Can Subgroup-Specific Mean Treatment Effects Explain Heterogeneity in Welfare Reform Effects? Evidence from Connecticut's Jobs First 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*

First version: June 2007
This version: November 12, 2007

Abstract

The welfare reform literature of the last decade has largely focused on mean treatment effects. Our own recent work, Bitler, Gelbach & Hoynes (2006*b*), provides evidence of pervasive heterogeneity in the effects of Connecticut's Jobs First. To the extent that other research in the welfare reform literature has addressed treatment effect heterogeneity, the focus has largely been on mean treatment effects within small numbers of subgroups. In this paper, we again use experimental data from Jobs First to evaluate whether subgroup-specific mean treatment effects can replicate the treatment effect heterogeneity we measured previously. We show how to construct null hypotheses under which the treatment group's earnings distribution can be consistently estimated using mean treatment effects and control group earnings data. This approach allows us to avoid making any parametric restrictions on the earnings distribution other than those implied by the null. We find substantial evidence that even relatively flexible subgroup-specific mean treatment effects specification are inadequate to generate synthetic treatment effect heterogeneity comparable to the observed heterogeneity. This conclusion holds even allowing mean treatment effects to vary according to key variables like race, education, and pre-reform earnings history, as well as by time since random assignment.

*Correspondence to Hoynes at UC Davis, Department of Economics, 1152 Social Sciences and Humanities Building, One Shields Avenue, Davis, CA 95616-8578, phone (530) 752-3226, fax (530) 752-9382, or hwhoynes@ucdavis.edu; Gelbach at gelbach@email.arizona.edu; or Bitler at mbitler@uci.edu. The data used in this paper are derived from data files made available to researchers by MDRC. The authors remain solely responsible for how the data have been used or interpreted. We are very grateful to MDRC for providing the public access to the experimental data used here. We would also like to thank David Brownstone, Julie Cullen, and Giuseppe Ragusa for helpful conversations, as well as participants of the SOLE meetings and the UCSD Applied Micro seminar.

1 Introduction

Welfare reform dramatically changed the system of cash support for low income families with children in the United States. By the end of 1998, all states were required to eliminate Aid to Families with Dependent Children (AFDC) and replace it with Temporary Assistance for Needy Families (TANF). This meant large changes in incentives facing welfare recipients including adding lifetime time limits for welfare receipt, as well as stringent work requirements and the threat of financial sanctions. Due to this policy change, and the concurrent changes in the Earned Income Tax Credit and the strong labor market of the 1990s, annual employment rates of less educated single mothers increased by fully 18 percentage points—from 64 percent in 1992 to 82 percent in 2000 (Eissa & Hoynes (2006)).[1]

An enormous literature has followed this important set of policy changes. Recent comprehensive reviews of the literature conclude that welfare reform led to a reduction in welfare caseloads and an increase in employment and earnings, with less consistent and less statistically significant findings for the the impacts on income, poverty, and child well-being (Grogger & Karoly (2005) and Blank (2002)).

There has been significant attention in this literature to exploring the extent of heterogeneity in the effects of welfare reform. One might expect to find heterogeneity, for example, because of differences in preferences, fixed costs of work, wage opportunities, or variation in the risk of being impacted by welfare.

In earlier work, we explored this heterogeneity by estimating the impact of welfare reform on the distribution of earnings, transfers, and income (Bitler et al. (2006*b*)). In particular, we estimated quantile treatment effects applied to a randomized experiment of an important welfare reform in Connecticut. The quantile treatment effects are estimated simply as the difference in outcomes for various quantiles of the treatment (welfare reform) group and the control (AFDC) group and capture estimates of the treatment across the distribution. We found evidence of substantial heterogeneity of Connecticut's reform: For example, the reform had no impact on the bottom of the earnings distribution, it increased earnings in the middle of the distribution, and, before time limits took effect, reduced earnings at the top of the distribution. Importantly, we argued that this

[1] Annual employment rates calculated from the March Current Population Survey, including a sample of single mothers between the ages of 19 and 44 with a high school education or less.

heterogeneity is consistent with predictions of static labor supply theory.

Quantile treatment effects offer one way to get at underlying heterogeneity; in the context of welfare reform a far more common approach is to estimate mean impacts for various demographic subgroups. For example, some studies such as Meyer & Rosenbaum (2001) find that the impacts of reform are greater for never-married women and women with young children. While much of the existing work that focuses on subgroups is motivated by a desire to uncover heterogeneity arising from differences in preferences or wage opportunities, in practice, choice of subgroups is often somewhat ad hoc and dictated in part by data constraints. Further, Grogger, Karoly & Klerman (2002) conclude that "the effects of reform do not generally appear to be concentrated among any particular group of recipients" (p. 231).

In this paper, we evaluate the ability of subgroup-specific, constant mean treatment impacts to generate the rich heterogeneity observed using the semi-parametric quantile treatment effects estimator we used in Bitler et al. (2006*b*). As with our earlier work, we use data from a randomized experiment in Connecticut. In that experiment, welfare recipients and applicants were randomly assigned to either Jobs First, Connecticut's welfare reform program, or the existing AFDC program.

We begin by presenting the mean treatment effects for the full sample and for subgroups. We construct subgroups based on the race, education, and marital status of the mother, the age of her youngest child, her number of children, and her welfare and employment history.[2] We compare synthetic distributions created from these mean impacts to the quantile treatment effects (QTE) estimated for the full sample and for various subgroups. To assess the importance of heterogeneity in these two estimation approaches, we test whether the mean treatment effects vary across subgroups, and whether the quantile treatment effects vary across quantiles. We then go on to formally test for the adequacy of mean impacts in capturing the heterogeneity that we find using the quantile treatment effects. In particular, we test whether the quantile treatment effects in the full sample are statistically distinguishable from synthetic results generated under the null hypothesis that reform has constant mean impacts within a variety of subgroups.

We have two key findings. First, substantial evidence of treatment effect heterogeneity in welfare reform's impact on earnings exists not only for the full sample, but also within virtually all subgroups we examine. Importantly, the nature of the heterogeneity seems to differ depending

[2]The vast majority of the participants in the experiment were women, thus we use "she" and "her."

on the woman's race, education and earnings history. Our analysis of mean treatment effects also shows evidence of heterogeneity—we can reject equality of mean impacts across many subgroups. Second, we resoundingly reject the null hypothesis that mean treatment effects by subgroup can explain the heterogeneity shown in the full-sample quantile treatment effects. This conclusion holds even when we define subgroups as three-way classifications between race, education, and earnings history.

The Jobs First experiment and data are ideal for this exercise. First, the use of experimental data means we can evaluate alternative estimators without concerns about identification, selection and the like. Second, we have rich administrative pre-treatment data on earnings and welfare history to supplement the usual demographic variables (race, ethnicity, education, age of youngest child, number of children, marital status) for forming subgroups. This rich dataset allows us to form subgroups that are not only more detailed than is possible in standard data sets but also more likely to characterize wage opportunities or incentives. Lastly, as is typical with experimental analyses of social programs, our experimental sample is drawn from the population of welfare recipients and applicants. This has the advantage of generating a more homogeneous sample than the typical non-experimental sample. Thus, if treatment effect heterogeneity cannot be explained with subgroup-specific mean treatment effects here, it is even less likely to be explicable in non-experimental settings.

Exploring heterogeneity is important insofar as it allows us to "get inside the black box" of welfare reform. Can we identify groups with larger or smaller responses to the reforms? If so, then this information could help improve targeting of certain features of the reform to the appropriate groups. Understanding heterogeneity in the effects of reform is important not just for the purposes of policy evaluation, but also because it allows us to test implications of economic models of behavior.

The paper provides an important contribution to both the welfare reform literature and the growing methodological literature on heterogeneous treatment effects. On the latter topic, there are many other applications where researchers rely on estimating means within subgroups (e.g., job training, anti-poverty programs like Mexico's Progresa, and social programs like Moving To Opportunity in the United States). Moreover, there is an extensive literature on treatment effect heterogeneity, including Imbens & Angrist (1994), Heckman & Vytlacil (2000), Heckman & Vytlacil

(1999) Hotz, Imbens & Klerman (Forthcoming), Abadie (2002), Crump, Hotz, Imbens & Mitnik (2006*a*), Crump, Hotz, Imbens & Mitnik (2006*b*) (surveyed in Angrist (2004)). 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), and Chernozhukov & Hansen (2005). Other papers that consider the distribution of treatment effects include Athey & Imbens (2006) and Wu & Perloff (2006).

However, to our knowledge, ours is the first paper to provide a testable null hypothesis under which all heterogeneity is—nonparametrically—driven by treatment effects that are constant within, but vary across, identifiable subgroups. This is an important innovation because so many applied researchers (reasonably) use ad hoc subgroups to try and isolate sample members likely to respond to program or policy changes. If our finding—that this approach falls woefully short of capturing the actual form of heterogeneous treatment effects—holds more generally, then the subgroup-specific effect may need to be reconsidered. Estimating mean impacts may miss a lot, even across multiple subgroups, so researchers should consider using distributional estimators such as those used here.

The remainder of the paper is organized as follows. In section 2 we provide an overview of welfare reform, the Connecticut Jobs First program, and its theoretically predicted effects. We then discuss our data in section 3. We discuss the empirical methods in section 4 and present the results for the mean treatment effect and quantile treatment effects in section 5. We present our tests for the adequacy of the subgroup-specific treatment effects compared to the QTE in section 6 and the results of those tests in section 6.1. We conclude in section 7.

2 Welfare Reform, Jobs First & Predicted Labor Supply Effects

The current era of welfare reform consists of two periods of policy change. First, in the early- to mid-1990s about half of the states were granted waivers to reform their AFDC programs. Second, state experimentation led to passage of the 1996 Personal Responsibility and Work Opportunity Act (PRWORA). PRWORA eliminated AFDC and replaced it with Temporary Assistance for Needy Families (TANF), representing a dramatic federal reform.

In hindsight, probably the most important feature of the PRWORA reform is the introduction of time limits—which limit the number of years over a woman's lifetime that she could receive cash support. The other central features of TANF are work requirements, financial sanctions, and enhanced earnings disregards. These changes were designed to increase work and reduce welfare participation.[3]

A federal requirement faced by states seeking welfare waivers was that the state evaluate the policy changes, which some states did using random-assignment experiments. This requirement has led to a wealth of data for "waiver states" allowing for experimental analyses of welfare policy changes. Interestingly, evaluation was not required when states implemented their TANF programs.

In this project we analyze Connecticut's waiver program, called Jobs First. We have chosen Connecticut because its waiver program is among the most TANF-like of state waiver programs. Importantly, the Jobs First waiver contained each of the key elements found in TANF programs: Time limits, work requirements, financial sanctions, and enhanced earnings disregards. In contrast, few state waivers contained time limits of any kind. By using the experimental data available for Connecticut, we are able to avoid the pitfalls of non-experimental analyses of welfare reform (Blank (2002)).

In the Connecticut experiment, a random sample of welfare recipients (current recipients or new applicants) were randomized into either the Jobs First program or the existing AFDC program. The programs differ in many ways. First, Jobs First has a time limit of 21 months compared to no time limit in the AFDC program. Second, Jobs First has a very generous earnings disregard policy: every dollar of earnings below the federal poverty line is disregarded in benefit determination, leading to an implicit tax rate of 0% for all earnings up to the poverty line. In contrast, the implicit tax rate under AFDC was two thirds in the first four months on aid, and 100 percent after.[4] Furthermore, work requirements and financial sanctions were strengthened in the Jobs First program relative to AFDC. For more information on these and other features of Jobs First see the final report on the Jobs First evaluation (Bloom, Scrivener, Michalopoulos, Morris, Hendra, Adams-Ciardullo &

[3]Other changes adopted by some states include: Expanding eligibility for two-parent families, family caps (freezing benefits at the level associated with current family size), and imposing residency and schooling requirements for unmarried teen recipients. For a detailed discussion of these policy changes, see Blank & Haskins (2001) and Grogger & Karoly (2005).

[4]These implicit tax rates in AFDC applied to all earnings above a monthly disregard of $120 during a woman's first 12 months on aid, and $90 thereafter.

Walter (2002)) or our earlier paper (Bitler et al. (2006*b*)).

This combination of short time limits (in fact, the Job's First time limits are the shortest in the U.S. [Office of Family Assistance (2003)]) and generous disregards leads to large changes in the incentives to work for welfare recipients. As such, Jobs First provides the perfect setting to examine the heterogeneous impacts of welfare reform.

Labor supply theory has strong and heterogeneous predictions concerning welfare reforms like those in Jobs First. To make the discussion more concrete, consider Figure 1 which shows a stylized budget constraint under Jobs First (represented by AF) and AFDC (represented by AB). Further, suppose that a woman has been on aid for fewer than 21 months, so that the time limit does not yet bind.[5]

What we have in mind is to compare the outcome of a woman if she were assigned to AFDC to the counterfactual outcome for that same woman if she were assigned to Jobs First. At the time of randomization, women will largely by definition[6] be on cash support and, most likely, not working. However, after random assignment, the AFDC and Jobs First groups are tracked for three to four years. Over that time period, women in the AFDC group will leave welfare—at different rates for different women. In fact, we find that about half of women in the AFDC control group have left welfare within two years after random assignment, which is similar to the pattern of welfare dynamics in the literature (Bane & Ellwood (1994)). So if we consider the full experimental period, we may find women in the AFDC group at a range of labor supply choices such as points $\{A, C, D, E, H\}$ in Figure 1. We want to then compare labor supply outcomes for women arrayed along these choices to the counterfactual outcome they would be predicted to have had if they had instead been assigned to Jobs First.

Applying the static labor supply model, we assume that the woman can freely choose hours of work at her (assumed fixed) wage. A woman who would not work (i.e., would locate at point A) when assigned to AFDC will will either stay out of the labor force or locates at some point on AF when she is instead assigned to Jobs First. This outcome, of course, will depend on her preferences, her fixed costs of work, and her wage opportunities.[7]

[5]The remainder of this section follows Bitler et al. (2006*b*).

[6]Anyone from the recipient sample is by definition already receiving support at the time of random assignment. A very large majority of women in the applicant sample do wind up receiving aid, though some apparently do not qualify for aid.

[7]Assuming that offered wages do not vary with hours worked, predictions about hours worked map one-to-one

Next, suppose a woman works positive hours and receives welfare if assigned to AFDC (so that she locates at point C). If she is assigned to Jobs First, her hours will increase as long as the substitution effect dominates the income effect. Now imagine that at some time after assignment to AFDC, a woman ends up at a point like D, where she is earning above the AFDC break-even point but below the poverty line. Assignment to Jobs First would make this woman eligible for welfare and the outward shift in the budget line would be predicted to reduce her hours of work.

Finally, consider a woman who, given assignment to AFDC, eventually ends up at point E or H. At E, as long as leisure and consumption are normal goods, the woman is predicted to decrease her hours to qualify for the windfall payment. At H, assignment to Jobs First could lead to no change or a reduction in hours, depending on her preferences.

The set of points $\{A, C, D, E, H\}$ represent the (qualitatively) possible hours/earnings outcomes under AFDC assignment. Therefore, we can summarize the impacts of Jobs First as follows: At the bottom of the earnings distribution, the Jobs First effect will be zero; it will then be positive over some range; then it will become negative; and finally at the very top of earnings distribution it may again be zero.

