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

Marianne P. Bitler

University of California, Irvine and NBER

Jonah B. Gelbach

University of Arizona

Hilary W. Hoynes

University of California, Davis and NBER*

This version: May 2010

*Correspondence to Hoynes at UC Davis, Department of Economics, 1152 Social Sciences and Humanities Building, One Shields Avenue, Davis, CA 95616-8578, phone (530) 564-0505, fax (530) 752-9382, or hwhoynes@ucdavis.edu; Gelbach at gelbach@email.arizona.edu; or Bitler at mbitler@uci.edu. This paper was previously circulated under the title "Can Constant Treatment Effects Within Subgroup Explain Heterogeneity in Welfare Reform Effects?" 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 Joe Altonji, Richard Blundell, Mike Boozer, David Brownstone, Moshe Buchinsky, Raj Chetty, Julie Cullen, David Green, Jeff Grogger, Jon Guryan, John Ham, Pat Kline, Thomas Lemieux, Bruce Meyer, Robert Moffitt, Enrico Moretti, Giuseppe Ragusa, Jeff Smith, Melissa Tartari, and Rob Valetta for helpful conversations, as well as seminar participants at the IRP Summer Research Workshop, the SOLE meetings, the Harris School, UBC, UC Davis, UC-Irvine, UCL, UCLA, UCSD, UCSB, the San Francisco Federal Reserve Bank, Tinbergen Institute, Toronto, and Yale University.

Abstract

Randomized control trials (RCTs) are the gold standard source of data for program evaluation in a wide range of applications. When using data from RCTs, researchers typically estimate mean impacts. The most common approach to exploring treatment effect heterogeneity is to estimate separate mean treatment effects across different subgroups. However, mean treatment effects may be ill suited or insufficient to capturing treatment effect heterogeneity when there is within- as well as across-group treatment effect heterogeneity. In Bitler, Gelbach & Hoynes (2006), we took a different approach and applied quantile treatment effect estimators to estimate experimental effects using data on Connecticut's welfare reform. Our results revealed heterogeneous impacts of welfare reform on labor supply. A key result is that the effects' patterns are consistent with theoretical predictions from static labor supply models. In the present paper, we use the same experimental data to comprehensively assess the ability of models assuming constant mean impacts within subgroups to explain observed heterogeneity, as measured by quantile treatment effects. A key contribution of this paper is that we show how to nonparametrically test the null hypothesis that all treatment effect heterogeneity measured via QTE is the result of cross-subgroup variation in mean treatment effects. We then implement this test using a variety of subgroup definitions. We find that mean impacts generally fail to replicate evidence consistent with strong predictions of labor supply theory, evidence which was revealed in our prior work by quantile treatment effect estimates. However, some approaches work better than others. Subgroups based on individual labor supply histories perform best. Further, adjusting subgroups' mean impacts for extensive margin impacts of reform leads to a substantial improvement in the ability of cross-subgroup mean treatment effects to explain distributional effects of welfare reform.

1 Introduction

Randomized control trials provide the gold standard data used for program evaluation in a wide range of areas such as education, development, medicine, and, increasingly, social programs. The appeal of randomized control trials is that simple treatment/control comparisons yield unbiased and consistent estimates of program effects so long as the treatment is randomly assigned. A universal starting point for any evaluation is to present average treatment effects (ATE) estimated unconditionally, i.e., simple differences in mean outcomes across the treatment and control groups. However, theory often predicts, and is almost always consistent with, heterogeneous impacts of a reform program. To explore the possibility of treatment effect heterogeneity, many researchers estimate average treatment effects separately for different sample subgroups. This approach yields multiple conditional average treatment effect estimates, where the conditioning information is subgroup membership. Like unconditional ATEs, these estimates are unbiased and consistent when treatment is randomly assigned. An alternative way to explore heterogeneity, and one that also relies only on randomization, is to estimate quantile treatment effects.

In previous work (Bitler et al. (2006)) we applied quantile treatment effects estimators to evaluate the impact of welfare reform on labor supply. The quantile treatment effects are estimated simply as the difference in outcomes at various quantiles of the treatment group and control group distributions and capture estimates of the impact of treatment on the outcome distribution. Our policy context was one where the predicted effects of labor supply varied across different groups in the population. Our quantile treatment effect estimates revealed behavioral responses consistent with theoretical predictions—increases in earnings in the middle of the distribution and decreases in earnings at the top of the distribution. In contrast, the mean impacts showed small earnings effects.

In this paper, we extend our earlier work. Our primary goal is to comprehensively assess the ability of models assuming constant mean impacts within subgroups to explain observed heterogeneity, as measured by simple quantile treatment effects. As in Bitler et al. (2006) we use data from a randomized experiment of welfare reform in Connecticut. We consider subgroups used commonly in the welfare reform literature and also subgroups suggested by theory and prior work to be proxies for varying wage opportunities and costs of leisure.

We take two primary approaches. First, we calculate group-specific QTE for different subgroups and examine whether there is evidence of heterogeneity in reform's effects on earnings *within* these

groups. Second, we develop and implement a formal test of whether constant treatment effects within subgroup (CTEWS) models are sufficient to characterize QTE estimated on the sample of women pooled across subgroups. If such a parametric model of treatment effects is correctly specified, then the QTE implied by that model should equal the QTE that would be calculated nonparametrically without imposing the model.

