects to investigate this heterogeneity in various contexts include Heckman, Smith & Clements (1997), Firpo (2007), Abadie, Angrist & Imbens (2002), Friedlander & Robins (1997), Koenker & Biliias (2001), Djebbari & Smith (2008), and Chernozhukov & Hansen (2005). Other papers that consider the distribution of treatment effects include Athey & Imbens (2006) Poirier & Tobias (2003), and Wu & Perloff (2006).

²Koenker & Biliias (2001) apply quantile regression (Koenker & Bassett (1978)) to an experiment evaluating unemployment benefits.

of youngest child, or earnings histories; but which may vary across such subgroups. Below we discuss extensions to allow for variation across time since random assignment and different intensive- and extensive-margin impacts. To evaluate our three constant treatment effects within subgroup (CTEWS) models, we ask whether the CTEWS model can generate the heterogeneity observed in the actual sample data. The trick in addressing this question is to find a way to compare observed nonparametric earnings distributions with counterfactual distributions that result when various CTEWS null hypotheses are imposed.

In Section C.1, we show how to construct a class of distributions that obtain under three null hypotheses meant to encompass, and extend, the CTEWS model. In Section C.2, we show how to (a) estimate the relevant null distributions using experimental data and (b) implement bootstrap-based inference procedure to test the null hypotheses developed in Section C.1. We present the results of these tests in Section 4.4 of the main text.

C.1 Formulating CTEWS Null Hypotheses

Under the time-constant CTEWS model, every person in group g has the treatment effect δ_g . Thus we can state our null hypothesis for the time-constant CTEWS model as

$$H_{01} : Y_{1igt} = \tilde{Y}_{1igt}^1, \text{ where } \tilde{Y}_{1igt}^1 \equiv Y_{0igt} + \delta_g, \quad (1)$$

where we recall that Y_{digt} is person i 's realized outcome during the t^{th} period since random assignment given that she is a member of group g and has treatment assignment $D_i = d$. The parameter δ_g is the average effect of treatment on women in group g . Under this null hypothesis, mean treatment effects do not vary with time since random assignment.

A more realistic null hypothesis that does allow treatment effects to vary with time is the time-varying CTEWS model, under which the null is

$$H_{02} : Y_{1igt} = Y_{0igt} + \delta_{gt} \text{ for all } i. \quad (2)$$

The only difference between H_{01} and H_{02} is that the former constrains δ_{gt} to be time-invariant. Nevertheless, H_{02} does not allow treatment effects to vary within subgroup-by-time cells. This is an important observation because, for a group- g woman whose time- t earnings equal 0 when assigned

to AFDC, both null hypotheses H_{01} and H_{02} impose exact earnings levels of δ_g (under H_{01}) or δ_{gt} (under H_{02}) on the woman's potential earnings when she is assigned to Jobs First. We have seen that basic theory predicts that some women will have zero earnings under both assignments, and both sample distributions do have a substantial share of women with zero earnings. Even after allowing for heterogeneity in treatment effects across time, for example, it would be unsurprising to reject H_{02} simply because there is a nonzero mean treatment effect. As Heckman et al. (1997) have noted, the sensitivity of constant mean treatment effects models to such rejection is both undeniable and rarely acknowledged.

One contribution of this paper is that we construct a third, more realistic null hypothesis that allows nonzero, constant mean treatment effects among women with positive AFDC-group earnings even as it allows potential outcomes to be zero under both assignments. Under this null, a woman's treated outcome Y_{1igt} equals a random draw from a non-degenerate probability distribution in the event that her control outcome Y_{0igt} equals zero. Let the share of group- g women whose potential time- t earnings would be zero under program-group assignment $D_i = d$ be written p_{dgt} , and let the conditional distribution of earnings among group- g women with positive earnings at time t and program assignment d be $F_{dgt}(\cdot|y > 0)$. Finally, define $\delta_{gt}^{positive}$ as the time- t mean treatment effect among group- g women with positive earnings, i.e., the difference in the means of the conditional distributions $F_{1gt}(\cdot|y > 0)$ and $F_{0gt}(\cdot|y > 0)$. Our third null hypothesis, for what we call the extended CTEWS model, is

$$H_{03} : \text{ for all } i, Y_{1igt} = \begin{cases} Y_{0igt} + \delta_{gt}^{positive}, & Y_{0igt} > 0 \\ X_{gt}, & \text{otherwise} \end{cases} \quad (3)$$

where the random variable X_{gt} equals 0 with probability p_{1gt}/p_{0gt} and, with probability $(1 - p_{1gt}/p_{0gt})$, equals $\delta_{gt}^{positive}$ plus a random draw from $F_{0gt}(\cdot|y > 0)$. Note that if this null hypothesis is true, then women who work only if they face the Jobs First rules have the same distribution of earnings as those women who work under either program assignment. It is hard to devise a theoretical model which predicts this. Nonetheless, it is a simple approach which preserves the overall mean treatment effect in the same spirit as the other nulls.

Null hypothesis H_{03} has two notable properties:

1. **Equal probability of work.** Aside from a minor weighting issue discussed below, the probability that a random earnings draw from the null distribution among group- g women equals zero necessarily equals p_{1gt} . To see this fact, observe that under H_{03} ,

$$Pr(Y_{1igt} = 0) = p_{0gt} \cdot Pr(X_{gt} = 0) \quad (4)$$

$$= p_{0gt} \cdot \frac{p_{1gt}}{p_{0gt}}, \quad (5)$$

which equals p_{1gt} .