It is useful to think about how these predictions will vary with differences in wage opportunities. Suppose we compare two hypothetical women with similar preferences and fixed costs of work, but where one woman has higher wage opportunities because, for example, she has more education or more extensive work experience. With a higher wage and similar preferences, the higher-wage woman will unambiguously be more likely to enter work from point A. In addition, the higher-wage woman will be more likely to experience a reduction in hours. Why? Importantly, the Jobs First phase-out point is the federal poverty line and thus (in income space) does not vary across the two women. So the higher-wage woman will reach the federal poverty line at a lower hours point than the lower-wage woman. When assigned to AFDC, then, the higher-wage woman will be more likely to locate at points like D or E where hours are predicted to decline.

imply to predictions about earnings. This fact is important since we observe earnings but not hours worked. In Bitler et al. (2006*b*) we discuss the possibility that "queuing" effects might cause women to reduce their reservation wages for working in order to secure employment before the time limit. However, we find little empirical evidence to support this theory. In other work analyzing Canada's Self-Sufficiency Program, or SSP, we do find evidence of a decline in wages at the top of the wage distribution (see Bitler, Gelbach & Hoynes (2006*a*)). However, SSP differs from Jobs First, most notably in that it provides a limited time period for experimental particpants to establish eligibility for the program's generous earnings subsidy. This feature is not present in Jobs First (or in any other waiver or TANF program).

In sum, simple labor supply theory yields important heterogeneous predictions for the impacts of welfare reform on labor supply and earnings. Some groups may remain out of the labor market with no increase in hours, either due to low labor market opportunities, high fixed costs of work or strong tastes for leisure over income. Other groups may experience increases in hours worked with entry into the labor market. Finally, some women who would otherwise exit welfare relatively quickly and work at a level above the AFDC break-even point may be driven to work fewer hours in the presence of Jobs First.

3 Data

The evaluation of the Connecticut Jobs First program was conducted by MDRC.[8] In this analysis, we use public-use data made available by MDRC on completion of an application process. The data include information on a total of 4,803 cases; 2,396 were assigned to Jobs First, with 2,407 assigned to AFDC. The sample includes both women who were assigned to the experiment when they applied to start a new spell on welfare (the "applicant" sample) and women who were already on aid when they were assigned to the experiment (the "recipient" sample). The experiment took place in the New Haven and Manchester welfare offices.

The public use data consist of administrative data on earnings, welfare receipt and welfare payments, and survey data on demographic variables. Data on quarterly earnings and monthly income from welfare and food stamps are available for most of the two years preceding program assignment as well as for at least 4 years after assignment.[9] In this paper, we use earnings data from only the first seven quarters after random assignment. We focus on this period because the time limit cannot bind for anyone in the sample during the first 21 months after random assignment; as we discuss above, labor supply predictions are cleanest before the time limit. Demographic data collected at experimental baseline include each woman's number of children, education, age, race, ethnicity, and marital status, all at the time of random assignment. Random assignment took place between January 1996 and February 1997. Our final sample has 4,773 cases–2,392 assigned to Jobs

[8] MDRC, formerly known as the Manpower Demonstration Research Corporation, identifies itself as "a nonprofit, nonpartisan social policy research organization with headquarters in New York City and a regional office in Oakland, California." MDRC has conducted many other social experiments in addition to Jobs First.

[9] For confidentiality purposes, MDRC rounded all earnings data. Earnings between \$1–\$99 were rounded to \$100, so that there are no false zeros. All other earnings amounts were rounded to the nearest \$100.

First and 2,381 assigned to AFDC–for which we observe earnings for the full 16-quarter follow-up period.[10]

Table 1 reports summary information concerning a number of baseline characteristics. As described in Bloom et al. (2002) and Bitler et al. (2006*b*), average values of several of these characteristics differ statistically by treatment assignment. In particular, the first two columns provide means for the Jobs First (column 1) and AFDC (column 2) groups. The third column reports the unadjusted difference in means across the program groups, with indicators as to when the difference is statistically significantly different from zero. The table shows that the Jobs First group is statistically significantly more likely than the AFDC group to have more than two children, be in the recipient sample (drawn from the current caseload of AFDC recipients), and has lower earnings for the period prior to random assignment. A standard test for joint significance of the 17 differences (including some missing indicators), however, leads to a χ^2 test statistic of 22.83 (p-value of 0.16), so we cannot reject that assignment was indeed random.

Despite our inability to reject random assignment, one might be concerned about the pre-treatment differences in earnings and welfare receipt. Mindful of this possibility, we deal with the unbalanced sample using inverse propensity score weighting, as in Bitler et al. (2006*b*).[11] We use a logit model to estimate the probability that person i is in the treatment group; we include as regressors the following pre-random assignment variables: Quarterly earnings in each of the 8 pre-assignment quarters, quarterly AFDC and quarterly Food Stamps payments in each of the 7 pre-assignment quarters, dummies indicating whether each of these variables is nonzero, and dummies indicating whether the woman was employed at all or on welfare at all in the year preceding random assignment. We also include dummies for being in the applicant sample, and race, marital status, education, number of children, and age of woman. Finally, we include dummies indicating whether education, number of children, or marital status is missing.

Denoting the estimated propensity score for person i as $\hat{p}_i$ and the treatment dummy as D_i,

[10]That is, in this version of the paper, we dropped 30 women from the sample because they are missing earnings data for the 16^{th} quarter after random assignment. Since we focus only on the first seven quarters after random assignment, we will add these women back to the sample in future versions.

[11]Firpo (2007) shows that this approach yields asymptotically consistent estimates of QTEs for continuous dependent variables. Because MDRC essentially rounds the earnings data we use to the nearest hundred dollars (see the appendix for more detail on the rounding procedure), our dependent variable is actually discrete. Gelbach (2005) shows that sample quantiles computed using inverse propensity score weighting are consistent for the population quantiles of the rounded earnings variable.

the estimated inverse-propensity score weight for person i is

$$\hat{\omega}_i \equiv \frac{D_i}{\hat{p}_i} + \frac{1 - D_i}{1 - \hat{p}_i}. \tag{1}$$

We use inverse-propensity score weights in all our estimators used below. We find that the weighting never changes the qualitative conclusions concerning the quantile treatment effects; it does, however, lead to some important changes in the mean treatment effects. Unweighted results are available upon request.

4 Average Treatment Effects and Quantile Treatment Effects

To introduce the alternatives for capturing heterogeneity, it is helpful to briefly introduce a model of causal effects. For the moment, ignore the need to adjust for propensity score differences. Let $D_i = 1$ 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 $D_i = d$. The fundamental evaluation problem is that for any i, at most one element of the pair $(Y_{it}(0), Y_{it}(1))$ can ever be observed: we cannot observe someone who is simultaneously treated and not treated. Evaluation methodology focuses on inferences concerning various features of the joint distribution of $(Y(0), Y(1))$. There is an enormous literature concerning this model, which is variously called the Roy Model, the Quandt Model, and the Rubin Causal Model, as well as the assumptions under which it is useful (see, for example, papers by Heckman et al. (1997) or Imbens & Angrist (1994) for further details and citations).

The treatment effect for person i in period t, is equal to the difference between her period-t outcome if treated and untreated: $\delta_{it} \equiv Y_{it}(1) - Y_{it}(0)$. We use δ_t to represent 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. Using overbars to denote sample means, random assignment allows us to estimate the average effect of the policy for period t consistently using the difference in sample mean outcomes: $\overline{\delta}_t \equiv \overline{Y}_t(1) - \overline{Y}_t(0)$. Likewise, $\overline{\delta} \equiv \overline{Y}(1) - \overline{Y}(0)$ is a consistent estimate of $E[\delta_{it}]$, where this expectation is taken over all i and t.

In the welfare reform literature, heterogeneity is most commonly introduced by estimating mean treatment effects for subgroups of the population. For example, subgroup g might consist of

women (in both the treatment and control groups) who share a certain race, education, or welfare history. Then, define $Y_{it}^{g(i)}(d)$ as the counterfactual outcome value in period t for person i who is a member of subgroup $g(i)$ 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: $\overline{\delta}_t(g) \equiv \overline{Y}_t^g(1) - \overline{Y}_t{}^g(0)$. Accounting for the unbalanced sample 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. 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 $\Delta_q = y_{q1} - y_{q0}$. We can account for inverse propensity score weighting by defining the empirical *cdf* as $\widehat{F}_d(y) \equiv \sum_{i:\, Y_i(d) \leq y} \hat{\omega}_i / \sum_i \hat{\omega}_i$ and then proceeding as before. The QTE for quantile q may be estimated very simply as the difference across treatment status in the two outcome quantiles. For instance, if we take the sample median for the treatment group and subtract from it the sample median for the control group, we have the QTE at the 0.5 quantile. Other quantile treatment effects are estimated analogously; we evaluate the distributions at all 99 centiles. As with mean treatment effects, we can estimate QTE for subgroups of the population by calculating quantiles within these subgroups and proceeding as above.

We make one final methodological note requiring quantile treatment effects. QTE capture heterogeneity in that they tell us how the distribution changes when we assign Jobs First treatment randomly. We wish to stress an important methodological distinction between the quantile treatment effects and quantiles or other features of the *treatment effect distribution*. Unlike quantile treatment effects, quantiles of the distribution of treatment effects cannot generally be written as features of the marginal distributions. Rather, they require more detailed knowledge of the joint distribution (i.e., further assumptions about it). Under some conditions, the distribution of treatment effects is recoverable from the quantile treatment effects. For example, if the treatment effect is equal for all observations, then the distribution of treatment effects is degenerate and is fully identified by the mean impact. Second, if women's ranks in the distributions are the same regard-

less of whether they are assigned to treatment or control group, i.e., there is rank preservation across treatment status, then the QTE at quantile q tells us the treatment effect for someone whose location is quantile q in the given distribution. Rank preservation is a strong assumption, however, and it will fail here if, for example, preferences for work do not map one-to-one with rank in the earnings distribution.[12]

5 Mean and Quantile Treatment Effects

In this section, we report mean and quantile treatment effects for the full sample, as well as for several subsamples. We also test whether mean impacts differ across subgroups and whether the set of QTE differ from the mean treatment effect. As noted above, we limit our analysis here to the pre-time limit period. Since we use the person-quarter as the unit of analysis, there are a total of $7 \times 4,773 = 33,411$ observations in our sample. Five of our subgroup classifications are commonly used in analyses of welfare programs. These subgroups are defined by educational attainment, by race, by age of youngest child, by number of children, and by marital history at the time of random assignment. We also construct two additional subgroups, based on welfare and on employment history. Because welfare and employment history data are rarely available in nonexperimental data sets, subgroups based on these variables are rarely used in the literature. Finally, we consider subgroups defined by cross-classifying some of the eight subgroup variables just described. For example, RACE-BY-ED....XXX....

5.1 Mean Treatment Effects

Table 2 reports estimated mean treatment effects. The first row, for the full sample, shows that Jobs First is associated with an increase in quarterly earnings of $80. This estimate is statistically insignificant, as the 95% confidence interval of $[-38, 122]$ includes 0. Moreover, the point estimate is small by comparison to the control group mean of $1,112.

The next three rows of Table 2 report estimated mean treatment effects separately for whites, blacks and Hispanics. The estimate of $228 for whites is both statistically significant and large relative to the control group mean. By contrast, estimates for blacks and Hispanics are statistically

[12] We plan to address empirical tests for rank preservation in future work.

insignificant and substantively small. The row below the Hispanics' treatment effect row reports an F statistic testing the null hypothesis that the mean treatment effects are the same across race; the test clearly rejects, so we can say with confidence that there is racial heterogeneity in the mean treatment effects. The next row reports an F statistic for testing the same null when we exclude observations for which race is either missing or different from white, black or Hispanic; there are 305 such observations.[13]

The next set of results reports mean treatment effects of $133 and $87 for highschool dropouts and women with either a high school diploma or GED (henceforth, "dropouts" and "high school graduates"). The former estimate is statistically significant, while the latter is not. Given that mean control group earnings for dropouts are less than half the mean for high school graduates, these estimates are substantively quite distinct. An F test clearly rejects equality of these estimates, though we do not reject when we use an F test based on the sample that does not include 284 women missing data on educational attainment. Results for women whose youngest child is aged 5 or younger show a statistically significant estimated mean treatment effect of $86. The estimate of $143 for those whose youngest child is aged 6 or older is also significantly different from zero. The reported F statistics show that we can reject equality of these estimates when we include 157 observations with missing data on youngest child's age, but not when we exclude these observations. Generally similar results appear when we define subgroups using the number of children in the case or marital history at the time of random assignment.

The final two sets of results concern subgroups defined using either AFDC or employment history at the time of random assignment. There are two AFDC-history subgroups: women who received any AFDC income in the quarter that occurred 7 quarters before random assignment, and those who did not. The employment-history subgroups are defined analogously, according to whether women had any earnings income in the quarter occurring 7 quarters before random assignment. Among women with AFDC income 7 quarters before random assignment, the estimated mean treatment effect was a statistically significant $109, which is moderately large relative to the control group mean. By contrast, women with no AFDC income 7 quarters before random assignment had a

[13]To conserve space, we do not report estimated mean treatment effects for women missing data on the variables we use to construct our subgroups. However, the first F statistic in each pair reported in Table 2 is based on estimates for both the reported subgroups and the set of women with missing data on the relevant subgroup variable. Moreover, in constructing the synthetic earnings variables we use below, we treat women with missing data as a separate category.

statistically insignificant and small estimated mean treatment effect of $46. The reported F test shows that these estimated mean treatment effects are not statistically different.[14] Classifying subgroups based on employment history, we see that the estimated mean earnings impact is $175 among those with no 7 quarters-lagged employment history and $126 among those with such a history. The reported F statistic shows that these estimates are clearly statistically significantly different from each other.

Overall, the estimated mean impacts show substantial and often statistically significant heterogeneity. Interestingly, there is no consistent pattern relating the magnitude of the treatment effects to, say, control group mean earnings. For example, some subgroups show larger impacts for more "advantaged" women (whites, ever married), but others show larger impacts for less advantaged women (less educated, less employment history).

5.2 Quantile Treatment Effects

We now turn to our results for the quantile treatment effects. We construct QTE for 98 centiles in graphical form.[15] As above, we use the person-quarter as the unit of analysis and analyze the 33,411 observations on quarterly earnings during the first seven quarters. Figure 2 reports QTE for the full sample.[16] The solid line plots the estimated QTE, the dotted lines plot upper and lower bounds for 95% pointwise confidence intervals,[17] the dashed (horizontal) line shows the estimated mean treatment effect, and the 0-line is provided for reference.

Heterogeneity in Jobs First's impact across the earnings distribution's quantiles is unmistakably significant, both statistically and substantively. Figure 2 shows that for quarterly earnings in the pre-time limit period, the QTE are zero below the median. This result occurs because quarterly

[14] There are no missing observations on the variables used to construct either AFDC or employment history.

[15] We computed QTE results at quantile 99 but omit them from the figures below because their variances are frequently large enough to distort the scale of the figures.

[16] The estimated QTE plotted in this figure are identical to those in Figure 3 of Bitler et al. (2006*b*).