We discuss the implementation of the parametric CTEWS models below and in detail in the appendix and also provide there an illustrative simulation. In brief, though, we assess these models via a simple four-step process. In the first step, we use earnings data on experimental treatment and control group observations to estimate subgroup-specific treatment effects. In the second step, we use these estimates together with data on the control group to create what we call a synthetic treatment-group earnings distribution. In the third step, we calculate synthetic quantile treatment effects by subtracting the estimated quantiles of the actual control group distribution from the corresponding quantiles of the synthetic treatment group earnings distribution. In the fourth and final step, we compare these estimated synthetic quantile treatment effects to the quantile treatment effects estimated by subtracting the sample quantiles of the control group from the sample quantiles of the actual treatment group (rather than the synthetic treatment group).[1] If any of the CTEWS models is a good description of the actual effects of welfare reform on the earnings distribution, then the synthetic and actual quantile treatment effects should be equal up to sampling variation.

As we discuss below, one requirement for a CTEWS model to replicate the QTE we see are that mean impacts vary systematically across subgroups. Therefore, we also discuss estimated mean impacts for detailed subgroups in detail. Moreover, when actual quantile treatment effects are heterogeneous, CTEWS models can be correct only if there is systematic sorting of subgroups into different parts of the pooled control and treatment group earnings distributions. We therefore also test for such sorting.

The Connecticut welfare reform experiment—Jobs First—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, the Jobs First welfare reform generates substantial, and salient, changes in incentives for labor supply. We therefore expect to be able to capture important labor supply responses in this setting. Third, we have administrative *pre-random assignment* data on earnings and welfare history to supplement the standard demographic variables for forming subgroups. This rich dataset allows us to form subgroups that are not only more detailed

[1] These latter estimates are computed using inverse propensity score weighting, as we discuss below.

than is possible in standard data sets but also more likely to characterize wage opportunities and incentives. This turns out to be very important. 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 by a CTEWS model here, it may be even less likely to be explicable in non-experimental settings.

Our choice of subgroups is guided by the predictions of labor supply theory: We select subgroups to proxy for wage opportunities, fixed costs of work, and preferences for income versus leisure. We start with standard demographic variables including education and marital status of the mother, and the number and ages of her children.[2] In addition, we make use of the earnings and welfare history data that are available in our experimental data which allows for analyses not possible with standard observational data. We also interact some of these subgroups to create two-way or three-way classifications such as education by work history. Lastly, we allow the standard, CTEWS model its best chance of success by (i) allowing for time-varying subgroup mean treatment effects and (ii) accounting for differences in the extensive margin effects of welfare reform within subgroup.

Our main finding is that mean impacts fail to replicate the strong predictions of labor supply theory that are revealed by the QTE. Our quantile treatment effect estimates show that 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 most by the woman's education and her earnings history. Further, whether the subgroup means are defined for the full time period or allowed to be time-varying, 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 two-way classifications between education, welfare history, or earnings history.

But not all subgroups fare equally well in this comparison. Prior labor supply decisions, as measured here by earnings prior to the treatment, turn out to be the most informative measure. In particular, subgroups based on prior labor supply decisions are the only ones whose synthetic QTE reveal a reduction in labor supply which labor supply predicts and is evident in the full sample QTE. In contrast, subgroups based on standard demographic variables (education, numbers and ages of children) are much less successful. This is important to point out because it is relatively

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

unusual to have earnings history data available for such an analysis.

Lastly, we explore whether a simple adjustment to the subgroup means for the extensive margin impacts of reform suffices for our synthetic distributions to resemble our true ones. This possibility is important since so many women in both the treatment and control groups have zero earnings. Here, the synthetic QTE based on mean effects within demographic subgroups again fail to replicate the true QTE, as do the synthetic QTE based on welfare history. However, again with subgroup means based on prior earnings or predicted wages and an adjustment for participation, the actual QTE and the synthetic QTE are similar and we cannot reject statistically that the actual and associated synthetic earnings distributions are the same.[3]

The paper provides important contributions to the program evaluation literature. To our knowledge, ours is the first paper to construct a testable, nonparametric null hypothesis under which all heterogeneity is driven by treatment effects that are constant within, but vary across, identifiable subgroups. This is an important innovation because many applied researchers (reasonably) use subgroups to try and isolate sample members likely to respond to program or policy changes. If our finding—that this approach falls short of capturing the actual form of heterogeneous treatment effects—holds more generally, then the approach of assuming constant treatment effects within subgroups will need to be reconsidered. Estimating mean impacts may miss a lot, even across multiple subgroups, so researchers should report not just mean impacts, but also estimators capable of capturing distributional effects, like those used here and in Bitler et al. (2006).

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. In Section 4, we discuss the empirical methods and present the results for the mean treatment effect and quantile treatment effects in Section 4. We also present there our tests for the adequacy of the subgroup-specific treatment effects compared to the QTE and the results of those tests. We conclude in Section 5.

2 Welfare Reform, Jobs First & Predicted Labor Supply Effects

We compare QTE and mean impacts by using data from a randomized experiment on welfare reform in Connecticut. The nature of this experiment and its implications for labor supply predictions

[3]Note that some part of this is mechanical as the extensive margin adjustment constrains the share of non-participants in the treatment and control groups to be the same.

make it an ideal setting for conducting this analysis. Before going into detail on the experiment and its data, we begin with some background on welfare reform.