2. **Equal conditional mean treatment effect.** Aside from weighting issues discussed below, the cross-program difference in the conditional mean of earnings among those with positive earnings among group- g women necessarily equals $\delta_{gt}^{positive}$ under the null. To see this fact, observe that under H_{03} ,

$$E(Y_{1igt}|Y_{1igt} > 0) = (1 - p_{0gt}) \cdot E[Y_{0igt} + \delta_{gt}^{positive}|Y_{0igt} > 0] + p_{0gt} \cdot (E[X_{gt}|X_{gt} > 0]), \quad (6)$$

with the second term arising because under H_{03} , a woman whose control group earnings equal zero has some chance of earning $\delta_{gt}^{positive}$ plus a random draw from $F_{0gt}(\cdot|y > 0)$. By construction, when X_{gt} is positive, it equals a random draw from this conditional distribution plus $\delta_{gt}^{positive}$, so $E[X_{gt}|X_{gt} > 0] = E[\delta_{gt}^{positive} + Y_{0gt}|Y_{0gt} > 0]$. It follows that

$$E(Y_{1igt}|Y_{1igt} > 0) = \left(\delta_{gt}^{positive} + E[Y_{0igt}|Y_{0igt} > 0] \right) \cdot (1 - p_{0gt} + p_{0gt}), \quad (7)$$

which is $\delta_{gt}^{positive} + E[Y_{0igt}|Y_{0igt} > 0]$, proving the claim.

The first property ensures that null hypothesis H_{03} cannot be rejected simply because the AFDC and Jobs First program distributions have different fractions of women with zero earnings. The second property ensures that H_{03} cannot be rejected simply because the AFDC and Jobs First program distributions have different conditional earnings means among those who work. These two properties are features: They imply that any rejection of H_{03} must be due to within-group distributional effects other than (i) differences in the share who work or (ii) the difference in conditional mean earnings. That is, rejection occurs only if there are distributional differences due to factors other than $\{p_{0gt}, p_{1gt}, \delta_{gt}\}$.

C.2 Implementing Tests of the Null Hypotheses

We now turn to the problem of how to test the three null hypotheses above. In so doing, it will be useful to develop some further notation. Let

$$Y_{igt} \equiv Y_{1igt} \cdot D_i + Y_{0igt} \cdot [1 - D_i] \quad (8)$$

be the observed value of person i 's outcome in period t . From above, recall that for each treatment status $d \in \{0, 1\}$ and $q \in (0, 1)$, the q^{th} quantile for program assignment d is

$$y_{qd} \equiv \inf_y \{y : F_d(y) \geq q\}. \quad (9)$$

where F_d is the *cdf* of earnings given treatment status $d \in \{0, 1\}$. Notice that these quantiles are defined relative to the earnings distribution for the population, pooled over the full period considered: They do not specify any particular subgroup g or time t .

We will define \widehat{F}_d to be the sample analogue of F_d , i.e., the empirical distribution function for women actually observed in treatment status d . Our goal is to find a way to estimate group-specific parameters that allow us to test whether the estimated null distribution given Jobs First assignment equals the empirical distribution, up to sampling error.

The basic approach is as follows. For each person i , we define \widetilde{Y}_{1igt}^j , which represents i 's potential outcome under the relevant null hypothesis H_{0j} , $j \in \{1, 2, 3\}$, as above. We call this outcome, which depends on the particular null hypothesis, the synthetic null earnings value given Jobs First treatment, or just the synthetic earnings value when there is no cause for confusion. Given knowledge of δ_{gt} and actual AFDC-assigned earnings Y_{0igt} , i 's value of \widetilde{Y}_{1igt}^j is known. Therefore, we can use observed data on both program groups to consistently estimate δ_g , δ_{gt} , and p_{1gt} and p_{0gt} in the case of H_{03} .

Together with the actual value of Y_{0igt} among only those women who are observed in the AFDC group, i.e., those with $D_i = 0$, these estimates allow us to estimate consistently the synthetic earnings level \widetilde{Y}_{1igt}^j under each null hypothesis. We can then estimate the population synthetic earnings distribution, \widetilde{F}_1^j , using the empirical distribution of synthetic earnings values, which we call $\widehat{\widetilde{F}}_1^j$. Finally, we can compare the estimate $\widehat{\widetilde{F}}_1^j$ to the empirical earnings distribution of actual earnings, \widehat{F}_1 , among women observed with Jobs First assignment. If the null hypothesis under study is true, then these distributions will be equal up to sampling error. Under the null H_{0j} ,

$\tilde{F}_1^j = F_1$. Since \hat{F}_1^j is consistent for \tilde{F}_1^j and \hat{F}_1 is consistent for F_1 regardless of whether any null holds, we can test the null consistently by comparing \hat{F}_1^j to \hat{F}_1 .