[17] We construct confidence intervals using the percentile bootstrap based on 999 bootstrap replications. We use a block bootstrap algorithm, so that we randomly sample entire 7-quarter earnings profiles. This re-sampling scheme replicates any within-person dependence in the data. The 95% confidence interval limits are given as follows. First, let $\widehat{y}_q$ be the q^{th} real-data sample quantile. Second, let $\widehat{y}^*_{q,\alpha/2}$ and $\widehat{y}^*_{q,1-\alpha/2}$ be the $\alpha/2$ and $1-\alpha/2$ quantiles of the empirical bootstrap distribution for the q^{th} quantile. The lower limit of a 95% confidence interval is given by $\widehat{y}_q - (\widehat{y}^*_{q,1-\alpha/2} - \widehat{y}_q)$, while the upper limit is given by $\widehat{y}_q - (\widehat{y}^*_{q,\alpha/2} - \widehat{y}_q)$. With 999 replications and $\alpha = 0.05$, $\widehat{y}^*_{q,\alpha/2}$ will be given by the 25th smallest bootstrap QTE for the q^{th} quantile and $\widehat{y}^*_{q,1-\alpha/2}$ will be given by the 25th largest. This percentile method does not impose symmetry, and the estimated confidence interval limits frequently are not symmetric.

earnings are 0 for 48% of person-quarters in the Jobs First group over the first 7 quarters and 55% of corresponding AFDC group person-quarters. For quantiles 49–82, Jobs First group earnings are greater than control group earnings, yielding positive QTE estimates. Between quantiles 83–87, earnings are again equal (though non-zero). Finally, for quantiles 88–98, AFDC group earnings exceed Jobs First group earnings, yielding negative QTE estimates. For quantiles 89–96, these negative estimates are statistically significantly different from zero based on individually applied tests. This pattern is consistent with the predictions of labor supply theory discussed above (we argue this point in detail in Bitler et al. (2006*b*)). Finally, we note that the QTE range from a minimum of -$300 to a maximum of $500, a considerable range. To address the possibility that the estimated QTE in Figure 2 might simply be noisy estimates of a common treatment effect, which would necessarily equal the mean treatment effect, we applied Procedure 1 below to the entire sample; this test rejects equality of the QTE at the 5 percent level. Thus, as in Bitler et al. (2006*b*), we conclude that a single mean treatment effect cannot explain the heterogeneity in quantile treatment effects that we estimate in Figure 2.

Next, we estimate the QTE for our seven subgroups. Figure 3 replicates Figure 2 when we restrict consideration to whites. The pattern of estimated QTE mirror that for the full sample: QTE are zero at the bottom of the distribution, rise in the middle, and then fall in the upper part of the distribution (although the QTE do not become negative at the top of the earnings distribution). In Figure 4, we plot estimated QTE for whites, blacks, and Hispanics on the same graph. This figure shows some important differences across race/ethnicity. Throughout the distribution, whites' QTE exceed both blacks' and Hispanics'. Moreover, QTE for blacks and Hispanics exhibit negative QTE at the top of the distribution, as in the full sample.[18]

While Figure 4 shows some variation in the amplitude and location of the positive QTE across race and ethnicity, the shapes of the three QTE profiles do resemble one another considerably. This finding is unsurprising given that the share of the control group that belonging to each race/ethnicity group is fairly evenly distributed across the control and treatment groups. Figure 5 illustrates this fact by plotting the race/ethnicity shares at each centile of the pooled control group earnings distribution (because an analogous figure for the treatment group looks qualitatively similar, we

[18]To avoid clutter, we omit confidence intervals from Figure 4 and all remaining QTE plots. Figures that include 95% confidence intervals are available on request.

present only the control group figure). The figure shows the share of observations within each centile that are black, white, or Hispanic.[19] While these shares do vary, they are fairly stably distributed across the control group distribution. Intuitively, if the members of a particular subgroup are spread evenly across the centiles of the treatment and control groups, then it seems unlikely that stratifying along these subgroups will explain much of the heterogeneity in observed quantile treatment effects.

Figure 6 plots QTE for dropouts and high school graduates. Each subgroup's QTE profile shows substantial variation across quantiles, suggesting mean impacts are inadequate to explain the QTE. The figure also shows that high school graduates, but not dropouts, exhibit negative QTE. This finding is in line with the expectataion that more educated women are more likely to locate at points like E or H in Figure 1 under AFDC assignment. Figure 7 repeats the group-shares analysis of Figure 5, with groups now defined by dropout and high school graduate status. We see that the dropout share falls, and the high school graduate share rises, as we move up the pooled control group earnings distribution. This pattern suggests that there may be more scope for mean impacts by education subgroups to capture some of the heterogeneity in the QTE. That said, the considerable within-subgroup heterogeneity evinced in Figure 6 makes it unlikely that variation in subgroup-specific mean treatment effects can explain the QTE heterogeneity observed in the pooled sample.

Figures 8 and 9 show that, in contrast to the results for race and education, there is little difference in the QTE by presence of child aged 5 or younger or by marital history. However, we do find substantial cross-group heterogeneity in estimated QTE based on welfare and earnings history. QTE estimates in Figure 10 show that women with no welfare income seven quarters prior to random assignment have both (i) a smaller range of positive QTE and (ii) everywhere smaller magnitudes for the earnings impacts, by comparison to women who had welfare income seven quarters prior to random assignment. Differences based on employment seven quarters before random assignment, reported in Figure 11, are striking. Among women with no employment income seven quarters prior to random assignment, the estimated QTE are zero for more than the bottom half of the earnings distribution, with estimated effects being positive higher in the earnings distribution. For women with employment income seven quarters before random assignment, the estimated QTE are zero only for the first 30 quantiles, are positive and small over the next 35 or so quantiles, and

[19] We report group shares only for quantiles 45–98, since earnings are zero for each group at all quantiles below 45.

are negative for the top third of the earnings distribution. As with more educated women, it is reasonable to think that women with an employment history are more likely to locate at points like E and H of Figure 1 under AFDC assignment. Thus we regard the substantial range of negative QTE for these women as additional evidence in favor of the usefulness of labor supply theory.

6 Testing for Adequacy of Mean Treatment Effects

Figures 2–11 document the fact that quantile treatment effects vary substantially across the distribution of earnings, both in the pooled sample and within subgroups commonly thought to be relevant for welfare reform. We now turn to the question of whether cross-subgroup heterogeneity in mean treatment effects, e.g., as documented in Table 2, can explain the heterogeneity in quantile treatment effects documented in these figures.

As we will see, systematically addressing this question requires finding a way to compare observed (nonparametric) earnings distributions with counterfactual distributions that result when a (parametric) null hypothesis is true. In the remainder of this section, we show how to construct a class of distributions that obtain under such null hypotheses. We also discuss bootstrap-based tests for the equivalence of these null distributions to the observed earnings distribution. We can state our simplest null hypothesis as

$$H_0: \; Y_{it}^{g(i)}(1) = Y_{it}^{g(i)}(0) + \overline{\delta}(g(i)) \text{ for all } i, \tag{2}$$

where we recall that $Y_{it}^{g(i)}(d)$ is person i's realized outcome t periods since random assignment given that she is a member of group $g(i)$ and has treatment assignment $D_i = d$, while $\overline{\delta}(g)$ is the effect of treatment on women in group g. (We will drop the i superscript when there is no potential for confusion.) Notice that this null implies that 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

$$H_0: \; Y_{it}^{g(i)}(1) = Y_{it}^{g(i)}(0) + \overline{\delta}_t(g(i)) \text{ for all } i, \tag{3}$$

Notice that, while this null allows the treatment effect to vary across both time period and

across subgroup, it does not allow treatment effects to vary within subgroup-by-time cells. Thus, for example, this null hypothesis assigns earnings $\overline{\delta}_t(g)$ to a woman whose potential outcome is 0 when assigned to AFDC. We have seen that basic theory predicts that some women will have zero earnings under both assignments, and both sample distributions have a substantial share of women with zero earnings. It would therefore be unsurprising to reject the null simply because (i) there is a nonzero mean treatment effect while (ii) both the treatment group and control group earnings distributions exhibit positive mass at zero earnings. 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 more realistic null hypothesis that allows nonzero, constant mean treatment effects given positive earnings even as potential outcomes can be zero under both assignments. To allow this possibility, we consider a third null hypothesis, which defines a probability distribution for $Y_{it}^{g(i)}(1)$ in the event that $Y_{it}^{g(i)}(0) = 0$. Note that the share of group-g women whose potential time-t earnings would be zero under treatment group assignment can be written as $p_{1gt} \equiv F_{1gt}(0)$, where F_{1g} is the group-g earnings distribution given treatment group assignment. Let the conditional distribution of control group earnings among group-g women with positive earnings at time t be $F_{0gt}(\cdot|y > 0)$. Finally, redefine $\overline{\delta}_t(g(i))$ as the time-t mean treatment effect among group-g women with positive earnings, i.e., the difference in the means of $F_{1gt}(\cdot|y > 0)$ and $F_{0gt}(\cdot|y > 0)$. Our most sophisticated null hypothesis is

$$H_0 : \; Y_{it}^{g(i)}(1) = \begin{cases} Y_{it}^{g(i)}(0) + \overline{\delta}_t(g(i)), & Y_{it}^{g(i)}(0) > 0 \\ X(gt), & \text{otherwise} \end{cases} \text{ for all } i, \tag{4}$$

where the random variable $X(gt)$ equals 0 with probability p_{1gt} and, with probability $(1 - p_{1gt})$, equals $\overline{\delta}_t(g)$ plus a random draw from $F_{0gt}(\cdot|y > 0)$. Notice that under this null hypothesis, the null value of $Y_{it}^{g}(i)(1)$ always equals its actual population value. Moreover, the conditional mean of treatment group earnings under the null also always equals its population value. These two facts imply equality of the null and population values of the unconditional mean of treatment group earnings. Thus by construction, the null hypothesis in (4) cannot be rejected due to differences across program assignment in either the share of zeros or the conditional mean treatment effects

given positive earnings. That is, any rejection of (4) must arise for distributional reasons unrelated to either $\{p_{1gt}, \overline{\delta}_t(g)\}_{g,t}$. This fact is obviously a feature of this null hypothesis.

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 notation. Let

$$Y_{it} \equiv Y_{it}^{g(i)}(1) D_i + Y_{it}^{g(i)}(0) \left[1 - D_i\right] \tag{5}$$

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 $q \in (0, 1)$, we define

$$y_{qd} \equiv \inf_y \left\{y : \ F_d(y_{qd}) \geq q\right\}. \tag{6}$$

where F_d is the *cdf* of outcome variable Y given treatment status $d \in \{0, 1\}$. Notice that these quantiles are defined relative to the earnings distribution for the entire population—that is, they do not specify any particular subgroup g. For notational simplicity, we will adapt the convention that when $q \in (1, 100)$, the quantile of interest is actually $q/100$. Thus, for example, we can refer to the median as either y_{50d} or $y_{0.5d}$. We will define $\widehat{F}_d$ to be the sample analogue of F_d, i.e., the empirical distribution function for observations actually observed in treatment status d. Next, we define

$$\widetilde{Y}_{0it} = Y_{it}^{g(i)}(0) \tag{7}$$

$$\widetilde{Y}_{1it} = \widetilde{Y}_{0it} + \delta(g(i)). \tag{8}$$

Thus $\widetilde{Y}_{0it}$ is the outcome value that person i would have in period t if assigned to the control group. Obviously, this definition entails no assumption about treatment effects, since it depends only on actual potential earnings given assignment $D_i = 0$. By contrast, the definition of $\widetilde{Y}_{1it}$ tells us the value person i's outcome in period t will take if the effect of treatment for this person is the time-constant $\overline{\delta}(g(i))$ for each time t. If instead we wish to allow the treatment effect to vary with t, then we define

$$\widetilde{Y}_{1it} = \widetilde{Y}_{0it} + \delta_t(g(i)). \tag{9}$$

Finally, if we wish to impose equal shares of zero earnings across program assignment, as in (4), we define

$$\widetilde{Y}_{1it} = \begin{cases} Y_{it}^{g(i)}(0) + \delta_t(g(i)), & Y_{it}^{g(i)}(0) > 0 \\ X(gt), & \text{otherwise} \end{cases} \tag{10}$$

where $X(gt)$ is defined above. Note that (8), (9), and (10) will hold for all i and t if and only if the null hypotheses (2), (3) or (4) are correct. Thus, we can rewrite each null hypothesis as

$$H_0 : \ \widetilde{Y}_{1it} = Y_{1it}, \tag{11}$$

where $\widetilde{Y}_{1it}$ is defined appropriately to the relevant null by either (8), (9) or (10). Our inferential procedure must therefore be able to construct consistent estimates of functionals of the distribution of $\widetilde{Y}_{1it}$ and distinguish these estimates from estimates that are consistent even when the null is false. It will be helpful to have notation for the true quantiles of the distributions of $\widetilde{Y}_{0it}$ and $\widetilde{Y}_{1it}$. Let $\widetilde{F}_d$ be the *cdf* of outcome $\widetilde{Y}_{dit}$ for $d \in \{0, 1\}$. Observe that since $\widetilde{Y}_{0it} = Y_{0it}$ by definition, $\widetilde{F}_0 = F_0$ is always true. When the null hypothesis is false, though, $\widetilde{F}_1$ will differ from F_1. We denote the relevant quantiles as

$$\widetilde{y}_{qd} \equiv \inf_y \left\{ y : \ \widetilde{F}_d(\widetilde{y}_{qd}) \geq q \right\}. \tag{12}$$

Under the null hypothesis, all quantiles of $\widetilde{F}_1$ 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:

$$H_0 : \ \widetilde{y}_{q1} = y_{q1} \text{ for all } q \in (0, 1). \tag{13}$$

Let $\widetilde{\Delta}_q \equiv \widetilde{y}_{q1} - y_{q0}$ be the QTE at quantile q when the null hypothesis 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 is

$$H_0 : \ \widetilde{\Delta}_q = \Delta_q \text{ for all } q \in (0, 1). \tag{14}$$

For given q, we can estimate both $\widetilde{\Delta}_q$ and Δ_q consistently using sample quantiles of the distributions of $\widetilde{Y}_{1it}$ and Y_{1it}. The following procedure shows how to estimate these sample quantiles and test the null hypotheses implied by our three definitions of $\widetilde{Y}_{1it}$.

Procedure 1 (Testing H_0 from (14)).

Let Q be some number 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 $q = 1, 2, \ldots, Q$ (the method applies more generally). The following procedure provides a consistent test of the null hypothesis in (2):

1. Calculate sample quantiles of the control and treatment group observations for $q = 1, 2, \ldots, Q$ (in practice, we use $Q \in \{97, 98, 99\}$). We denote the set of sample quantiles for assignment status $d \in \{0, 1\}$ as $\{\widehat{y}_{q0}, \widehat{y}_{q1}\}_{q=1}^{Q}$. Defining $\widehat{F}_d$ as the empirical distribution function of observations with $D_i = d$, these sample quantiles are defined implicitly as

$$\widehat{y}_{qd} \equiv \inf_y \left\{ y : \widehat{F}_d(\widehat{y}_{qd}) \geq q \right\}. \tag{15}$$

2. Estimate the mean treatment effect for each subgroup g and period t. When we constrain the mean treatment effects to be constant across all t, as in (8), we use

$$\overline{\delta}(g) \equiv \overline{Y}_1^g - \overline{Y}_0^g \tag{16}$$

When we allow treatment effects to vary with time since random assignment, as in (9), we use

$$\overline{\delta}_t(g) \equiv \overline{Y}_{1t}^g - \overline{Y}_{0t}^g. \tag{17}$$

Finally, when we define $\delta_t(g)$ as the conditional treatment effect given positive earnings, as in (10), we use

$$\overline{\delta}_t(g) \equiv \frac{\overline{Y}_{1t}^g}{1 - \widehat{p}_{1gt}} - \frac{\overline{Y}_{0t}^g}{1 - \widehat{p}_{0gt}}, \tag{18}$$

where $\widehat{p}_{1gt}$ and $\widehat{p}_{0gt}$ are the estimated share of treatment and control group observations with zero earnings.