In the 1990s, welfare reform dramatically changed the system of cash support for low income families with children in the United States. Federal welfare reform occurred in 1996 with passage of the Personal Responsibility and Work Opportunity Act (PRWORA). This law required all states to eliminate Aid to Families with Dependent Children (AFDC) and replace it with Temporary Assistance for Needy Families (TANF). Under TANF, welfare recipients now face lifetime time limits for welfare receipt, stringent work requirements and the threat of financial sanctions. Many states, including Connecticut, also increased the financial incentives to work—by raising so-called earnings disregards.[4] Considerable state welfare reform also occurred prior to the 1996 federal welfare reform. In fact, more than half of states experimented with their AFDC programs in the period leading up to PRWORA, receiving waivers to modify existing policy rules.

In this project we analyze Connecticut's waiver program, called Jobs First (which later became its TANF program). This is an excellent setting to compare QTE to mean impacts for several reasons. First, we can examine the impact of welfare reform in an experimental setting—here individuals were randomly assigned to either the treatment group (Jobs First) or the control group (AFDC).[5] By using experimental data, we are able to avoid the pitfalls of identification in the context of non-experimental analyses (Blank (2002)). Second, the Jobs First program implemented very large changes to incentives to work, making it ideal for examining labor supply responses. Finally, the Connecticut experiment is appealing because its waiver program is among the most TANF-like of state waiver programs, including 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.

Jobs First differs from the pre-existing AFDC program in several key 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

[4]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).

[5]We use data from the waiver period rather than the TANF period because of the availability of randomized experimental data. Federal law required states seeking welfare waivers to evaluate the effects of the waiver 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. However, evaluation was not required when states implemented their TANF programs.

poverty guideline 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.[6] Furthermore, work requirements and financial sanctions were strengthened in the Jobs First program relative to AFDC.[7]

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). This figure illustrates the Jobs First-induced dramatic change in the budget constraint, increasing the net of tax and transfer wage considerably. In this discussion, and in our empirical implementation below, we consider women who have been on aid for fewer than 21 months, so that the time limit does not yet bind.

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. What we have in mind here 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 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. 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.[8] A woman who would not work (i.e., would locate at point A)

[6] 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.

[7] For more information on these and other features of Jobs First see Bloom, Scrivener, Michalopoulos, Morris, Hendra, Adams-Ciardullo & Walter (2002) and Bitler et al. (2006).

[8] Assuming that offered wages do not vary with hours worked, predictions about hours worked map one-to-one to predictions about earnings. This fact is important since we observe earnings but not hours worked. In Bitler et al. (2006) 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

when assigned to AFDC will either stay out of the labor force or will locate at some point on AF when she is instead assigned to Jobs First. A woman who combines work and welfare if assigned to AFDC (i.e., would locate at point C) will increase her hours if assigned to Jobs First as long as the substitution effect dominates the income effect. Consider next a woman who, at some time after random assignment, 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. After time limits hit, the treatment group reverts to the no welfare budget set. In this analysis, we consider only the period prior to time limits binding.

This discussion makes clear that the effect of welfare reform for a given woman will depend on her wage opportunities, preferences for income versus leisure, and fixed costs of work. This, of course, gives us guidance about the expected nature of the heterogeneous responses to welfare reform. In particular, we expect larger extensive margin employment responses for women with higher wages and lower fixed costs of work. Further, women with strong tastes for income relative to leisure and those with high wages are more likely to end up at relatively high hours points (above B) and thus are predicted to reduce their hours of work when assigned to the welfare reform. This informs our choice of subgroups—which we choose to predict wages, preferences, and fixed costs of work and therefore proxies for position on the budget set.

---

wages at the top of the wage distribution (see Bitler, Gelbach & Hoynes (2008)). Card & Hyslop (2005), in another study of SSP, provide a dynamic model suggesting such behavior and also find empirical evidence in support of it. However, SSP differs from Jobs First, most notably in that it provides a limited time period for experimental participants 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).

3 Data

The evaluation of the Connecticut Jobs First program was conducted by MDRC. In this analysis, we use public-use data made available by MDRC. The data include information on 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, with random assignment taking place between January 1996 and February 1997.

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] This earnings and welfare history data turns out to be very important in this analysis. The administrative data are augmented by demographic data collected at experimental baseline that include each woman's number of children, education, age, marital status, race, ethnicity, etc., all at the time of random assignment. Our analysis uses the full sample of 4,803 women.

Our outcome variable is quarterly earnings over 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 and labor supply predictions are cleanest before the time limit. In our analyses, we pool all seven quarters of data for each woman, leading to a total of $4,803 \text{ x } 7 = 33,621$ observations.

The above discussion concludes that the expected effects of Jobs First on labor supply will vary with three key factors: a woman's wage, her preferences for leisure versus income, and her fixed costs of work. This suggests identifying variables that provide proxies for—or predictors of—women's wages including her education, work history, age, and marital status. Proxies for preferences for leisure versus work include number and ages of children and welfare and earnings history. In particular, age of youngest child is an important predictor of fixed costs of work (child care) and the value of time out of work. Welfare history may also be a useful predictor of long term dependency and possibly valuation of leisure.

[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.