It will be helpful to have notation for the true quantiles of the distributions of \tilde{Y}_{0igt}^j and \tilde{Y}_{1igt}^j . Observe that since $\tilde{Y}_{0igt}^j = Y_{0igt}$ by definition, $\tilde{F}_0^j = F_0$ is always true. When the null hypothesis is false, though, \tilde{F}_1^j will differ from F_1 . We denote the relevant quantiles as

$$\tilde{y}_{qd}^j \equiv \inf_y \left\{ y : \tilde{F}_d^j(y) \geq q \right\}. \quad (10)$$

Under the null hypothesis, all quantiles of \tilde{F}_1^j and F_1 will be equal; under the alternative hypothesis, some (and perhaps all) quantiles will differ. We thus have another equivalent way to specify the null hypothesis of interest:

$$H_{0j} : \tilde{y}_{q1}^j = y_{q1} \text{ for all } q \in (0, 1). \quad (11)$$

Let $\tilde{\Delta}_q^j \equiv \tilde{y}_{q1}^j - y_{q0}$ be the QTE at quantile q when the null hypothesis H_{0j} is correct, and recall that $\Delta_q \equiv y_{q1} - y_{q0}$ is the population QTE at quantile q . Our final way to write the null hypothesis H_{0j} is thus

$$H_{0j} : \tilde{\Delta}_q^j = \Delta_q \text{ for all } q \in (0, 1). \quad (12)$$

For given q and null hypothesis H_{0j} , we can estimate both $\tilde{\Delta}_q^j$ and Δ_q consistently using sample quantiles of the distributions of \tilde{Y}_{1igt}^j and Y_{1igt} . The following procedure provides an overview of how we estimate these sample quantiles and test the null hypotheses implied by our three definitions of \tilde{Y}_{1igt}^j .³

³For null hypotheses H_{01} and H_{02} , there is another, equivalent approach: We could simply estimate a set of Q quantile regression models in which the left hand side variable is individual-level earnings, and the right-hand side variables include a full set of dummies for subgroup and interactions with the experimental program dummy (we thank Pat Kline for making this point to us). The null hypothesis H_{01} could be tested by testing the joint null that all interactions involving the program dummy equal zero. In the case of H_{02} , we would have to include interactions of all subgroup and time dummies with each other, as well as with the program dummy. We would then need to estimate a variance matrix with dimensionality equal to the product of Q and the number of fully interacted dummies; given the panel nature of our data, the only feasible way to do this appears to be to use a bootstrap approach, which we do below for our preferred methodology as well. We take the approach we do because (a) we see little computational gain in the alternate approach; (b) our approach allows a direct test using the metric in which we are most interested, which is the difference between null and observed earnings distributions; and (c) there is no obvious parametric quantile regression analog to the alternative approach in the case of H_{03} .

Procedure 1 (Testing H_{0j} from (11)).

Let Q be some integer greater than 0 and less than 100. For ease of exposition, we will focus on the case in which we are interested in all quantiles $q = 1, 2, \dots, Q$ (the method applies more generally), and for notational simplicity, adapt the convention that when $q \in [1, 99]$, the quantile of interest is actually $q/100$. The following procedure provides a consistent test of the null hypotheses.

1. Calculate the sample quantiles of the actual data for the Jobs First group, $\{\hat{y}_{q1}\}_{q=1}^Q$.
2. Estimate the mean treatment effect for each subgroup g and period t under null hypothesis H_{0j} ; call these $\hat{\delta}_{gt}^j$.
3. Use the estimates $\hat{\delta}_{gt}^j$ (and, for H_{03} , \hat{p}_{1gt} and \hat{p}_{0gt}) to estimate \tilde{Y}_{igt}^j under each null H_{0j} , and then calculate estimates, \tilde{y}_{qd}^j , of the quantiles of \tilde{F}_1^j , which are correct only under null H_{0j} .
4. For each null hypothesis H_{0j} , calculate a test statistic, \hat{S}^j , based on how close the set of quantiles \tilde{y}_{qd}^j is to the set of quantiles \hat{y}_{q1} .
5. Use a bootstrap procedure to estimate a critical value for each \hat{S}^j under null hypothesis H_{0j} , and reject the null if \hat{S}^j exceeds this critical value.

□

We now discuss the details of each of these steps.

C.2.1 Step 1: Estimating the Jobs First sample quantiles

To calculate sample quantiles for the Jobs First group, recall that \hat{F}_1 is the empirical distribution function of observations with $D_i = 1$. The sample quantiles are defined implicitly as

$$\hat{y}_{q1} \equiv \inf_y \left\{ y : \hat{F}_1(y) \geq q \right\}. \quad (13)$$

C.2.2 Step 2: Estimating mean treatment effects

We use three approaches to estimating the mean treatment effect δ_{gt} . In testing the time-constant CTEWS model, we constrain the mean treatment effects to be constant across all t , so we use

$$\hat{\delta}_{gt}^1 \equiv \bar{Y}_{1g} - \bar{Y}_{0g} \text{ for all } t, \quad (14)$$

where \bar{Y}_{dg} is the sample mean earnings in program group d and demographic subgroup g , pooling over all time periods. When we test the time-varying CTEWS model, we allow treatment effects to vary with time since random assignment, so we use

$$\widehat{\delta}_{gt}^2 \equiv \bar{Y}_{1gt} - \bar{Y}_{0gt}, \quad (15)$$

where the sample mean earnings here have the obvious definitions. Finally, when we test the extended CTEWS model, we estimate a time-varying conditional treatment effect given positive earnings:

$$\widehat{\delta}_{gt}^3 \equiv \frac{\bar{Y}_{1t}^g}{1 - \widehat{p}_{1gt}} - \frac{\bar{Y}_{0t}^g}{1 - \widehat{p}_{0gt}}, \quad (16)$$

where \widehat{p}_{1gt} and \widehat{p}_{0gt} are the estimated share of Jobs First and AFDC group observations with zero earnings.