3. Construct estimates of the quantiles of F_1 that are correct only under the null, i.e., construct estimates of $\widetilde{F}_1$. To do so, first estimate $\widetilde{Y}_{1it}$ among control group observations, i.e., only

those with $D_i = 0$, using the appropriate choice of

$$\widehat{\widetilde{Y}}_{1it} \equiv Y_{it} + \widehat{\delta}\left(g(i)\right). \tag{19}$$

$$\widehat{\widetilde{Y}}_{1it} \equiv Y_{it} + \widehat{\delta}_t\left(g(i)\right). \tag{20}$$

$$\widehat{\widetilde{Y}}_{1it} \equiv \begin{cases} Y_{it}^{g(i)}(0) + \widehat{\delta}_t(g(i)), & Y_{it}^{g(i)}(0) > 0 \\ \widehat{X}(gt), & \text{otherwise} \end{cases} \tag{21}$$

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. To do so, we let $\widehat{\widetilde{Y}}_{1it} = 0$ whenever $Y_{it} = 0$ and $Y_{it} + \widehat{\delta}_t(g(i))$ whenever $Y_{it} > 0$. Recalling that the inverse propensity score weight for observations i defined above is $\hat{\omega}_i$, we multiply this weight by $D_i + \hat{\rho}_{it}(1 - D_i)$, where

$$\hat{\rho}_{it} \equiv \frac{\widehat{p}_{1gt}}{\widehat{p}_{0gt}} 1(Y_{it} = 0) + \frac{1 - \widehat{p}_{1gt}}{1 - \widehat{p}_{0gt}}[1 - 1(Y_{it} = 0)], \tag{22}$$

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. Thus it effects a consistent estimate of $\widetilde{Y}_{1it}$ as defined in (10). We note that the various estimates of $\widetilde{Y}_{1it}$ are consistent regardless of the null's correctness, since the null concerns the relationship between $\widetilde{Y}_{1it}$ and Y_{1it}, whereas consistency of $\widehat{\widetilde{Y}}_{1it}$ for $\widetilde{Y}_{1it}$ follows simply from the fact that $\widehat{\delta}_t(g)$ is consistent for $\delta_t(g)$ given random assignment. With these estimates in hand, we estimate $\widetilde{F}_1$ using the empirical distribution function, $\widehat{\widetilde{F}}_1$, of $\widehat{\widetilde{Y}}_{1it}$ among those in the control group. Finally, we estimate the quantiles of $\widetilde{F}_1$ by calculating the relevant quantiles of $\widehat{\widetilde{F}}_1$, using the values of $\widehat{\widetilde{Y}}_{1it}$ we constructed from (19). These sample quantiles are defined implicitly as follows:

$$\widehat{\widetilde{y}}_{qd} \equiv \inf_y \left\{ y : \widehat{\widetilde{F}}_d(\widehat{\widetilde{y}}_{qd}) \geq q \right\}. \tag{23}$$

4. Calculate the test statistic

$$\widehat{S} \equiv \sqrt{n} \left(\sup_{q=1 \text{ to } Q} \left\{ \frac{\left| \widehat{y}_{q1} - \widehat{\widetilde{y}}_{q1} \right|}{(\widehat{V}(q))^{1/2}} \right\} \right), \tag{24}$$

where n is the overall sample size and $\widehat{V}(q)$ is a consistent estimate of the variance of the discrepancy term $\widehat{y}_{q1} - \widehat{\widetilde{y}}_{q1}$. The discrepancy terms themselves are the differences between the always-consistent and consistent-only-under-the-null estimates of the quantiles. Chernozhukov & Fernandez-Val (2005, henceforth, CF) establishes the distribution of the supremum of the set of discrepancies, i.e., the statistic $\widehat{S}$. As CF discuss, dependence in the data-generating process causes this distribution to have non-standard properties, complicat-

ing inference using standard methods. However, CF show that a bootstrap procedure provides a basis for consistent inference on the Kolmogorov-Smirnov-like statistic $\widehat{S}$. We characterize this bootstrap procedure in the next step, and we discuss calculation of the estimated variance term $\widehat{V}(q)$ below.[20]

5. Do the following B times (where B is the number of bootstrap replications):

   (a) Re-sample the data in a manner consistent with the data generating process using the nonparametric block bootstrap. That is, we 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.

   (b) Repeat steps 1-3 using the re-sampled data. We use a superscript to indicate that the estimated sample quantile $\widehat{y}^b_{q1}$ or $\widehat{\widetilde{y}}^b_{q1}$ is based on the b^{th} bootstrap re-sample rather than the real data.

   (c) Calculate the bootstrap estimate of the statistic $\widehat{S}$, denoted

$$\widehat{S}^b \equiv \sqrt{n}\left(\sup_{q=1 \text{ to } Q}\left\{\frac{\left|\widehat{y}^b_{q1} - \widehat{\widetilde{y}}^b_{q1} - \left(\widehat{y}_{q1} - \widehat{\widetilde{y}}_{q1}\right)\right|}{(\widehat{V}(q))^{1/2}}\right\}\right). \tag{25}$$

   This statistic differs in form from $\widehat{S}$ in a key way: For each quantile, we create the bootstrap supremand by subtracting the real-data discrepancy term $\widehat{y}_{q1} - \widehat{\widetilde{y}}_{q1}$ from the corresponding bootstrap discrepancy term. This step is what allows Chernozhukov & Fernandez-Val's (2005) method to overcome the noncentrality problem alluded to above. Heuristically, the relevant noncentrality term can be consistently estimated using the real-data sample's estimate of the discrepancy $\widehat{y}_{q1} - \widehat{\widetilde{y}}_{q1}$. 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 properties under the null (and local alternatives).

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

$$s_{1-\alpha} \equiv \inf_s \{s : G(s_{1-\alpha}) \geq 1 - \alpha\}. \tag{26}$$

[20] As applied to our context, CF's assumptions require certain asymptotic normality properties for the sample quantiles. Given the discrete nature of our earnings data, our use of CF'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 yet proved that the bootstrap can also be used to estimate the null distribution of the statistic $\widehat{S}$.

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

$$\widehat{G}(s) \equiv \frac{1}{B+1} \sum_{b=1}^{B} 1\left(\widehat{S}^b \leq s\right). \tag{27}$$

The critical value for a level-α test is thus the smallest s such that $\widehat{G}(s) \geq 1-\alpha$, which we can write

$$\widehat{s}_{1-\alpha} \equiv \inf_s \left\{ s : \widehat{G}(\widehat{s}_{1-\alpha}) \geq 1-\alpha \right\}. \tag{28}$$

7. We reject the null hypothesis as stated in (13) as follows:

$$\text{Reject } H_0 \text{ iff } \widehat{S} \geq \widehat{s}_{1-\alpha}. \tag{29}$$

For instance, if we are interested in a level-0.05 test, with $B = 999$ we would find the 50^{th} largest bootstrap estimate of the test statistic (since $(1-0.05) \times (999+1) = 950$, and the 950^{th} smallest estimate is the 50^{th} largest when $B = 999$) and then reject the null hypothesis if and only if the real-data test statistic $\widehat{S}$ exceeded that value.

□

The only remaining task is to calculate the estimated variances given by $\widehat{V}(q)$. 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

$$\widehat{r}_q \equiv \widehat{y}_{q1} - \widehat{\widetilde{y}}_{q1}. \tag{30}$$

Similarly, define the iteration-b re-sampled estimate of this discrepancy term for the q^{th} quantile as

$$\widehat{r}_q^b \equiv \widehat{y}_{q1}^b - \widehat{\widetilde{y}}_{q1}^b. \tag{31}$$

We can estimate the variance of each discrepancy term $\widehat{r}_q$ with the sample variance of the re-sampled estimates of this vector. That is, for each q, we calculate

$$\widehat{V}(q) \equiv \frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{r}_q^b - \overline{r}_q\right)^2 \tag{32}$$

where $\widehat{r}_q^b$ is the bootstrap analogue of $\widehat{r}_q$ calculated using the b^{th} re-sample and $\overline{r}_q$ is the bootstrap sample mean of this statistic. We can then use this estimate of $\widehat{V}(q)$ in steps 4 and 5.[21]

6.1 Empirical evidence: Synthetic and Sample QTE

Before proceeding to the test statistics, it is useful to illustrate the difference between the sample QTE and the estimated synthetic QTE based on $\widetilde{Y}_{1it}$. We begin by constructing synthetic QTE using mean treatment effects by race/ethnicity, based on (19). In this and all subsequent figures, the sample QTE are calculated using the pooled sample, so that these estimates repeat those plotted in Figure 2.

In panel (a) of Figure 12, we allow treatment effects to vary only by race, constraining the within-race mean treatment effect to be the same for all time periods; thus this panel corresponds to the definition of $\widetilde{Y}_{1it}$ provided in (8). In panel (b), we allow treatment effects to vary across race and time period, as in (9). For both of these panels, the figure shows that the synthetic QTE do a very poor job of replicating the actual QTE. Not only are the synthetic QTE negative at the the bottom of the earnings distribution, but they also fail to replicate the large range of quantiles over which the QTE are zero due to the fact that about half the person-quarters in each program group exhibit no earnings. Moreover, the synthetic QTE in panels (a) and (b) fail to achieve as great a maximum QTE as the sample results, and neither synthetic specification exhibits negative QTE at the top of the distribution.

We turn now to panel (c), which reports actual estimated QTE together with estimated synthetic QTE based on the definition of $\widetilde{Y}_{1it}$ in (9). Recall that this specification of synthetic treatment earnings imposes the constraint that the synthetic and true earnings distributions have the same share of zeros. Not surprisingly, then, the panel (c) synthetic QTE are essentially the same as the sample QTE over the bottom half of the distributions.[22] Over quantiles 50-80 or so, the panel (c) synthetic QTE do a reasonably good job of replicating the shape of the sample QTE, though they fail to achieve the same amplitude as the sample QTE profile. However, the synthetic QTE

[21] Notice 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.

[22] The negative synthetic QTE at the very bottom of the distribution occur because the mean treatment effect is sufficiently negative for some time period(s) so that some women whose actual earnings are small and positive have negative synthetic earnings.

clearly fail to replicate the negative QTE at the top of the earnings distribution. As we have seen, this feature of the sample QTE is clearly predicted by labor supply theory. Thus the failure of the panel (c) synthetic QTE to replicate this feature is, in our view, a substantial mark against even the most flexible null hypothesis we consider.

Figures 13–17 report analogous results for the synthetic QTE for our subgroups defined by educational attainment, youngest child's age, marital history at the time of RA, receipt of AFDC 7 quarters before RA, and employment 7 quarters before RA. Results in these figures are qualitatively identical to those for the results for subgroups defined by race/ethnicity. Notably, the panel (a) and panel (b) results that do not constrain the share of zeros to be the same for synthetic as for population QTE fail entirely to replicate the shape of the sample QTE profile. The panel (c) results do much better in all cases, though they never do replicate the amplitude of the sample QTE or the negative effects at the top of the distribution. Finally, Figure 18 reports synthetic QTE for subgroups defined by race/ethnicity and educational attainment categories. Including missing values as a category, this classification involves 9 subgroup categories. Allowing the added flexibility makes little difference in any of the three panels.

These figures suggest that constant within-group treatment effects do a poor job of capturing the actual heterogeneity in treatment effects. Of course, the results in the figures are point estimates, and no systematic inferences can be drawn without some notion of the sampling variability of these point estimates. Thus in table 3, we report the values of formal test statistics and bootstrapped critical values for level-0.05 tests based on the panel (a) specifications. In each case, we easily reject the null hypothesis that the synthetic distribution equals the true treatment group distribution. We stress that we have not yet calculated test results for the more flexible null hypotheses underlying the panel (b) and panel (c) specifications; we plan to do this in our next draft.

Our preliminary results concerning the null hypotheses (2) and (3), corresponding to panels (a) and (b) of the figures, provide very strong evidence against the idea that constant within-group mean treatment effects are sufficient to characterize the heterogeneity in treatment effects that we estimate using QTE. Even the synthetic estimates based on the panel (c) null hypothesis (4) fail to replicate the key feature of negative earnings QTE at the top of the distribution. Given that this null itself allows an important form of treatment effect heterogeneity within subgroup-time cells, we believe that the panel (c) results also cast doubt on the plausibility of constant within-group

treatment effects to adequately evaluate key theoretical predictions.

7 Conclusion

The welfare reforms of the mid-1990s in the U.S. radically reformed the cash assistance system, moving from a system with strong work disincentives and no time limits to a time limited system aimed at encouraging work. A vast literature evaluates these reforms, concluding that there is considerable heterogeneity in some effects of these reforms. In previous work, we have explored this heterogeneity in the effect of a reform in Connecticut using quantile treatment effects (QTE) and experimental data (Bitler et al. (2006*b*)), finding evidence of considerable heterogeneity in the effects of this TANF-like reform on earnings.

In this paper, we explore the extent to which this heterogeneity can be explained by heterogeneity across subgroups (defined by demographics, welfare use, or earnings histories). We conclude that the heterogeneity uncovered in our earlier work simply cannot be explained by mean treatment effects that are constant within subgroups—even when we define subgroups richly in terms of race, education, and earnings history, and even when we allow the null hypothesis to treat zeros differently from positive earnings amounts. This finding provides further important evidence that means miss a lot. Distributional measures like QTE seem indispensable for measuring heterogeneity in the treatment effects of key reforms.

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.

Bane, M. J. & Ellwood, D. T. (1994), *Welfare Realities: From rhetoric to reform*, Harvard University Press, Cambridge and London.

Bitler, M. P., Gelbach, J. B. & Hoynes, H. W. (2006*a*), Distributional impacts of the Self-Sufficiency Project. Typescript.

Bitler, M. P., Gelbach, J. B. & Hoynes, H. W. (2006*b*), 'What mean impacts miss: Distributional effects of welfare reform experiments', *American Economic Review* **96**(4), 988–1012.

Blank, R. M. (2002), 'Evaluating welfare reform in the United States', *Journal of Economic Literature* **40**(4), 1105–1166.

Blank, R. M. & Haskins, R., eds (2001), *The New World of Welfare*, Brookings Institution Press, Washington, DC.

Bloom, D., Scrivener, S., Michalopoulos, C., Morris, P., Hendra, R., Adams-Ciardullo, D. & Walter, J. (2002), *Jobs First: Final Report on Connecticut's Welfare Reform Initiative*, Manpower Demonstration Research Corporation, New York, NY.

Chernozhukov, V. & Fernandez-Val, I. (2005), 'Subsampling inference on quantile regression processes', *Sankya: 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. (2006*a*), Moving the goalposts: Addressing limited overlap in estimation of average treatment effects by changing the estimand, Working paper. Typescript, UC Berkeley.

Crump, R., Hotz, V. J., Imbens, G. & Mitnik, O. (2006*b*), Nonparametric tests for treatment effect heterogeneity, Working Paper DP2091, IZA.

Eissa, N. & Hoynes, H. W. (2006), Behavioral responses to taxes: Lessons from the EITC and labor supply, *in* J. Poterba, ed., 'Tax Policy and the Economy', Vol. 20, MIT Press, Cambridge, MA, pp. 74–110.

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`.

Grogger, J. & Karoly, L. A. (2005), *Welfare Reform: Effects of a Decade of Change*, Harvard University Press, Cambridge, MA.

Grogger, J., Karoly, L. A. & Klerman, J. A. (2002), Consequences of welfare reform: A research synthesis, Working Paper DRU-2676-DHHS, RAND.

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. (2000), 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. (Forthcoming), 'Evaluating the differential effects of alternative welfare-to-work training components: A re-analysis of the California GAIN program', *Journal of Labor Economics* .

Imbens, G. W. & Angrist, J. D. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467 – 75.

Meyer, B. D. & Rosenbaum, D. (2001), 'Welfare, the earned income tax credit, and the laor supply of single mothers', *Quarterly Journal of Economics* **116**(3), 1063–1114.

Office of Family Assistance (2003), Temporary Assistance for Needy Families (TANF) fifth annual report to Congress, Working paper. http://www.acf.hhs.gov/programs/ofa/annualreport5/index.htm.