To connect our subgroups more explicitly to economic theory, we also use an estimated wage equation, based on a sample of low education female heads of household from the 1992–1995 Current Population Survey, to predict wages for each women in our sample. We use this to construct subgroups based on predicted wages.[10] Another way to connect our subgroups more explicitly to theoretical predictions would be to use pre-random assignment earnings to "place women on the budget set." This is appealing, but is hampered by Ashenfelter's dip, which is severe in the case of welfare receipt. This is exacerbated by the fact that half the sample is in the middle of a welfare spell when they are randomized into the experiment. The net effect is that only one-third of the sample has any earnings in the seventh quarter prior to random assignment (which is the furthest back we can go).[11]

In order to explore whether the experiment is balanced, Table 1 reports means of the baseline characteristics. The first two columns provide means for the Jobs First (column 1) and AFDC (column 2) groups. The third column reports the unadjusted differences in means across the program groups, with indicators as to when the differences are statistically significantly different from zero.[12] As described in Bloom et al. (2002) and Bitler et al. (2006), average values of some of these characteristics differ statistically by treatment assignment. 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 and higher welfare benefits for the period prior to random assignment. A standard test for joint significance of the differences in subgroup indicators (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

[10] We use a sample of all women without a 4-year college degree aged 16–54 living with a child aged 0–15 who are not currently married and living together with their spouses. We recode the CPS variables to match the variables in the experimental data. We then estimate a simple parametric Heckman model to deal with selection into work, and predict potential wages using average hourly wages for working women. We control for a rich set of interactions of race/ethnicity, marital status, education and age. Identification comes not only from functional form but also from instruments for the age of youngest child and number of children which are good predictors of selection into work.

[11] A further difficulty in mapping pre-random assignment earnings into locations on the budget constraint is that we have family size and composition only at random assignment, and we have earnings and welfare history only for the woman while she was living in Connecticut. If for example, she became divorced, but was previously married to a worker, we might place her at a point A when she really should have been located elsewhere.

[12] For time-invariant pre-random assignment characteristics, these are the results of simple t-tests. For time-varying pre-random assignment characteristics, they are the result of regressions of that characteristic on the program variable, with robust standard errors that allow for arbitrary within-person correlations.

the unbalanced sample using inverse propensity score weighting, as in Bitler et al. (2006).[13] 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, 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, 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. Column 4 of Table 1 shows that with the inverse propensity score weighting, all the mean differences are very close to zero, and never close to statistically significant. 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 Results

In Section 4.1, we provide basic mean-impact results. We then turn in Section 4.2 to observed estimates of quantile treatment effects for both the full sample and various subgroups. We also investigate whether women in each subgroup are spread evenly throughout the overall treatment or control group earnings distributions, or whether they are over-represented in some parts of these distributions. Then in Section 4.3, we report plots of the synthetic QTE generated by different versions of CTEWS models. Finally, in Section 4.4, we test the null hypothesis that the synthetic quantile treatment effect estimates generated by various CTEWS models are equal to the simple quantile treatment effect estimates plotted in Section 4.2.

[13] Firpo (2007) shows that this approach yields asymptotically consistent estimates of QTE for continuous dependent variables. Because MDRC essentially rounds the earnings data we use to the nearest hundred dollars, 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. We ignore this issue henceforth.

4.1 Results: Mean impacts

The Jobs First experiment generates significant changes in the financial incentives to work. Static labor supply theory predicts that the reform will lead to increases in earnings for some—those with lower wage levels and/or weak tastes for income versus leisure—and decreases in earnings for others—those with higher wage levels and/or strong tastes for income versus leisure. Those with high fixed costs of work, or very high disutility of work, may experience no change in earnings.

To begin, we explore whether the data is consistent with these heterogeneous predictions by estimating mean impacts of the reform within subgroups of the population. This is a common approach taken in the welfare reform literature as well as in the evaluation literature more broadly.

At this point, it is helpful to briefly introduce the usual 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 treatment effect for person i in period t is equal to the difference between her period-t outcome when treated and untreated: $\delta_{it} \equiv Y_{it}(1) - Y_{it}(0)$. Of course, 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. As we discuss below, the CTEWS model is one way of filling in the missing data.

Using overbars to denote sample means, random assignment allows us to obtain consistent estimates of the average effect of the reform using the difference in sample means between treatment (Jobs First) and control (AFDC) groups: $\overline{\delta} \equiv \overline{Y}(1) - \overline{Y}(0)$. This same approach can be used to estimate average effects for a given period t across subgroups ($\overline{\delta}_t \equiv \overline{Y}_t(1) - \overline{Y}_t(0)$) or within subgroup g ($\overline{\delta}_t(g) \equiv \overline{Y}_t^g(1) - \overline{Y}_t^g(0)$). Accounting for the slight imbalances in random assignment discussed above simply requires calculating weighted means using the inverse propensity scores as weights.

We posit that the response to welfare reform may depend on a woman's wage opportunities, her preferences for income versus leisure, and her fixed costs of work. Our first set of subgroups are based on standard demographic variables that are often used when estimating mean impacts by subgroup: educational attainment, marital status, age of youngest child, and number of children. We also use these variables to construct subgroups based on predicted wages constructed using the CPS-estimated wage equation. Our second set of subgroups is derived from employment and welfare history variables, which are available in our experimental data set but are not available in

most public-use data sets (e.g., the CPS).

In Table 2, we report estimated mean treatment effects for the full sample and for the demographic subgroups. We report the estimated mean treatment effect in column 1, its 95 percent confidence interval in column 2, and the AFDC control group mean in column 3. The top panel shows the overall number of observations in control and treatment groups in columns 4 and 5, while the other panels show the share of the control and treatment groups in each subgroup in columns 4 and 5. At the bottom of the panel for each subgroup, we present an F-statistic and p-value for testing the null that the subgroup means are equal.