C.2.3 Step 3: Estimating Quantiles of \widetilde{F}_1^j

To estimate the quantiles of F_1 under each null, we first use the appropriate CTEWS model to estimate potential earnings given Jobs First assignment, among women actually assigned to the AFDC program group. That is, we estimate \widetilde{Y}_{1igt}^j among those with $D_i = 0$. To do so, we use

$$\widehat{Y}_{1igt}^1 \equiv Y_{igt} + \widehat{\delta}_{gt}^1. \quad (17)$$

$$\widehat{Y}_{1igt}^2 \equiv Y_{igt} + \widehat{\delta}_{gt}^2. \quad (18)$$

$$\widehat{Y}_{1igt}^3 \equiv \begin{cases} Y_{igt} + \widehat{\delta}_{gt}^3, & Y_{igt} > 0 \\ \widehat{X}_{gt}, & \text{otherwise,} \end{cases} \quad (19)$$

where \widehat{X}_{gt} is a random variable drawn from a consistent estimate of the distribution of X_{gt} . In practice, we can use a reweighting scheme to avoid taking random draws when constructing \widehat{Y}_{1igt}^3 in (19). To do so, we let $\widehat{Y}_{1igt}^3 = 0$ whenever $Y_{0igt} = 0$ and $Y_{0igt} + \widehat{\delta}_{gt}^3$ whenever $Y_{0igt} > 0$. Recalling that the inverse propensity score weight for observations i defined above is $\widehat{\omega}_i$, we construct a new weight for AFDC-group observations by multiplying $\widehat{\omega}_i$ by

$$\widehat{\rho}_{igt} \equiv \frac{\widehat{p}_{1gt}}{\widehat{p}_{0gt}} \cdot 1(Y_{0igt} = 0) + \left(1 - \frac{\widehat{p}_{1gt}}{\widehat{p}_{0gt}}\right) \cdot [1 - 1(Y_{0igt} = 0)]. \quad (20)$$

This reweighting ensures that the actual treatment and synthetic group will have both the same share of observations with zero earnings and the same mean treatment effect within group-by-time cells.⁴ The resulting empirical distribution function, \widehat{F}_1^3 , is a consistent estimate of \widetilde{F}_1^3 , the population distribution of \widetilde{Y}_{1igt}^3 .

With these estimates in hand, we estimate \widetilde{F}_1^j using the empirical distribution function, \widehat{F}_1^j , of \widehat{Y}_{1igt}^j among those in the control group (notice that we pool over all subgroups and time periods to construct \widehat{F}_1^j). We note that the various estimates \widehat{F}_1^j are consistent for the true values \widetilde{F}_1^j regardless of the null's correctness: The null concerns the relationship between \widetilde{F}_1 and F_1 , whereas consistency of \widehat{F}_1^j for \widetilde{F}_1^j follows simply from the fact that $\widehat{\delta}_{gt}^j$ is consistent for δ_{gt} and each \widehat{p}_{dgt} is consistent for p_{dgt} , given random assignment. Finally, we estimate the quantiles of \widetilde{F}_1^j by calculating the relevant quantiles of \widehat{F}_1^j , using the values of \widehat{Y}_{1igt}^j we constructed from (17)–(19). These sample quantiles are defined implicitly in the usual way:

$$\widehat{y}_{q1}^j \equiv \inf_y \left\{ y : \widehat{F}_1^j(y) \geq q \right\}. \quad (21)$$

C.2.4 Step 4: Inference

To test each null hypothesis, we use a Kolmogorov-Smirnov-type test statistic and estimate its distribution under each null hypothesis using a bootstrap method developed by Chernozhukov & Fernandez-Val (2005, henceforth, CFV). Define

⁴In fact, because the inverse propensity score weights sum to slightly different values across AFDC and Jobs First groups, these shares are not quite identical, but they are very close; imposing equality would almost certainly make no difference.

$$\widehat{S}^j \equiv \sqrt{n} \left(\sup_{q=1 \text{ to } Q} \left\{ \frac{|\widehat{y}_{q1} - \widetilde{y}_{q1}^j|}{(\widehat{V}^j(q))^{1/2}} \right\} \right), \quad (22)$$

where n is the overall sample size and $\widehat{V}^j(q)$ is a consistent estimate of the variance of the discrepancy term $\widehat{y}_{q1} - \widetilde{y}_{q1}^j$. The discrepancy terms themselves are the differences between the always-consistent and consistent-only-under-the-null estimates of the q^{th} quantiles among Jobs First recipients.