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

Figure 1: Stylized Connecticut budget constraint under AFDC and Jobs First

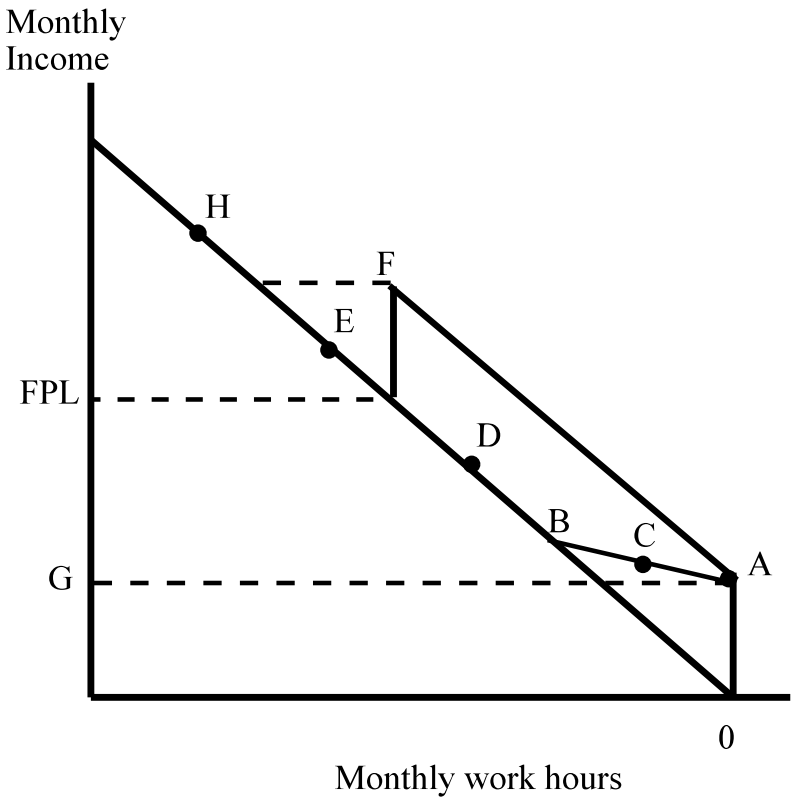
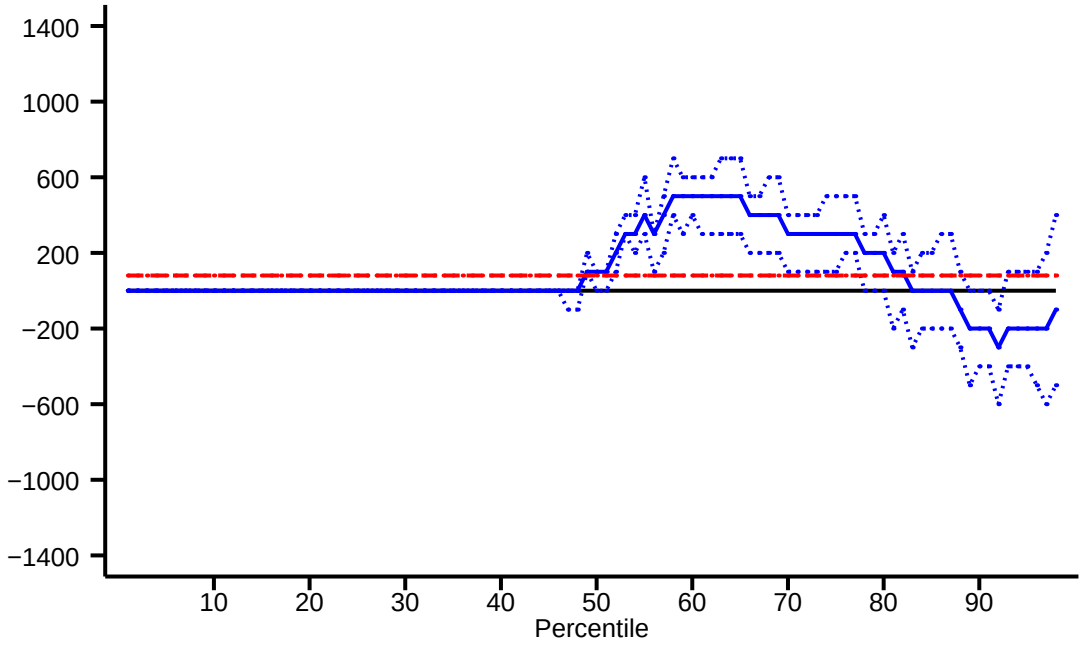
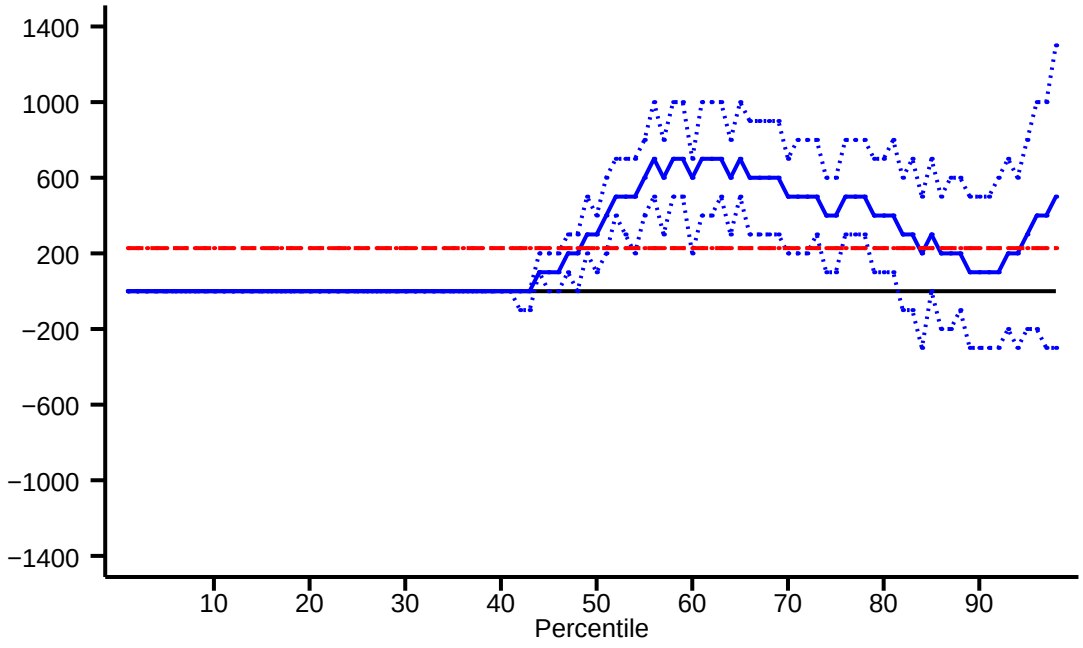


Figure 2: Quantile treatment effects on the distribution of earnings, quarters 1–7, all observations



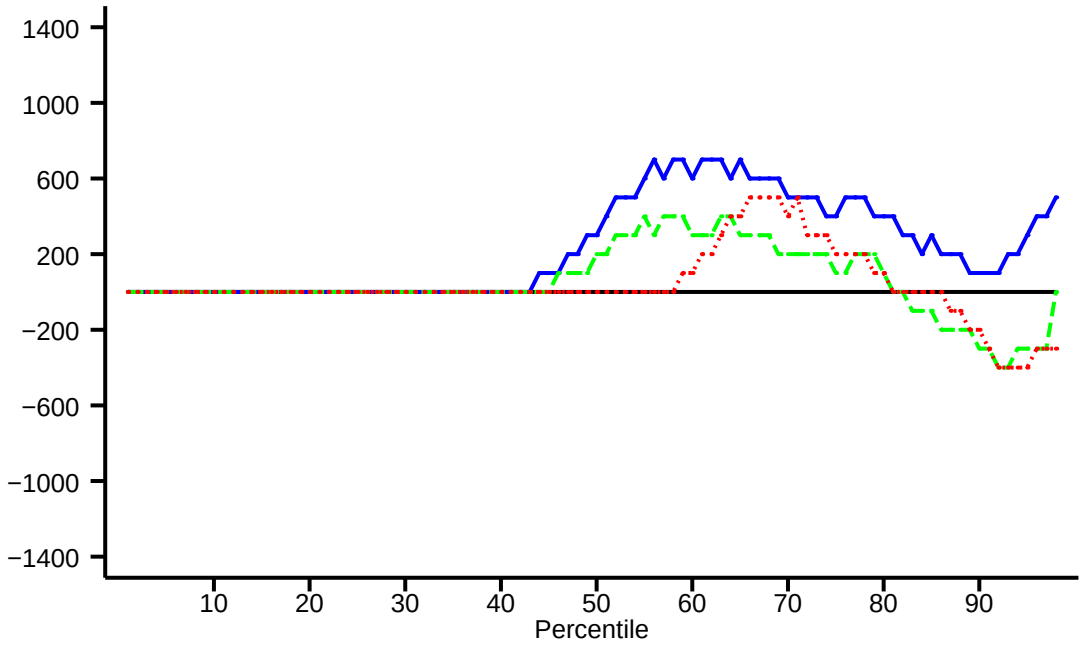
Notes: Solid line is QTE estimate, and dashed line is mean difference, and dotted line is 95% confidence interval.

Figure 3: Quantile treatment effects on the distribution of earnings, quarters 1–7, case head is white



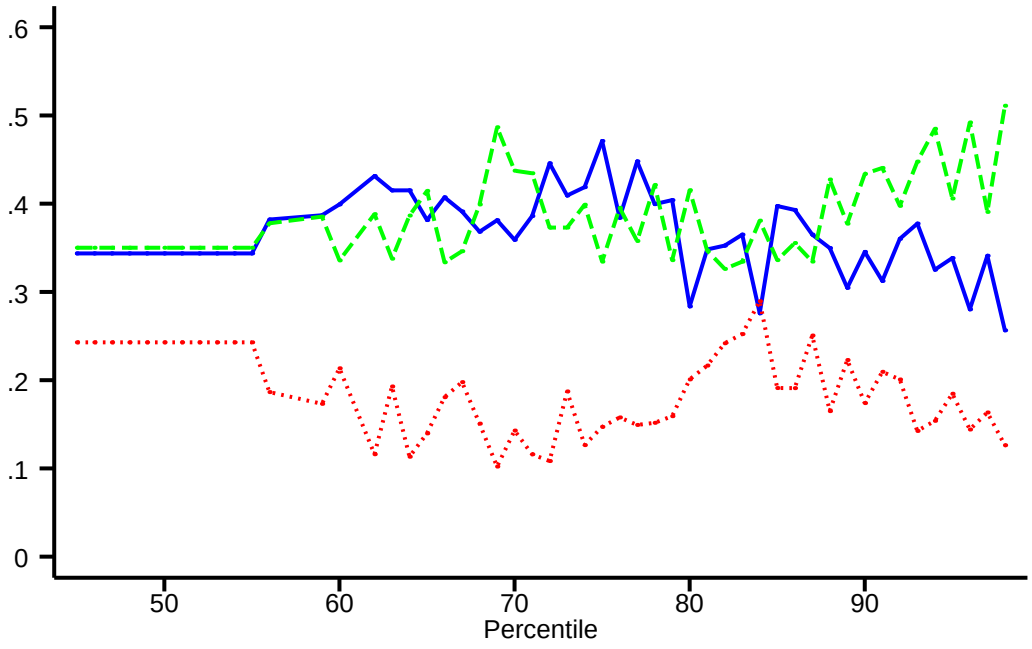
Notes: Solid line is QTE estimate, and dashed line is mean difference, and dotted line is 95% confidence interval.

Figure 4: Quantile treatment effects on the distribution of earnings, quarters 1–7, by race/ethnicity



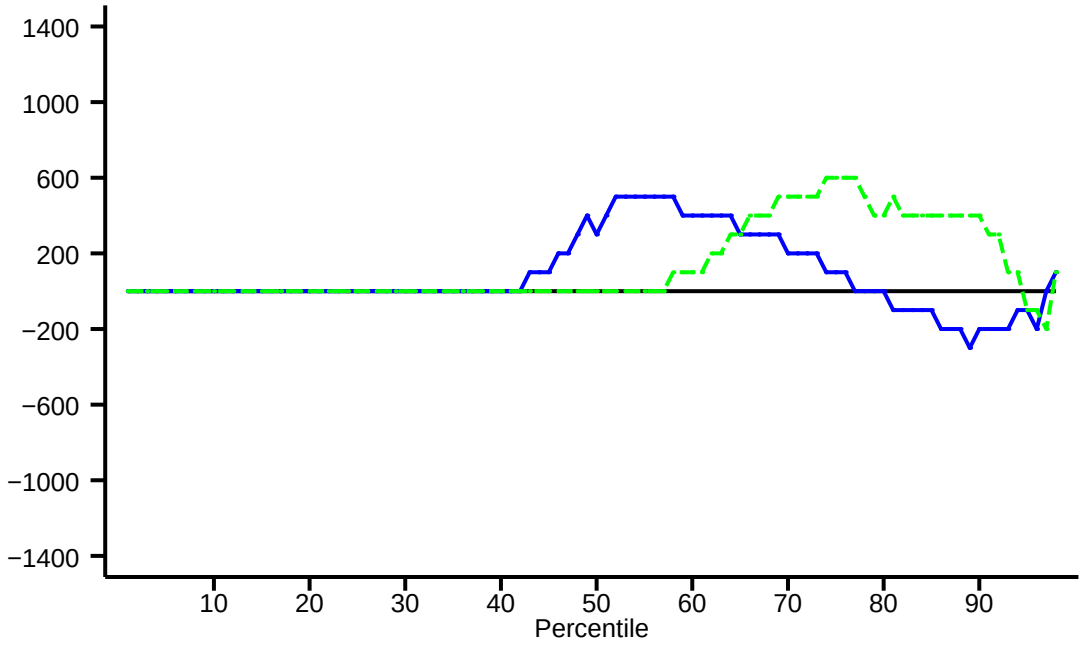
Notes: Solid line is QTE for whites, dashed line is QTE for blacks, and dotted line is QTE estimate for Hispanics.

Figure 5: Race/ethnicity breakdown of the control group distribution of earnings, quarters 1–7



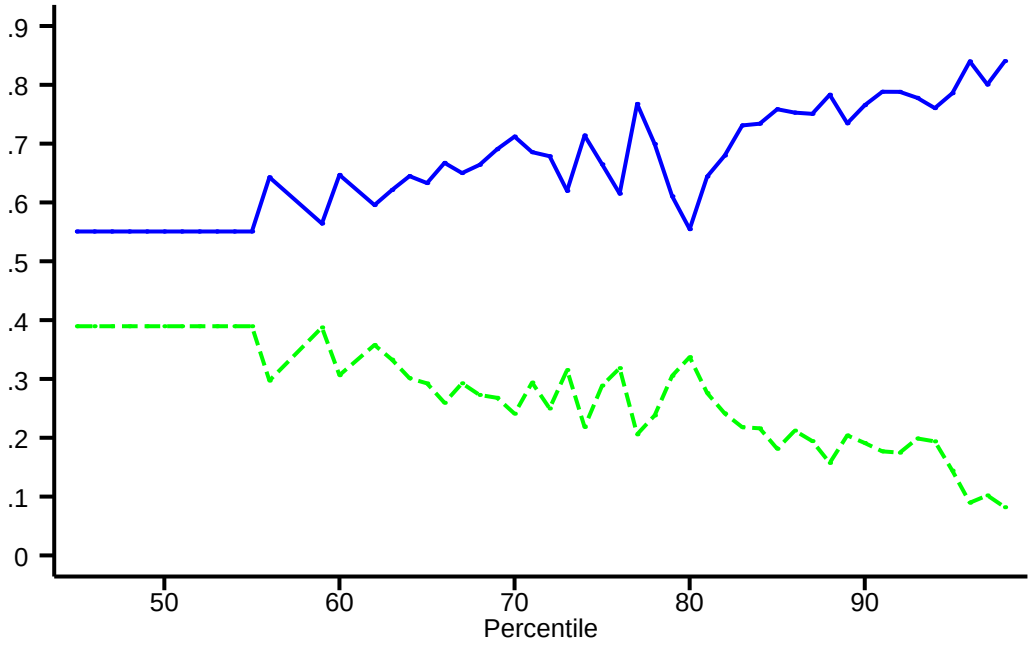
Note: Figure shows share of each centile of control group earnings distribution that is white, black, or Hispanic. Solid line is share white, dashed line is share black, and dotted line is share Hispanic. Shares only shown for for centiles 45–98 of the earnings distribution for the control group. Earnings are zero for all centiles below centile 45 as well. There is no variation in shares for which earnings are zero (includes all centiles below centile 45).