The first row, for the full sample, shows that Jobs First is associated with a statistically insignificant increase in quarterly earnings of $82, representing a 7% increase over the control group mean of $1,113. This small, statistically insignificant effect of reform on earnings is a typical finding in the welfare reform literature (Blank (2002) and Grogger & Karoly (2005)).

The remaining rows in the table show mean impacts for subgroups defined by education, marital status, and the number and ages of children. The results show some differences in the point estimates across groups, with larger mean impacts for those with lower education levels, older children, more children, and for those who have been previously married. These differences in point estimates are broadly consistent with the labor supply predictions showing smaller impacts for those with higher wage opportunities (as illustrated by the education results) or with high fixed costs of work (illustrated by the results for age of youngest child). However, there are no subgroups with negative mean impacts, which is theoretically predicted for relatively high wage women.

Further, and importantly, the differences across these groups are not statistically significant. For example, we cannot reject the equality of the mean treatment effects of $131 for high school dropouts and $90 for women with high school graduates ($F = 0.83$, implying a p-value of 0.36). The same is true for the subgroups based on marital status, number of children, and ages of children (see the table for F-statistics). This lack of heterogeneity in the underlying mean impacts is not unique to the Connecticut experiment. In their comprehensive review of the welfare reform literature, 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).[14]

[14] A relatively small subset of the sample has missing values for these demographic variables, having not completed the background questionnaire. If we include these observations and form separate mean impacts for the "missing data" subgroups, we can reject that the means are equal across the subgroups. For example, including the 284 women with missing data on educational attainment as a separate subgroup, we can reject the equality of mean impacts for high school dropouts, high school graduates, and those missing education. See the notes to Table 2 for details. Note that in constructing the synthetic earnings variables we use below, we treat women with missing data as a separate

Table 3 presents similar mean impacts and tests for equality of mean impacts for subgroups based on pre-random assignment earnings and welfare history, as well as predicted wages. Because a large fraction of the sample is in the middle of a welfare spell at random assignment, we find it most compelling to use earnings and welfare history measured in the seventh quarter prior to random assignment (the farthest back we can go)—in an attempt to measure outcomes before Ashenfelter's dip and actual program participation takes place, although we also use the number of quarters of earnings in these seven before random assignment.

The results using earnings history show striking and statistically significant differences across subgroups. We see that the estimated mean earnings impact is $174 among those with no earnings in the seventh quarter prior to random assignment and -$119 among those with earnings in the seventh quarter prior to random assignment. The reported F-statistic shows that these estimates are statistically significantly different from each other. These estimates are notable because a reduction in labor supply at the top of the distribution is a key prediction of static labor supply theory. We explore this further and use earnings in the seventh quarter prior to random assignment to create three groups: Zero earnings, low earnings, and high earnings. Similarly, we take the share of the seven quarters prior to random assignment with positive earnings and construct three groups: Zero quarters with positive earnings, low share, high share. In both cases, the "low" and "high" groups are constructed to be above and below the median among the observations with nonzero values. Mean treatment effects using these additional earnings subgroups show large negative mean impacts for those with high pre-random assignment earnings (-$358) and a high share of pre-random assignment quarters with positive earnings (-$134). Conversely, they show large positive effects for those with zero/low earnings or zero quarters with earnings. Again, the F tests show that the differences in these estimates are statistically significant.

The results using presence of AFDC income in the 7th quarter prior to random assignment show a larger treatment effect ($115) for those with AFDC income in the seventh quarter prior to random assignment compared to those without AFDC income at that point ($43). Finally, we take the demographics and predict wages for each woman in the sample, using the CPS-estimated wage equation.[15] We construct groups (low, medium, high predicted wages) so that the share of women in each group matches the share in the three subgroups based on earnings in the 7th quarter prior to random assignment. Mean impacts are very small for the high predicted wage group ($7), a bit

category. This ensures we always have the same sample for all comparisons.

[15]The estimates from the wage prediction equation are available upon request.

larger for the middle predicted wage group, and then in between the other two values for the zero predicted wage group. The F-statistic shows they are statistically significantly different from one another.

Overall, we find that using the common approach of estimating mean impacts within standard demographic subgroups yields little evidence of the heterogeneous treatment effects implied by the theory. This is true whether we look at proxies for wages such as completed education or marital history, or at proxies for preferences for leisure versus work such as number and ages of children. However, the differences in mean impacts using earnings history, and to a lesser extent, welfare history, are economically and statistically important. Unfortunately, few existing studies use earnings history because it is rarely available. Hardly any large repeated cross-sectional survey data sets (e.g., the Current Population Survey) would contain such retrospective data. Longitudinal data sets which do collect such data often have relatively small samples or substantial attrition. Even experiments sometimes fail to collect detailed earnings histories.

4.2 Results: Quantile treatment effects

Quantile treatment effects are the simple distributional analog of mean impacts. Let $F_d(y)$ be the population earnings distribution for women when assigned to program group d. For any $q \in (0,1)$, the q^{th} quantile of F_d is the smallest value y such that $F_d(y) \geq q$.[16] The q^{th} quantile treatment effect is the simple difference between the quantiles of the treatment (Jobs First) and control (AFDC) distributions: $\Delta_q = y_{q1} - y_{q0}$.

Aside from weighting, then, the q^{th} quantile treatment effect may be estimated very simply as the difference across treatment status in the two outcome quantiles.[17] 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 quantile treatment effect at the median. Other quantile treatment effects are estimated analogously; we evaluate the distributions at all 99 centiles. To estimate QTE for subgroups, we calculate quantiles from the sample of women belonging to each subgroup and then proceed as above.[18] quantile treatment effect at a

[16]Note that for expediency, we adopt the convention that we will use $q \in (0,1)$ or 100 times this value to denote the q^{th} quantile.