As CFV discuss, dependence in the data-generating process causes the distribution of \widehat{S}^j to have non-standard properties, complicating inference using standard methods. This matters in our case because we have repeated observations on each woman in the sample. CFV show that a bootstrap procedure provides a basis for consistent inference on the Kolmogorov-Smirnov-like statistic \widehat{S}^j . This bootstrap procedure involves doing the following procedure B times, where B is the number of bootstrap replications (we discuss calculation of the estimated variance term $\widehat{V}^j(q)$ below).⁵

1. Re-sample the data in a manner consistent with the data generating process using the non-parametric block bootstrap. That is, re-sample entire individual earnings profiles. Since individuals are assigned to treatment or control status in an *iid* fashion, this re-sampling approach reproduces the properties of the underlying data generating process.
2. Repeat the steps described in Sections C.2.1–C.2.3 using the re-sampled data. We use a b superscript to indicate that the estimated sample quantile \widehat{y}_{q1}^{jb} or \widetilde{y}_{q1}^{jb} is based on the b^{th} bootstrap re-sample rather than the real data.
3. Calculate the bootstrap estimate of the statistic \widehat{S} , denoted

$$\widehat{S}^{jb} \equiv \sqrt{n} \left(\sup_{q=1 \text{ to } Q} \left\{ \frac{|\widehat{y}_{q1}^{jb} - \widetilde{y}_{q1}^{jb} - (\widehat{y}_{q1}^j - \widetilde{y}_{q1}^j)|}{(\widehat{V}^j(q))^{1/2}} \right\} \right). \quad (23)$$

This statistic differs in form from \widehat{S}^j in a key way: For each quantile, we create the bootstrap supremum by subtracting the real-data discrepancy term $\widehat{y}_{q1}^j - \widetilde{y}_{q1}^j$ from the corresponding

⁵As applied to our context, CFV's assumptions require certain asymptotic normality properties for the sample quantiles. Given the discrete nature of our earnings data, our use of CFV's method thus appears not to be justified by their results. However, Gelbach (2005) shows that the bootstrap can be used to consistently estimate the distribution of quantile treatment effects with discrete data. We believe but have not proved that the bootstrap can also be used to estimate the null distribution of the statistic \widehat{S} .

bootstrap discrepancy term. This step is what allows CFV's method to overcome the non-centrality problem alluded to above. Heuristically, the relevant noncentrality term can be consistently estimated using the real-data sample's estimate of the discrepancy $\hat{y}_{q1}^j - \tilde{y}_{q1}^j$. The bootstrap re-sample's noncentrality term has the same asymptotic distribution, so subtracting the real-data sample discrepancy term from the bootstrap term yields a statistic with conventional asymptotic properties under the null (and local alternatives).

Next, we use the bootstrap distribution $\{\hat{S}^{jb}\}_{b=1}^B$ to estimate the relevant critical value of the distribution of \hat{S} . In other words, letting G^j be the null distribution of \hat{S}^j , we will show how to estimate the quantile $s_{1-\alpha}^j$ defined by

$$s_{1-\alpha}^j \equiv \inf_s \{s : G^j(s) \geq 1 - \alpha\}. \quad (24)$$

To estimate this critical value, we must estimate the distribution function G^j , which we can do using the empirical bootstrap distribution $\{\hat{S}^{jb}\}_{b=1}^B$. Defining the estimated bootstrap distribution as \hat{G}^j , we have

$$\hat{G}^j(s) \equiv \frac{1}{B+1} \sum_{b=1}^B 1(\hat{S}^{jb} \leq s). \quad (25)$$

The critical value for a level- α test is thus the smallest s such that $\hat{G}^j(s) \geq 1 - \alpha$, i.e.,

$$\hat{s}_{1-\alpha}^j \equiv \inf_s \{s : \hat{G}^j(s) \geq 1 - \alpha\}. \quad (26)$$

Finally, we reject null hypothesis H_{0j} if and only if $\hat{S}^j \geq \hat{s}_{1-\alpha}^j$. For instance, if we are interested in a level-0.05 test, with $B = 999$ we would find the 50th largest bootstrap estimate of the test statistic (since $[1 - 0.05] \cdot [999 + 1] = 950$, and the 950th smallest estimate is the 50th largest when $B = 999$) and then reject the null hypothesis if and only if the real-data test statistic \hat{S}^j exceeded that value.

The only remaining task is to calculate the estimated variances given by $\hat{V}^j(q)$. An easy way to do this is to use the bootstrap distribution of the discrepancy terms. To see how, define the discrepancy term for the q^{th} quantile as $\hat{r}_q^j \equiv \hat{y}_{q1}^j - \tilde{y}_{q1}^j$. Similarly, define the iteration- b re-sampled estimate of this discrepancy term for the q^{th} quantile as $\hat{r}_q^b \equiv \hat{y}_{q1}^{jb} - \tilde{y}_{q1}^{jb}$. We can estimate the variance of each discrepancy term \hat{r}_q^j with the sample variance of the re-sampled estimates of this

vector. That is, for each q , we calculate

$$\widehat{V}^j(q) \equiv \frac{1}{B-1} \sum_{b=1}^B \left(\widehat{r}_q^{jb} - \bar{r}_q^j \right)^2 \quad (27)$$

where \widehat{r}_q^{jb} is the bootstrap analogue of \widehat{r}_q^j calculated using the b^{th} re-sample and \bar{r}_q^j is the bootstrap sample mean of this statistic. We can then use this estimate of $\widehat{V}^j(q)$ in the appropriate places above.⁶

These are the test statistics that we report in Table 4.