Figure 6: Quantile treatment effects on the distribution of earnings, quarters 1–7, by education of head



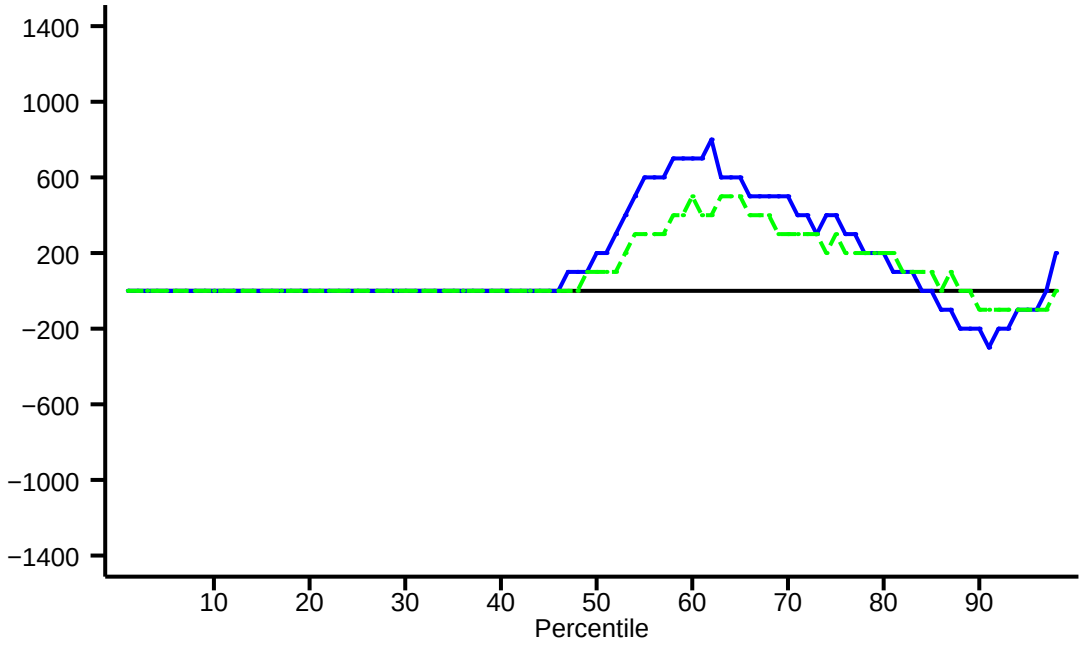
Notes: Solid line is QTE for case heads with at least a high school degree, dashed line is QTE for high school dropout case heads.

Figure 7: Education breakdown of the control group distribution of earnings, quarters 1–7



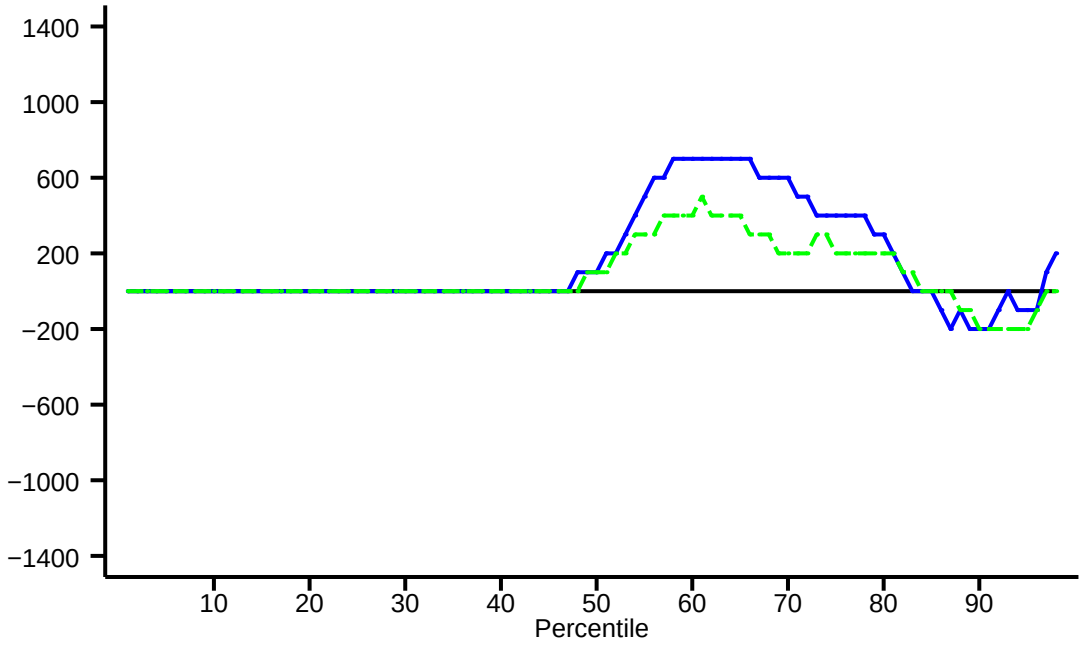
Note: Figure shows share of each centile of control group earnings distribution that has at least a high school degree or are high school dropouts. Solid line is share with at least a high school degree, dashed line is share of high school dropouts. Shares only shown for centiles 45–98 of the earnings distribution for the control group. Earnings are zero for all centiles below centile 45 as well. There is no variation in shares for which earnings are zero (includes all centiles below centile 45).

Figure 8: Quantile treatment effects on the distribution of earnings, quarters 1–7, by whether youngest child is under 6 at RA



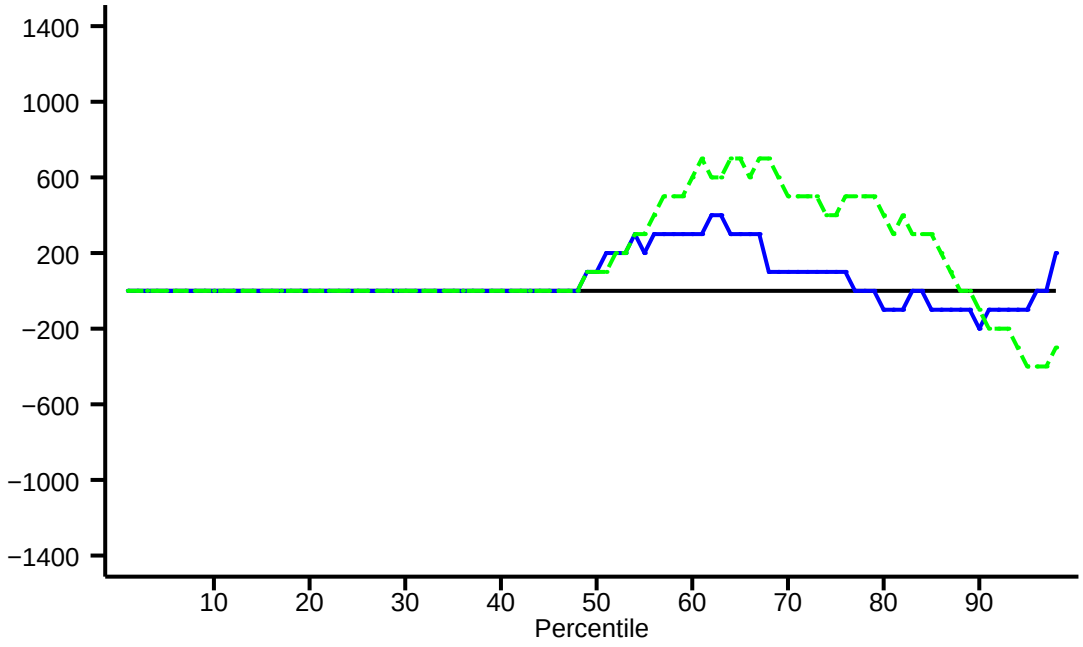
Notes: Solid line is QTE for cases with youngest kid ≥ 6, dashed line is QTE for cases with youngest child ≤ 5.

Figure 9: Quantile treatment effects on the distribution of earnings, quarters 1–7, by marital status of head at RA



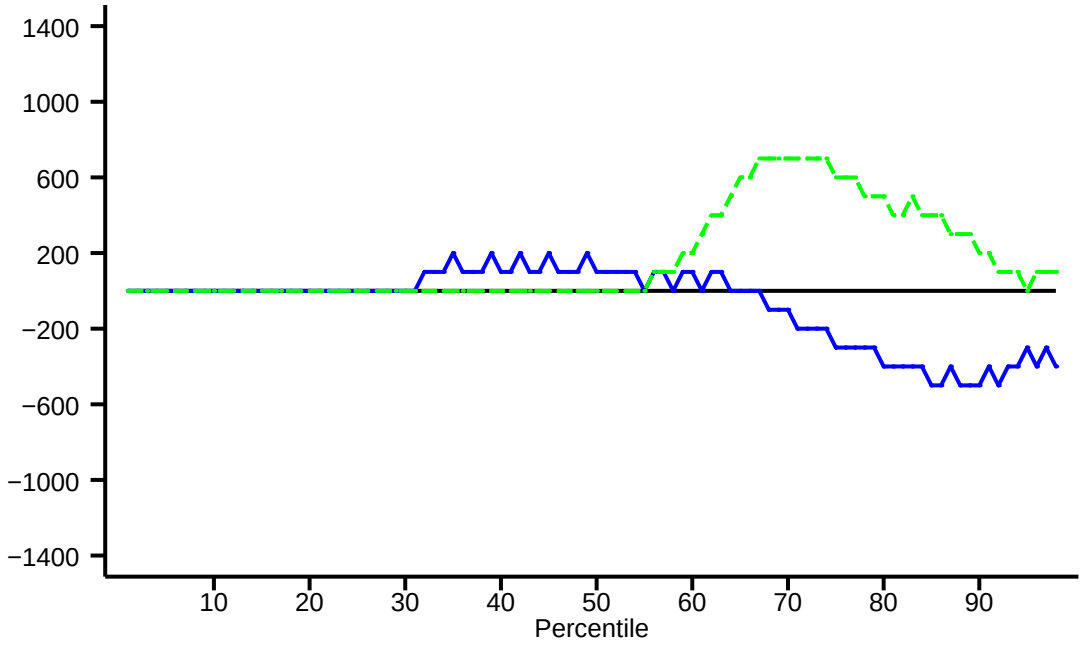
Notes: Solid line is QTE for cases with ever-married head at RA, dashed line is QTE for cases with never-married head at RA.

Figure 10: Quantile treatment effects on the distribution of earnings, quarters 1–7, by receipt of AFDC 7 quarters before RA



Notes: Solid line is QTE for cases with no AFDC 7 quarters before RA, dashed line is QTE for cases with any AFDC 7 quarters before RA.

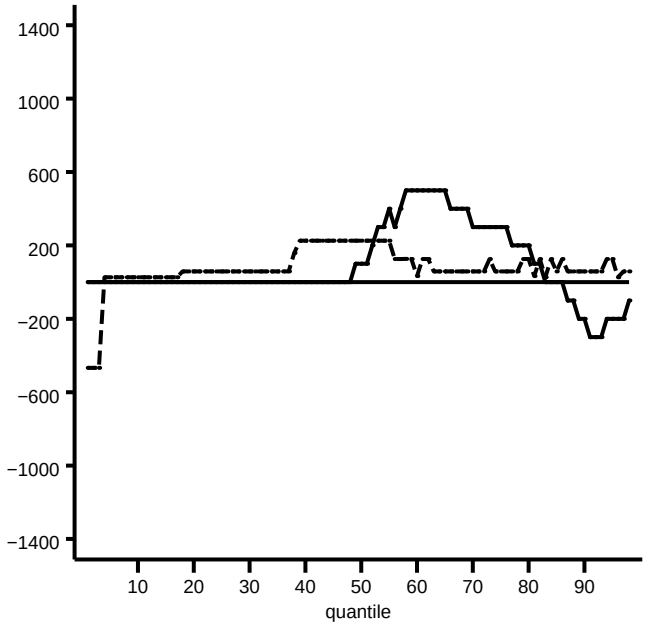
Figure 11: Quantile treatment effects on the distribution of earnings, quarters 1–7, by whether had any earnings 7 quarters before RA



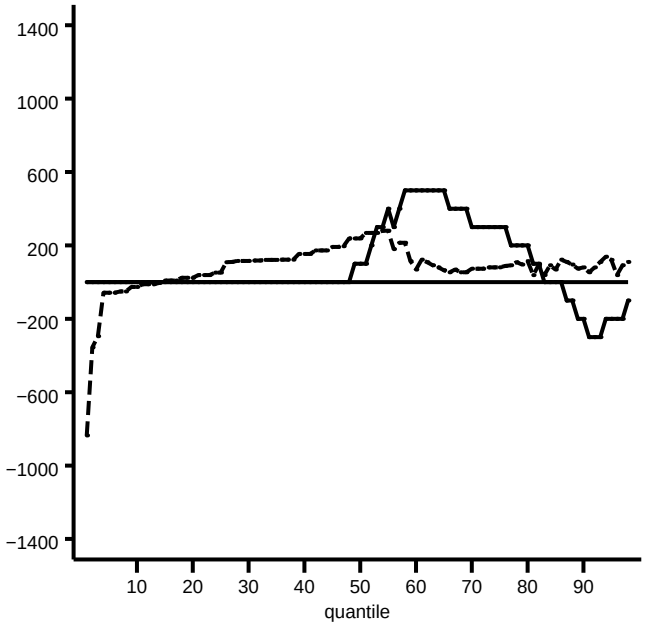
Notes: Solid line is QTE for cases with any earnings 7 quarters before RA, dashed line is QTE for cases with no earnings 7 quarters before RA.

Figure 12: Actual and synthetic QTE by race

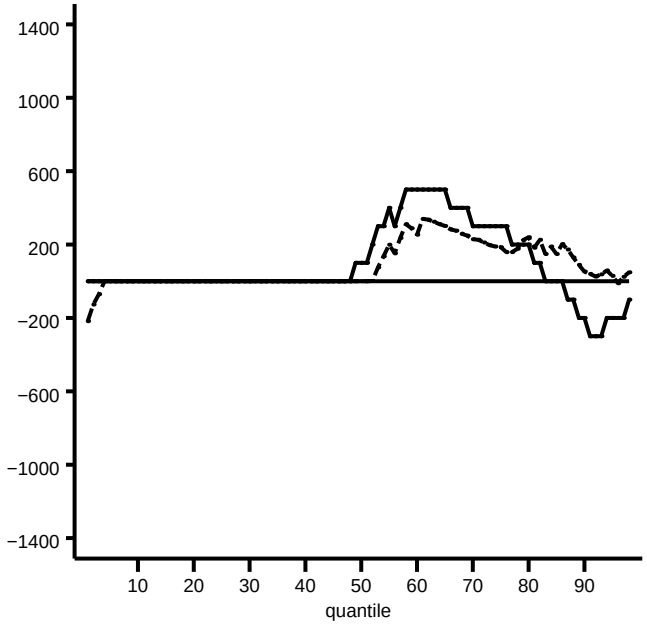
Solid lines: actual QTE.
Dashed line: synthetic QTE.