[17]As with mean impacts, accounting for the imbalance in random assignment simply requires calculating the distributions F using the inverse propensity scores as weights. The weighted empirical distribution function is defined as $\widehat{F}_d(y) \equiv \sum_{i=1}^{n} 1(D_i = d)1(Y_i(d) \leq y)\hat{\omega}_i / \sum_i 1(D_i = d)\hat{\omega}_i$, where $\hat{\omega}_i$ is the estimated inverse propensity score weight.

[18]One can find pooled means using the population-weighted average of subgroup means. The relationship between

QTE capture heterogeneity in that they tell us how the distribution changes with random assignment of the Jobs First treatment. As with mean impacts, random assignment is sufficient for identification of quantile treatment effects and consistency of sample QTE. There is, however, an important distinction between the quantile treatment effects and quantiles of the distribution of δ_i, the individual treatment effect. Because the q^{th} quantile of $\{\delta_i\}$ is a nonlinear function of the joint distribution of $(Y_i(0), Y_i(1))$, the q^{th} quantile treatment effect does not generally equal the q^{th} quantile of $\{\delta_i\}$. In some special cases, the q^{th} quantile of $\{\delta_i\}$ does equal the corresponding quantile treatment effect. Two such cases are homogeneous treatment effects (i.e., the treatment effect for everyone is identical) and rank preservation, meaning that each woman's rank in the distribution is unaffected by treatment assignment. Homogeneous treatment effects are clearly rejected by the data discussed below. Rank assignment is a strong assumption, and it will fail here if, for example, preferences for work do not map one-to-one with rank in the earnings distribution. While QTE do not identify the distribution of treatment effects, they are still appealing (e.g., see reasons discussed in Bitler et al. (2006) and Heckman, Smith & Clements (1997)).

We begin by presenting the QTE for the full sample, shown in Figure 2. For this and the subsequent QTE, we show the QTE for 98 centiles.[19] As above, we use the person-quarter as the unit of analysis and analyze the 33,621 observations on quarterly earnings during the first seven quarters. The solid line plots the estimated QTE, the dotted lines plot upper and lower bounds for 95% pointwise confidence intervals, the dashed (horizontal) line shows the estimated mean treatment, and the 0-line is provided for reference.[20]

Heterogeneity in Jobs First's impact across the earnings distribution's quantiles is unmistakably significant, both statistically and substantively. Figure 2 replicates Figure 3 of Bitler et al. (2006), which shows that for quarterly earnings in the pre-time limit period, the QTE are zero for all

within-group and pooled quantiles is more complicated, owing to the fact that any given quantile is a highly nonlinear functional of a distribution. For a useful illustration of the relationship between pooled and conditional quantiles when the conditional quantile function is linear in covariates, see Machado & Mata (2005).

[19] We computed quantile treatment effect results at quantile 99 but omit them from this figure and those below because their variances are frequently large enough to distort the scale of the figures.

[20] 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 $\widehat{y}_q - (\widehat{y}^*_{q,\alpha/2} - \widehat{y}_q)$; see, e.g., Hall (1988) for a discussion of these formulas. With 999 replications and $\alpha = 0.05$, $\widehat{y}^*_{q,\alpha/2}$ will be given by the 25th smallest bootstrap quantile treatment effect 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 (this result is unsurprising given the rounded nature of our data).

quantiles $q \leq 48$. This result occurs because quarterly 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, several of which are statistically significant.

This pattern is consistent with the predictions of labor supply theory discussed above (we argue this point in detail in Bitler et al. (2006)). In particular, the reduction in earnings at the top of the earnings distribution, which leads to negative QTE, is a key prediction of the theory. Finally, we note that the quantile treatment effect point estimates 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 performed a joint test of equality of the QTE to the mean treatment effect ($82); this test rejects equality of the QTE at the 5 percent level.[21] Thus, as in Bitler et al. (2006), we conclude that a single mean treatment effect cannot explain the heterogeneity in quantile treatment effects that we estimate in Figure 2.

Given the heterogeneity evident in the full-sample QTE and the lack of heterogeneity in mean impacts, the natural next step is to estimate QTE within subgroups. In the top left panel of Figure 3, we plot estimated QTE for high school graduates (solid line) and high school dropouts (dashed line).[22] The pattern of estimated QTE for each of the education subgroups 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. 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 at the top of the distribution. This finding is in line with the expectation that more educated women are more likely to locate at points like D, E or H in Figure 1 under AFDC assignment, where theory predicts a decline in earnings.

While the figure for QTE for subgroups defined by education shows some variation in the location of the positive QTE across education group, the shapes of the two QTE profiles do resemble one another considerably. To explore this further, we plot each education group's share of the sample

[21] See Appendix 2 of Bitler et al. (2006) for details of this test.

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

at each centile of the earnings distribution estimated by pooling all women in the control group.[23] This plot is interesting because subgroup-specific quantiles must be equal when each subgroup's share of a population is constant across the population's quantiles.[24] Thus, if the education groups are evenly distributed across the quantiles of the treatment and control group earnings distributions, then the subgroup QTE will be identical to the full sample QTE.