D Comparison of Synthetic and Actual QTE with Participation-Adjusted CTEWS Data Generating Process and with Data Generating Process with Heterogeneity within Subgroup

To illustrate the basic ideas here, we use a simulation based on data generating processes we can control directly. We assume there are two subgroups, indexed by subscripts A and B . If a person from subgroup A is assigned to the control group, her potential outcome is $X_A(0) = \max[0, Z]$, where Z is a standard normal random variable. If the same woman is assigned to the treatment group, her potential outcome is $X_A(1) = \max[0, X_A(0)] + 0.3 \cdot 1[X_A(0) > 0]$. We simulate subgroup- B women's potential outcomes similarly, with $X_B(0) = \max[0, 0.1 + Z]$ and $X_B(1) = \max[0, X_B(0)] + 0.5 \cdot 1[X_B(0) > 0]$. We use the notation $X(0)$ to refer to a generic draw from the control group, where subgroups A and B , and thus $X_A(0)$ and $X_B(0)$, are represented in proportion to the groups' population shares. We use the notation $X(1)$ analogously.

Each woman in group A has a treatment effect of 0.3 if her control group outcome $X_A(0)$ is pos-

⁶Notice that we use the same matrix in both steps, even though the real-data discrepancy term is subtracted from each bootstrap discrepancy term; this is appropriate because the real-data term is constant over all bootstrap iterations.

itive, and a treatment effect of zero otherwise. Group B women have a treatment effect of 0.5 when their control group outcomes $X_B(0)$ are positive, and zero otherwise. The most restrictive CTEWS model assumes that treatment effects are constant within a given subgroups. The data generating process described above is therefore obviously not an example of such a CTEWS model. However, one can reasonably think of this data generating process as equivalent to a generalized CTEWS model, in which the pooled treatment-group earnings distribution can be fully characterized by accounting for (i) a conditional mean impact among those with nonzero control-group realizations, $X_g(0)$, and (ii) a treatment impact on the share of observations with nonzero realizations. In other words, after adjusting for any subgroup-specific effects of treatment on the probability of nonzero realizations, it is sufficient under this model to add conditional mean impacts to observations with nonzero values of $X_g(0)$. As we will see, an algorithm generates a synthetic earnings distribution in this fashion does very well in replicating the observed treatment group’s earnings distribution, and thus the observed set of QTE estimates.

To mimic an experiment with 20,000 women drawn from this data generating process, we drew 5,000 realizations each of $X_A(0)$ and $X_B(0)$. To draw 5,000 realizations each of $X_A(1)$ and $X_B(1)$, we drew respectively from $\max[0, Z] + 0.3 \cdot 1[Z > 0]$ and $\max[0, 0.1 + Z] + 0.5 \cdot 1[0.1 + Z > 0]$. These latter 10,000 draws thus come from the marginal distributions for $X_A(1)$ and $X_B(1)$.⁷ We then calculate the set of quantile treatment effects based on these 20,000 observations, which we plot as the thin line in the top left panel of Figure 1. This graph shows that the QTE equal zero for the $q \leq 48$, since roughly 48 percent of both the treatment and control groups have $X = 0$. For $q > 48$, the QTE are positive, increasing slightly as q does, with a hump shape at values of q close to one.

⁷An alternative would have been to simply calculate the potential outcomes given treatment-group assignment corresponding to the 10,000 realized values of $X_A(0)$ and $X_B(0)$, since these are known given the data generating processes described above. However, actual experiments never have this form: Instead, they consist of draws from the marginals. Thus, the approach we take corresponds to what would happen in a real-world context.

By comparison, the thick line in the top left panel of Figure 1 involves synthetic QTE estimates that we calculated after imposing a restrictive version of the CTEWS model. To calculate these QTE estimates, we did the following:

1. For each subgroup A and B , we estimated the usual within-group mean treatment effect. For subgroup A , this effect was 0.135; for group B , it was 0.266.
2. We added the 0.135 to each of the 5,000 realization of $X_A(0)$, yielding 5,000 estimated synthetic realizations of $X_A(1)$. We added 0.266 to each of the 5,000 realizations of $X_B(0)$ to obtain estimated synthetic realizations of $X_B(1)$. Call these estimated synthetic realizations $\tilde{X}_A(1)$ and $\tilde{X}_B(1)$, and use $\tilde{X}(1)$ to refer to a random draw from the pooled set of realizations on $\tilde{X}_A(1)$ and $\tilde{X}_B(1)$.
3. We calculated synthetic QTE estimates by subtracting the sample quantiles of $X(0)$ from the sample quantiles of $\tilde{X}(1)$.

The resulting synthetic QTE estimates in the top left panel of Figure 1 look nothing like the QTE estimates based on the actual realizations of $X_A(1)$ and $X_B(1)$. Perhaps the most obvious difference is that there are no values of q for which the synthetic QTE equal 0. This result occurs because the version of the CTEWS model implemented here requires us to add the subgroup-specific mean treatment effect to every realization of $X_A(0)$ and $X_B(0)$. Thus, the least value of $\tilde{X}_A(1)$ is 0.135, and the least value of $\tilde{X}_B(1)$ is 0.266, as is clear from the synthetic QTE plot for values of $q \leq 48$. In addition to this failure of the CTEWS model to replicate the fact that many actual treatment-group realizations equal zero, the synthetic QTE for $q > 48$ are all clustered around 0.2. This occurs because this version of the CTEWS model fails to distinguish the zero treatment effect for group- g members with $X_g(0) = 0$ from the non-zero treatment effect for those with $X_g(0) > 0$.