(a) Time−constant treatment effects



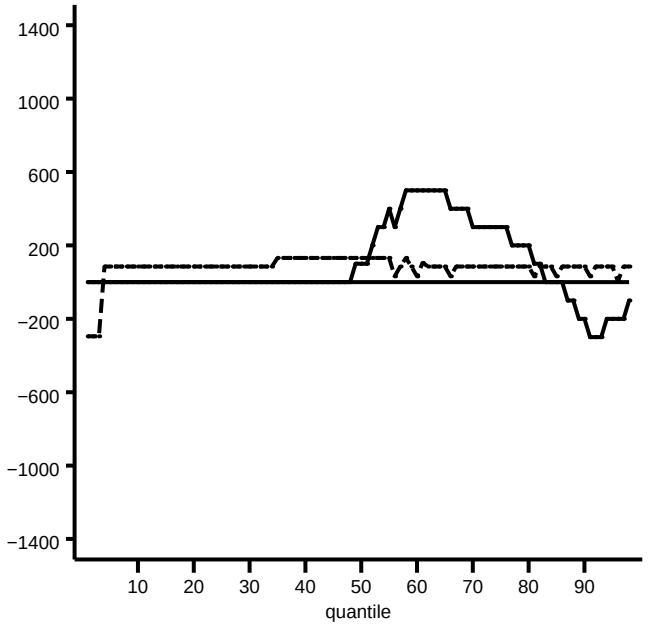
(b) Time−varying treatment effects



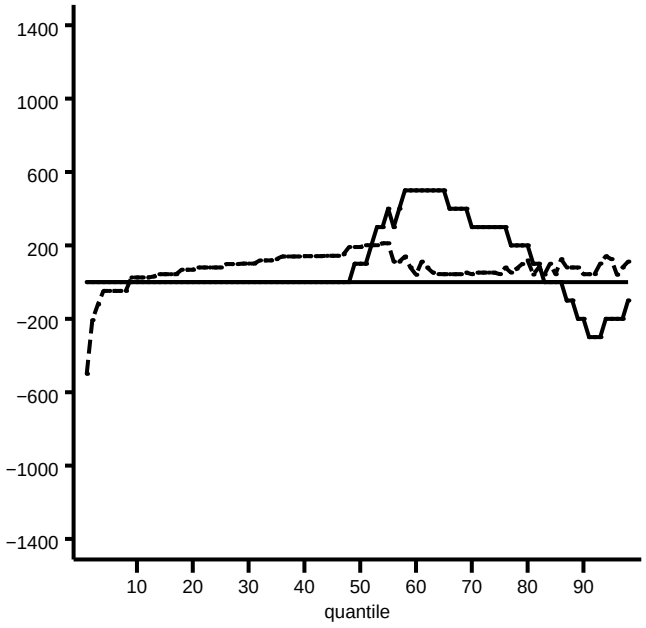
(c) Time−varying treatment effects, equal zero shares

Figure 13: Actual and synthetic QTE by education

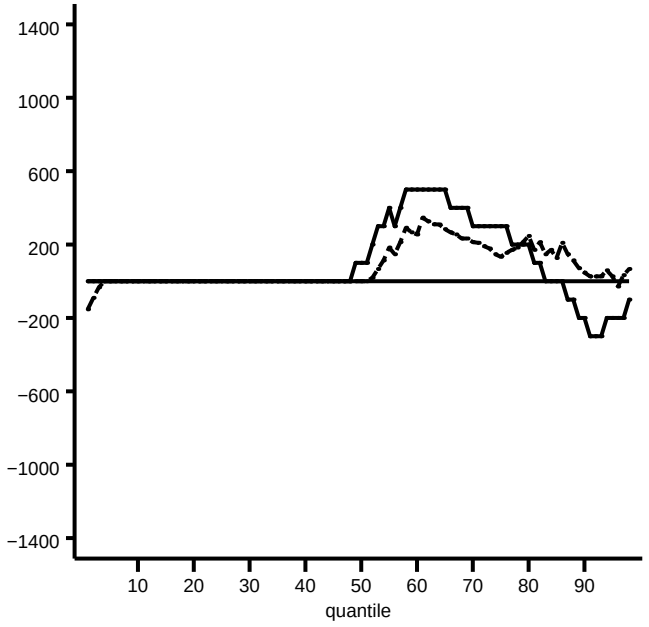
Solid lines: actual QTE.
Dashed line: synthetic QTE.



(a) Time−constant treatment effects



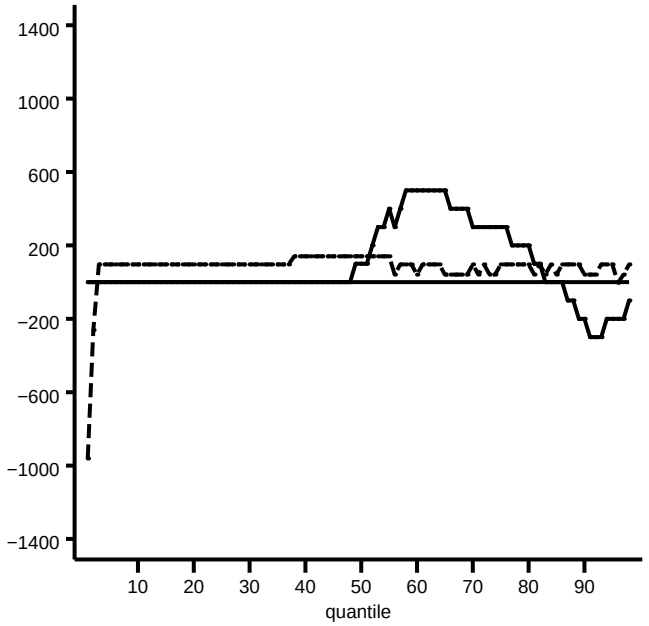
(b) Time−varying treatment effects



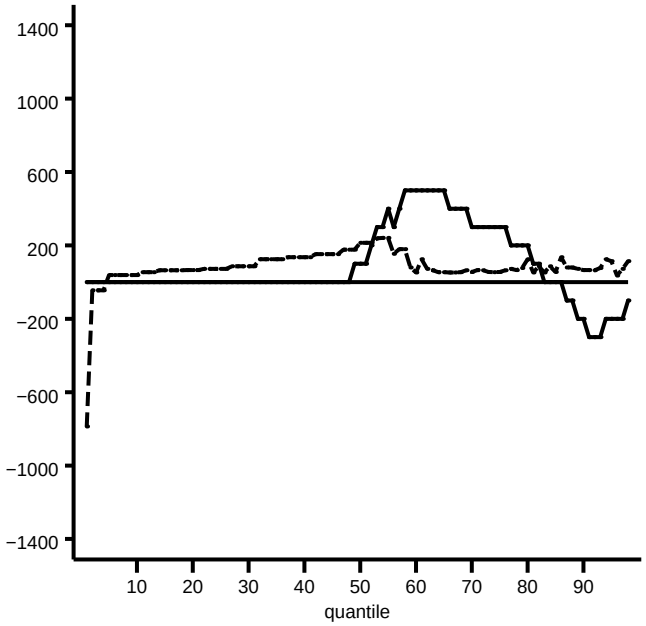
(c) Time−varying treatment effects, equal zero shares

Figure 14: Actual and synthetic QTE by youngest child's age

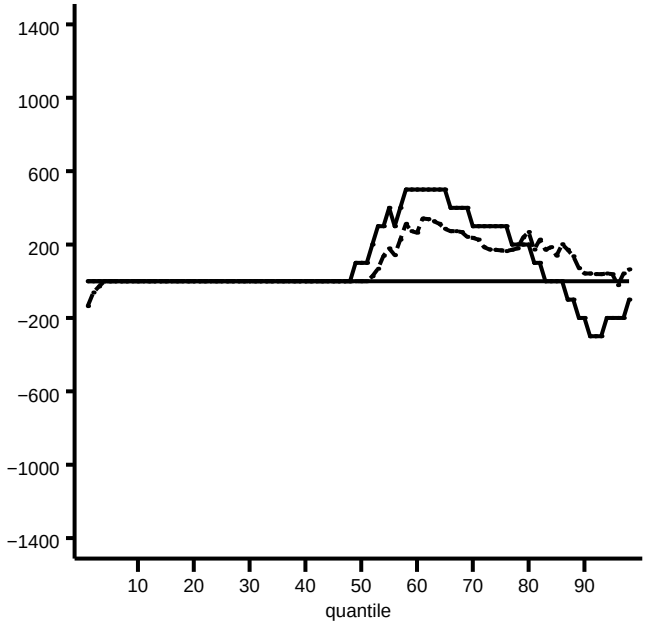
Solid lines: actual QTE.
Dashed line: synthetic QTE.



(a) Time−constant treatment effects



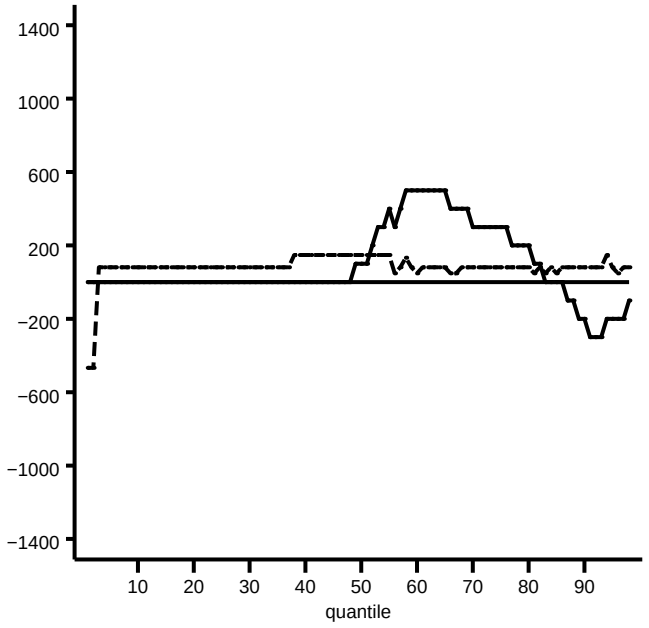
(b) Time−varying treatment effects



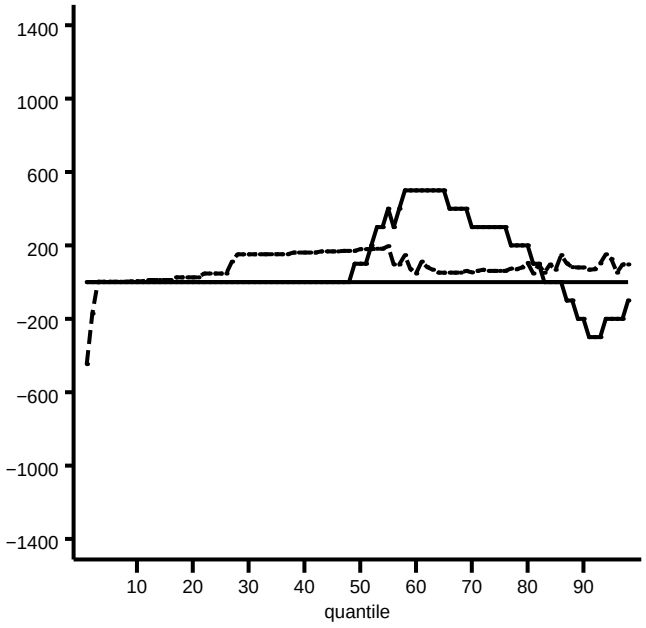
(c) Time−varying treatment effects, equal zero shares

Figure 15: Actual and synthetic QTE by marital history

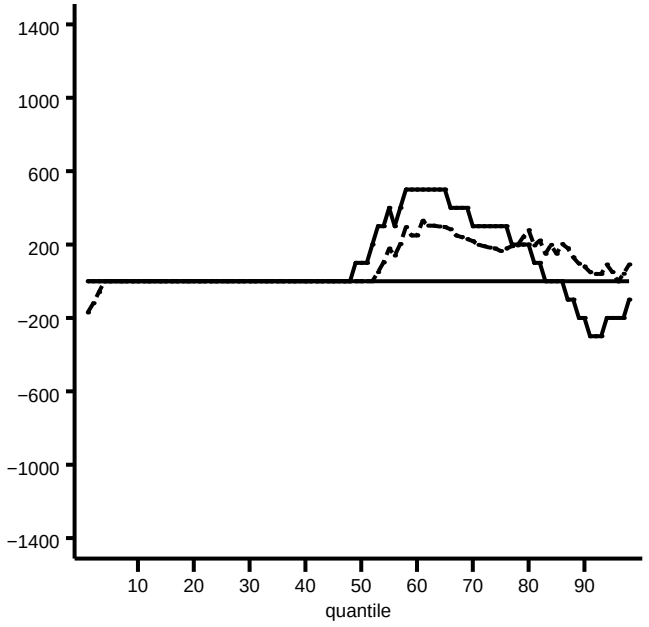
Solid lines: actual QTE.
Dashed line: synthetic QTE.



(a) Time−constant treatment effects



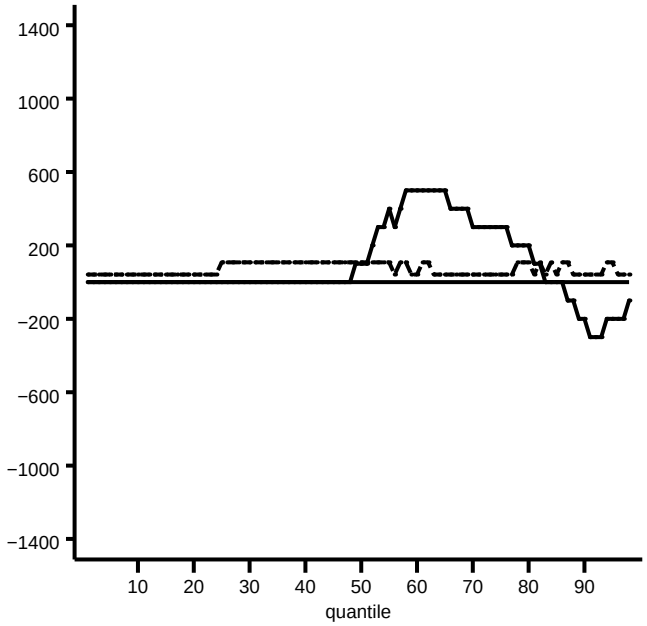
(b) Time−varying treatment effects



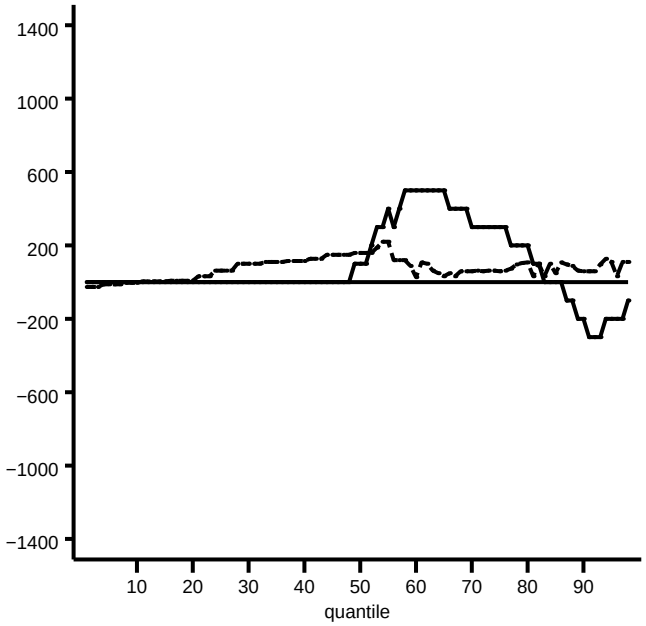
(c) Time−varying treatment effects, equal zero shares

Figure 16: Actual and synthetic QTE by 7 quarters-before-RA AFDC receipt

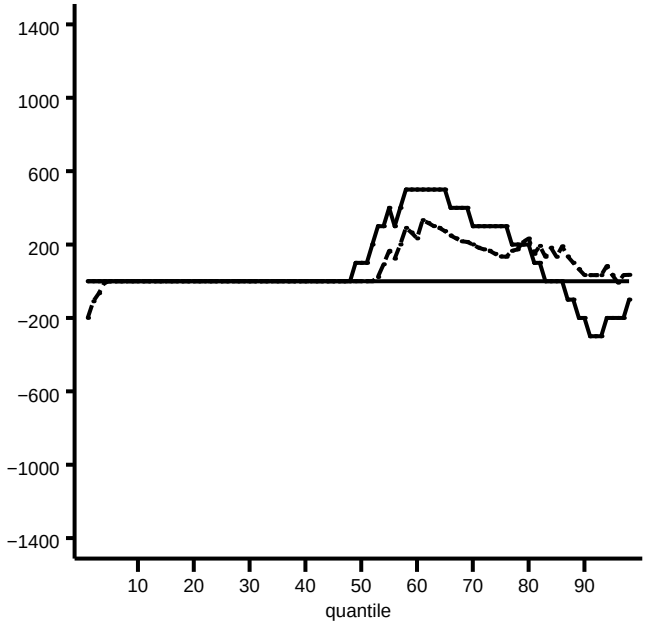
Solid lines: actual QTE.
Dashed line: synthetic QTE.