In the right panel of Figure 3, we see that—as expected—the dropout share falls, and the high school graduate share rises, as we move to higher centiles of the earnings distribution. This pattern suggests that there may be more scope for CTEWS models based on education subgroups to explain some of the heterogeneity in treatment effects. However, as shown in Appendix Figure 1, the other demographic variables (marital status, age of youngest child) exhibit little difference in the conditional QTE or the subgroup shares across the earnings distributions. This suggests that marital status and number and ages of children are of limited use in explaining the heterogeneity evident in the full sample QTE.

In Figure 4, we explore the within-group variation in earnings impacts using the earnings and employment history variables to define subgroups. In the top half of Figure 4, we present the subgroup QTE and shares across the earnings distribution using earnings in the seventh quarter prior to random assignment. We include subgroups based on high earnings (solid line), low earnings (dashed line), and zero earnings (dotted line). These figures show substantial within- and across-group heterogeneity in estimated QTE for these subgroups. Among women with no employment income in the seventh quarter prior to random assignment, the estimated QTE are zero for more than the bottom half of the earnings distribution, with estimated effects being positive and sometimes large higher in the earnings distribution, before turning negative at the very top of the distribution. A reasonable interpretation is that these women are most likely to locate at a point like A in Figure 1 when not assigned to Jobs First. This means they increase their hours and

[23]The analogous figure looks qualitatively similar for the treatment sample. We note that we plot group shares only for $q \geq 45$, since earnings quantiles are zero for all women in each group for all $q < 45$.

[24]Formally, let Y be a random variable with K points of support given by $y \in \{y^1, y^2, \ldots, y^K\}$. Let f_{pooled} be the probability mass function for the pooled population, so that $f_{\text{pooled}}(y)$ is the share of the overall population that has $Y = y$. For each group $g \in \{1, 2, \ldots, G\}$, let $f_g(y)$ be the share of group-g women with $Y = y$. Also, let π_g be group g's share of the pooled population; thus, $\sum_g \pi_g = 1$. Define $s^g(y)$ to be group g's share of women who have value $Y = y$. Observe that for each y_k, $s^g(y_k) = \pi_g f_g(y_k) / \sum_{h=1}^{G} \pi_h f_h(y_k)$. Because the denominator of this identity equals $f_{\text{pooled}}(y_k)$, $f_{\text{pooled}}(y_k) = \pi_g f_g(y_k)/s^g(y_k)$. If subgroup g's share of the population is constant across the pooled population's quantiles, then it must be true that $s_g(y_k) = \pi_g$ for any $k \in \{1, 2, \ldots, K\}$, so that $f_{\text{pooled}}(y_k) = f_g(y_k)$ for all k. Thus, a subgroup with a constant population share at every quantile also has the same probability mass function as the pooled population. This result implies equal *cdf*s, which implies equal quantile functions (e.g., see lemma 21.2 of van der Vaart (1998)).

earnings under Jobs First relative to their hours and earnings under AFDC.

For women with high levels of employment income in the seventh quarter before random assignment, the estimated QTE are zero only for the first third of the distribution, are positive and small for only about five centiles, and are negative for the top two-thirds 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 D, E and H of Figure 1 under AFDC assignment. This interpretation is supported by the top right panel of Figure 4, which shows that the women with high earnings in the seventh quarter prior to random assignment are disproportionately found near the top end of the control group earnings distribution, those with low earnings in the seventh quarter pre-RA near the middle, and those with 0 earnings near the bottom. (This is also true for the analogous figure for the treatment group, which is not shown here.) By contrast, for women with low but non-negative earnings, the QTE are positive for middle third of the distribution, and then the QTE become negative.

The results in the bottom of Figure 4, which present similar plots using the share of quarters prior to random assignment with positive earnings, show qualitatively similar findings. We regard the substantial range of negative QTE for the women with earnings history and the large gains for those without as additional evidence of behavior consistent with labor supply theory. In Appendix Figure 2, we present similar figures using welfare history and predicted wages. Those figures show more similarity across subgroups.

Taking the results in this and the prior section together, we find striking evidence consistent with labor supply predictions. The QTE by subgroup results show that more highly educated women and women with more earnings history are more likely to have negative QTE at the top of the earnings distribution. Those with less education and less work history show positive QTE throughout more of their earnings distribution. However, demographic variables and mean impacts are not sufficient to reveal these important results. Figures breaking down the control group earnings distribution by subgroup show that few of the demographic variables can be used to pin down women's location in the earnings distributions, though one would be more successful using the earnings history variables. Further, the similarity in the overall shape of the QTE across subgroups and the substantial within-group heterogeneity suggest that, like mean impacts overall, mean impacts by subgroups are likely to miss a lot.

4.3 QTE, mean impacts, and heterogeneity

We now turn to the question of whether the cross-subgroup heterogeneity in mean treatment effects documented in Tables 2–3 can explain the heterogeneity in quantile treatment effects documented in figures 2–4. To do so, we use several variations on the CTEWS treatment effects model to generate estimates of the synthetic treatment group earnings distribution that would hold if the CTEWS model were true. In the interest of brevity, we relegate a detailed discussion of the underlying assumptions, theory, and formal null hypotheses to the appendix, where we also present some illustrative simulations.