In the top right panel of Figure 1, we generalize the CTEWS model to ensure that the share

of zeros in the actual QTE will be replicated exactly by the synthetic QTE. We achieve this replication in two steps. First, we set $\tilde{X}_g(1) = 0$ whenever $X_g(0) = 0$. Second, we weight the synthetic realizations so that the weighted synthetic share of zeros equals the share of zeros among the set of realized actual treatment group observations, $X(1)$ (we discuss the weights we use above in Appendix Section C.2.3). Next, we use the actual data on observations in both the treatment and control group to calculate the estimated conditional mean treatment effect among observations with $X_g(0) > 0$. For subgroup A , this conditional mean treatment effect was 0.284, while it was 0.500 for subgroup B . To construct the synthetic value, $\tilde{X}_g(1)$, for group- g observations with $X_g(0) > 0$, we then added the conditional mean treatment effect to the observed value for each such observation on $X_g(0)$. We then calculated the (weighted) sample quantiles for the pooled distribution of synthetic realizations, $\{\tilde{X}(1)\}$. Finally, we subtracted the sample quantiles for the actual control group (this step is unchanged from our previous approach).

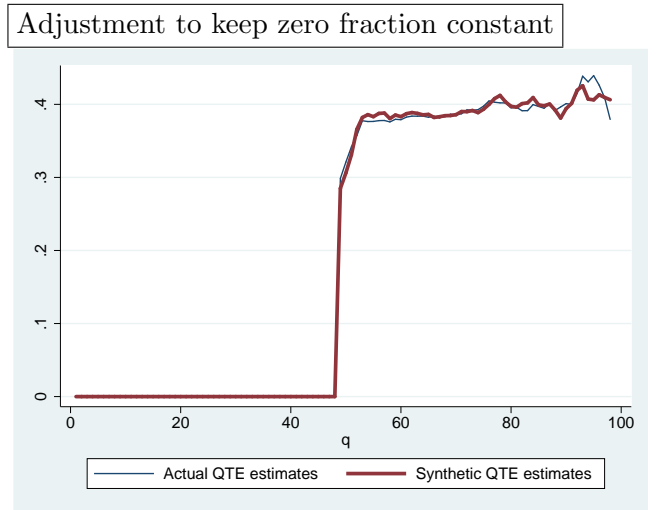
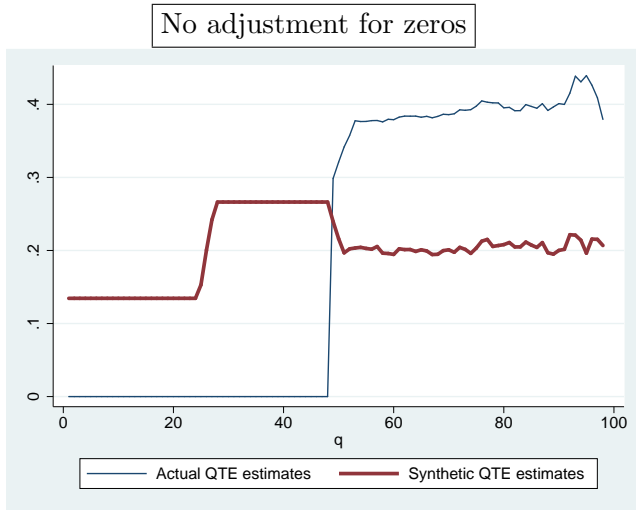
The resulting synthetic QTE estimates are given by the thick line in the top right panel of Figure 1. These synthetic QTE estimates line up almost perfectly with the actual QTE estimates. This result is to be expected, because the approach we used to construct the synthetic earnings distribution is identical to the underlying data generating process, up to sampling variation. This example shows that our second algorithm for calculating synthetic earnings among treated observations does very well when it corresponds to the underlying model. This result suggests that if the generalized CTEWS model is correct for the Jobs First experiment, then the resulting synthetic QTE should look similar to the actual estimated QTE.

The bottom panels of Figure 1 replicate the synthetic QTE estimation exercise just described. However, in each case, the true data generating process now allows for within-group heterogeneity in the true treatment effect among those with $X_g(0) > 0$. All control group realizations in these panels are identical to those in the top two panels. To construct the treatment group realizations