(a) Time−constant treatment effects



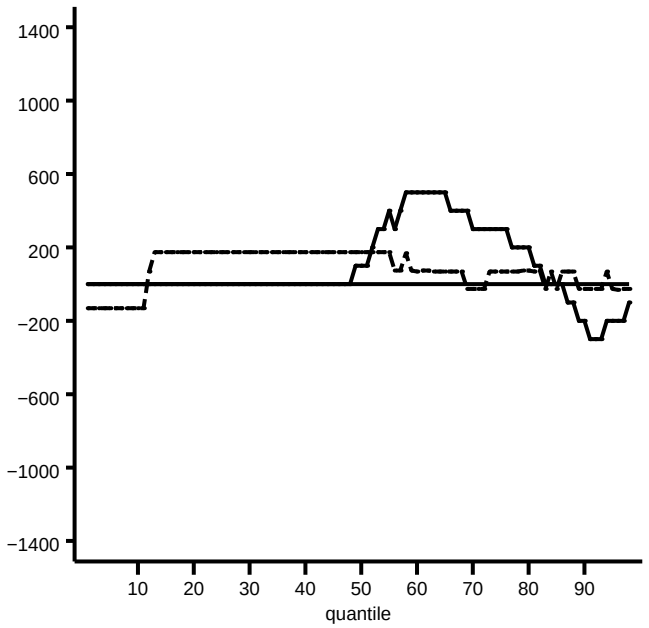
(b) Time−varying treatment effects



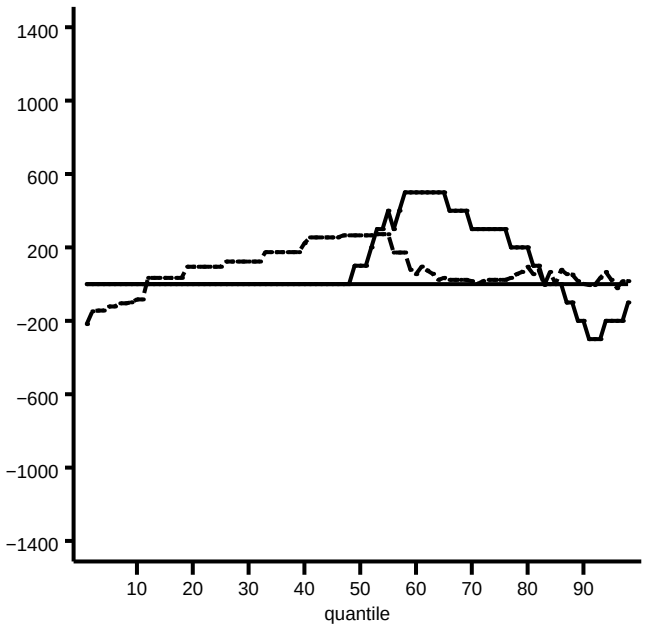
(c) Time−varying treatment effects, equal zero shares

Figure 17: Actual and synthetic QTE by 7 quarters-before-RA employment

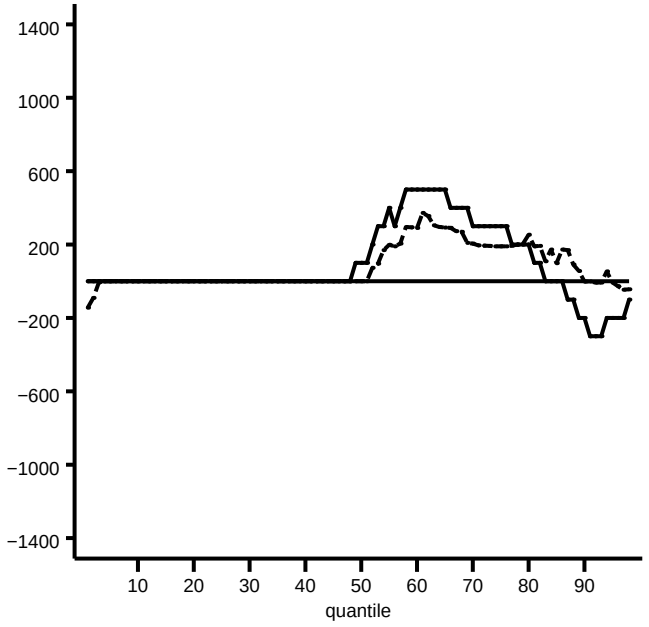
Solid lines: actual QTE.
Dashed line: synthetic QTE.



(a) Time−constant treatment effects



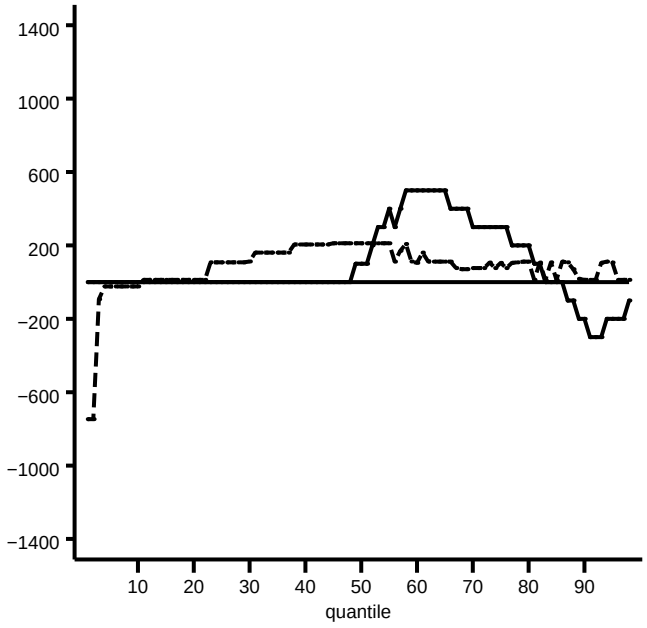
(b) Time−varying treatment effects



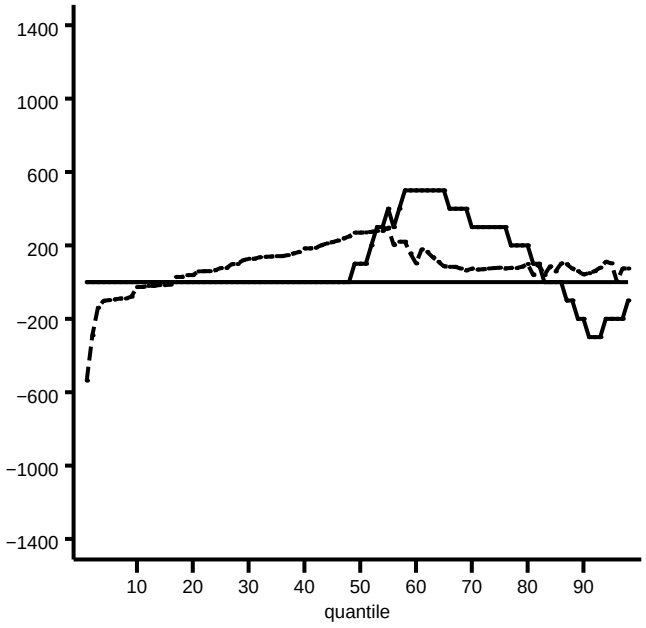
(c) Time−varying treatment effects, equal zero shares

Figure 18: Actual and synthetic QTE by race and education

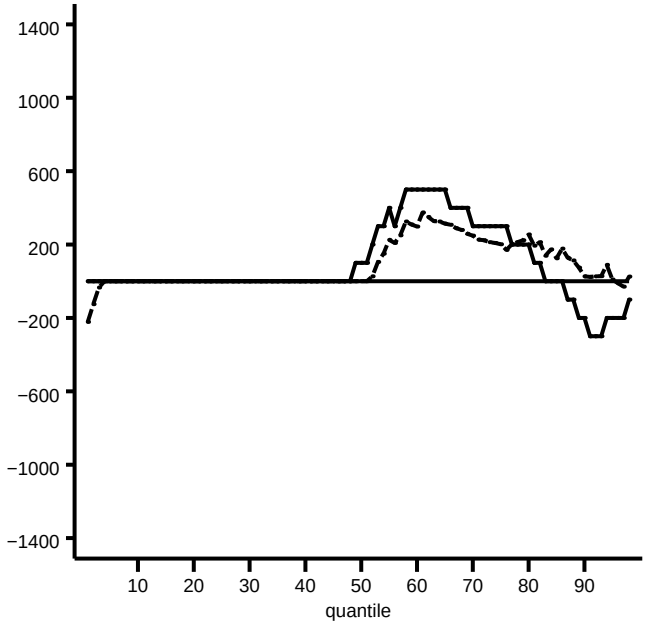
Solid lines: actual QTE.
Dashed line: synthetic QTE.



(a) Time−constant treatment effects



(b) Time−varying treatment effects



(c) Time−varying treatment effects, equal zero shares

Table 1: Characteristics of experimental sample

	Levels		
	Jobs First	AFDC	Difference
Demographic characteristics			
White	0.363	0.349	0.014
Black	0.366	0.369	-0.002
Hispanic	0.208	0.217	-0.009
HS dropout	0.331	0.313	0.018
HS diploma/GED	0.550	0.565	-0.015
More than HS diploma	0.063	0.059	0.004
At least HS diploma/GED	0.613	0.624	-0.011
More than two children	0.227	0.206	0.021*
At least two children	0.484	0.470	0.014
Youngest child 5 or younger	0.536	0.525	0.011
Never married	0.625	0.630	-0.005
Div./wid./sep./living apart	0.317	0.314	0.003
Div./wid./sep./married	0.330	0.325	0.005
Any AFDC 7 quarters before RA	0.546	0.526	0.020
Earnings are zero 7 quarters before RA	0.700	0.673	0.027**
Mother younger than 25	0.290	0.296	-0.006
Mother aged 25–34	0.411	0.416	-0.005
Mother older than 34	0.299	0.287	0.011
Recipient (stock) sample	0.622	0.591	0.032**
Average quarterly pre-treatment values			
Earnings	678	789	-111***
	(1,304)	(1,548)	(41)
Cash welfare	888	832	56**
	(806)	(785)	(23)
Fraction of pre-treatment quarters with			
Any earnings	0.322	0.351	-0.030**
	(0.362)	(0.372)	(0.014)
Any cash welfare	0.571	0.542	0.029**
	(0.452)	(0.450)	(0.013)

Notes: Standard errors in parentheses for lower panels. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively (significance indicators provided only for difference estimates). Standard deviations omitted for binary variables in top panel because all variables are binary. For earnings, 8 quarters of pre-treatment data are used in lower panel. For cash welfare, only 7 quarters are available for all observations. Baseline data on a small number of observations for some variables are missing.

Table 2: Mean differences in earnings during quarters 1–7 by subgroup

Subgroup	Mean difference	95 % CI	Control group mean	N_C	N_T
All	**80**	[-38, 122]	1112	16,744	16,667
By race of case head:					
White	**228**	[161, 294]	1079	5845	6048
Black	**62**	[-5, 129]	1202	6174	6104
Hispanic	**26**	[-48, 100]	882	3633	3472
F-statistic [p-value]	19.6	[0.0000]			
F-statistic excluding NA† [p-value]	7.87	[0.0004]			
By education of case head:					
No HS degree/GED	**133**	[79, 187]	647	5243	5516
At least HS/GED	**87**	[-34, 141]	1322	10,444	10,220
F-statistic [p-value]	9.75	[0.0001]			
F-statistic excluding NA [p-value]	0.99	[0.3187]			
By whether youngest child is ≤ 5:					
Youngest child ≤ 5	**86**	[33, 139]	1084	8792	8939
Youngest child ≥ 6	**143**	[73, 213]	1150	5572	5810
F-statistic [p-value]	40.35	[0.0000]			
F-statistic excluding NA [p-value]	2.24	[0.1347]			
By number of children in case:					
2 or more	**156**	[100, 212]	1042	7868	8071
1 or pregnant	**76**	[20, 132]	1118	8302	8071
F-statistic [p-value]	40.90	[0.0000]			
F-statistic excluding NA [p-value]	2.21	[0.1370]			
By marital status of case head:					
Never married	**82**	[36, 128]	1038	10,542	10,416
Ever married	**151**	[75, 227]	1185	5439	5502
F-statistic [p-value]	15.80	[0.0000]			
F-statistic excluding NA [p-value]	3.36	[0.0667]			
By whether on AFDC 7 quarters before RA:					
Yes	**109**	[54, 165]	956	8813	9100
No	**46**	[-18, 110]	1294	7931	7467
F-statistic [p-value]	2.16	[0.1413]			
By whether earnings are zero 7 quarters before RA:					
Yes	**175**	[-18, 110]	751	11,277	11,669
No	**-126**	[-224, -28]	1903	5467	4998
F-statistic [p-value]	44.6	[0.0000]			

Mean differences overall and by subgroup for earnings during quarters 1–7, with 95 percent CIs, the control group mean, and the number of observations in the treatment and control groups. Each set of subgroup differences also contains 2 rows with tests for the mean treatment effect being the same across the subgroups, the first of which includes the the missing data category and the second excluding it. Differences are for treatment (Jobs First) versus control (AFDC) group using 4773 observations with data for all 16 quarters, with inverse propensity score weighting. Means are differences in quarterly earnings for all sample members in the subgroup mentioned in the column label for quarters 1–7. Numbers in **bold** are mean differences.

† NA denotes data missing for relevant variable.

Table 3: Test of whether QTEs deviate from those calculated by adding mean TE within subgroup to the control group distribution

Subgroup	Real data $\hat{S}$	Critical value (5% level)	Reject?
Full sample	2976	804	Yes
Race	3810	1107	Yes
Education	3657	1029	Yes
Age of youngest child	3927	1173	Yes
Marital status	3876	1116	Yes
Race by education	3198	1212	Yes
Education by age of youngest child	2925	1179	Yes

Notes: Modified K-S test for whether QTEs deviates from null of constant average treatment effects within subgroup added to control group. Label reports subgroup for row. First column reports the real data difference, second column reports the critical value, and the third column reports the results of the test (reject/do not reject). All estimates in this table for significance level of 0.05. All estimates using 4,773 observations. Means within subgroup estimated with propensity score weighting. Differences for bootstraps have real data difference subtracted out. Statistics normalized by bootstrap estimate of the variance of the difference terms. For more information, see text.