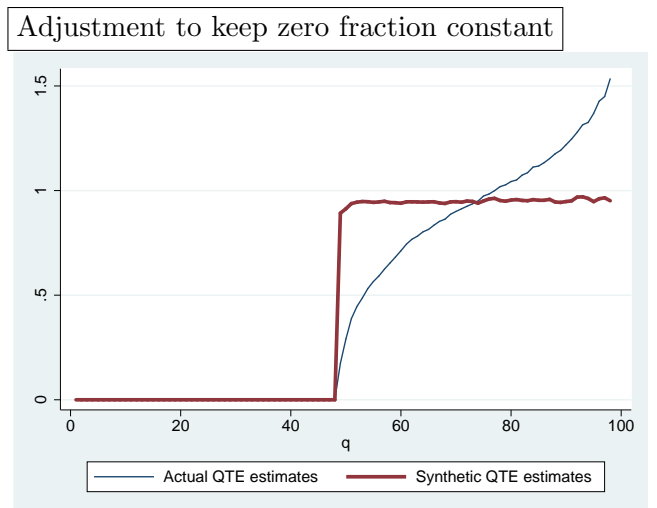
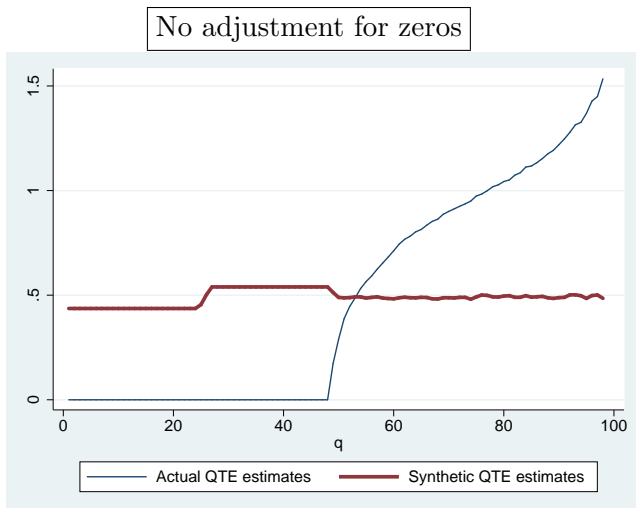
for subgroup A , we took each of the 5,000 original realizations of $X_A(1)$ and added a random draw from the truncated normal distribution with lower bound -0.3 . This ensured that all resulting treated values were above 0. Likewise, for subgroup B , we added to each original realization of $X_B(1)$ a random draw from the truncated normal distribution with lower bound -0.5 . The graphs in the lower two panels of the figure show that the resulting synthetic QTE do a very poor job of replicating the actual QTE for this data generating process. This result occurs not just when we naively ignore the fact that many treatment group observations have $X_g(0) = 0$, but also when we adjust the data to impose equality of the zero shares across synthetic and actual treatment group earnings distributions. Of course, this result is not surprising: when the true data generating process does not involve constant within-group treatment effects, intuition suggests that we should not expect to be able to replicate the treatment group earnings distribution using only subgroup-specific mean impacts.

Figure 1: Actual and synthetic earnings QTE with two simulated data generating processes

True DGP is constant treatment effects within subgroups



True DGP involves heterogeneous within-group impacts



References

- Abadie, A. (2002), 'Bootstrap tests for distributional treatment effects in instrumental variable models', *Journal of the American Statistical Association* **97**, 284–92.
- Abadie, A., Angrist, J. D. & Imbens, G. (2002), 'Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings', *Econometrica* **70**(1), 91–117.
- Angrist, J. D. (2004), 'Treatment effect heterogeneity in theory and practice', *Economic Journal* **114**, C52–C83.
- Athey, S. & Imbens, G. (2006), 'Identification and inference in nonlinear difference-in-differences models', *Econometrica* **74**(2), 431–497.
- Chernozhukov, V. & Fernandez-Val, I. (2005), 'Subsampling inference on quantile regression processes', *Sankhya: The Indian Journal of Statistics* **67**, part 2, 253–256.
- Chernozhukov, V. & Hansen, C. (2005), 'An IV model of quantile treatment effects', *Econometrica* **73**(1), 245–261.
- Crump, R., Hotz, V. J., Imbens, G. & Mitnik, O. (2008), 'Nonparametric tests for treatment effect heterogeneity', *Review of Economics and Statistics* **90**(3), 389–406.
- Crump, R., Hotz, V. J., Imbens, G. & Mitnik, O. (2009), 'Dealing with limited overlap in estimation of average treatment effects', *Biometrika* **96**(1), 187–199.
- Djebbari, H. & Smith, J. (2008), 'Heterogenous program impacts of the PROGRESA program', *Journal of Econometrics* **145**(1–2), 64–80.
- Firpo, S. (2007), 'Efficient semiparametric estimation of quantile treatment effects', *Econometrica* **75**(1), 259–276.

- Friedlander, D. & Robins, P. K. (1997), 'The distributional impacts of social programs', *Evaluation Review* **21**(5), 531–553.
- Gelbach, J. B. (2005), Inference for sample quantiles with discrete data. Available at <http://glue.umd.edu/~gelbach/papers/working-papers.html>.
- Heckman, J. J., Smith, J. & Clements, N. (1997), 'Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts', *Review of Economic Studies* **64**, 487–535.
- Heckman, J. J. & Vytlacil, E. J. (1999), 'Local instrumental variables and latent variable models for identifying and bounding treatment effects', *Proceedings of the National Academies of Science* **96**, 4730–4734.
- Heckman, J. J. & Vytlacil, E. J. (2001), Local instrumental variables, in J. Heckman & E. Leamer, eds, 'Nonlinear Statistical Inference: Essays in Honor of Takesha Ameniya', North Holland, Amsterdam.
- Hotz, V. J., Imbens, G. & Klerman, J. (2006), 'Evaluating the differential effects of alternative welfare-to-work training components: A re-analysis of the California GAIN program', *Journal of Labor Economics* **24**(3), 521–566.
- Imbens, G. W. & Angrist, J. D. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467 – 75.
- Koenker, R. & Bassett, G. (1978), 'Regression quantiles', *Econometrica* **46**, 33–50.
- Koenker, R. & Biliias, Y. (2001), 'Quantile regression for duration data: A reappraisal of the Pennsylvania reemployment bonus experiments', *Empirical Economics* **26**(1), 199–220.

Poirier, D. & Tobias, J. (2003), 'On the predictive distributions of outcome gains in the presence of an unidentified parameter', *Journal of Business and Economic Statistics* **21**(2), 258–268.

Wu, X. & Perloff, J. M. (2006), Information-theoretic deconvolution approximation of treatment effect distribution.