We consider three CTEWS models. In the first, overall mean impacts are estimated for each subgroup. In the second, we allow the subgroup mean impacts to vary over time. In the third, we adjust for participation effects (the extensive margin) of Jobs First. In each case, to assess the performance of these CTEWS models using Jobs First data, we compare a synthetic earnings distribution to the actual control group distribution.[25]

To illustrate, suppose we want to calculate the synthetic QTE implied by our most restrictive—yet simplest—CTEWS model, where the mean impacts vary only across subgroups. In particular, suppose that the subgroups are based on educational attainment, broken into those with less than a high school education, high school or greater, or missing data on education. First, we estimate the mean impacts for these subgroups (these are reported in Table 2). Second, we take each earnings observation in the AFDC group, and to it we add the estimated mean impact corresponding to the woman's education subgroup. The resulting sum is the synthetic treatment value for each woman. Third, we calculate the sample quantiles of the synthetic earnings distribution. Finally, we calculate the synthetic quantile treatment effect estimates by subtracting the corresponding sample quantiles for the actual control group earnings distribution. Adjusting this method to allow for time-varying subgroup mean impacts (the second of the CTEWS models we consider) involves simple changes to steps one and two above; namely adding the estimated mean impacts for each quarter and subgroup to each woman's earnings.

We can estimate our most generalized CTEWS model, which allows Jobs First assignment to affect both the share of observations with zero earnings and the conditional mean of earnings among

[25] On the one hand, these models make strong assumptions about the estimates using mean impacts: all important variation in the effects of treatment must be captured by the subgroup mean treatment effects. On the other hand, these models do make use of the full variation within the control group distribution in forming the synthetic treatment distribution. A richer specification for the mean impact model might add an error term, drawn from some distribution to the subgroup specific mean treatment effect.

those with nonzero earnings, by modifying the first step and reweighting the data (see Appendix Section C for details on the reweighting). Of course, we can use these approaches to create the synthetic QTE using subgroups based on any other variable. Further, we can create more detailed subgroups by combining, say, education and earnings history.

We begin by considering subgroups defined using demographic variables. In Figure 5, we present synthetic QTE where the treatment effects vary across education (top panels) and age of youngest child (bottom panels). The figures in the left most panels constrain the means to be constant across the seven quarters (e.g., we pool quarters), while those in the right most panels allow the means to vary by quarter since random assignment. In each of these (and all subsequent) figures, we compare the synthetic QTE (dashed line in the figures) to the full sample QTE (solid line). The full sample QTE included here are identical to those plotted in Figure 2. These figures make visual comparisons of the estimated synthetic QTE and actual sample QTE. In the next section we present the results of formal tests for comparing the actual and synthetic treatment distributions; the relevant nulls and procedures for constructing the test statistic are in the Appendix.

Figure 5 shows that the synthetic QTE described above do a very poor job of replicating the actual QTE. The synthetic QTE show no evidence of negative QTE at the top of the distribution and do not achieve as great a maximum QTE as the actual QTE. Further, the synthetic QTE also fail to replicate the large range of quantiles over which the actual QTE are zero, which occurs because about half the person-quarters in each program group exhibit no earnings. This happens because in the restrictive version of the CTEWS model, all observations have the same (nonzero) treatment effect. These qualitative results also obtain for subgroups based on education, age of youngest child, and others that we omit for brevity, like marital status and number of children.

One obvious enhancement to this simple CTEWS model just used is to allow subgroup mean impacts to vary with time since random assignment. After all, we have earnings data for seven quarters after random assignment, and treatment effects may differ across time. To implement this less-restrictive CTEWS model, we estimate mean treatment effects for each quarter-by-subgroup cell. We then add the appropriate treatment effect to each quarterly control-group earnings observation, which yields the synthetic treatment-group potential outcome for the observation. We then calculate the sample quantiles of the synthetic treatment group distribution. As above, we calculate estimated synthetic QTE by subtracting actual control-group sample quantiles from corresponding synthetic treatment-group sample quantiles. We plot the resulting estimated synthetic QTE in Figure 5, with subgroups based on educational attainment in the top-right panel, and subgroups

based on age of youngest child in the bottom-right panel. Allowing for time-varying mean impacts creates a bit more of a hump shape in the synthetic QTE, but the synthetic QTE again fail to reflect the full range of qunatile treatment effect heterogeneity we see in the actual QTE.

In Figure 6, we present additional graphs in which we again estimate synthetic earnings allowing distinct subgroup-by-time treatment effects. Subgroups in this figure involve additional cuts of the data based on earnings and welfare history, as well as predicted earnings. The subgroups are defined as those with zero, low, or high earnings in the seventh quarter prior to random assignment (top-left panel); zero, low, or high share of quarters with earnings prior to random assignment (top-right panel); those with or without any welfare income in the seventh quarter prior to random assignment (bottom-left panel); and those with low, medium, or high predicted wages (bottom right panel). As with those shown above based on demographic subgroups, the synthetic QTE in these figures do not match the key features of the actual QTE. In particular, there is little or no evidence of a systematic decline in earnings at the top of the distribution.

4.3.1 Adjusting the synthetic distributions for participation

One obvious and striking point of difference between the actual and synthetic QTE just discussed arises from the fact that 48 percent of person-quarters in the treatment group, and more than that share in the control group, exhibit zero earnings. This feature generates the result that all actual QTE for $q \leq 48$ are zero. By comparison, the synthetic QTE do not have this feature, for a simple reason: When we add a nonzero mean treatment effect to all control-group observations in the relevant cell, observations with zero actual earnings wind up with nonzero synthetic treatment-group earnings. Without accounting for the substantial mass point at zero earnings, we would not expect synthetic QTE to match the actual QTE.[26]

To address this issue, we consider a final version of the CTEWS model, which we call the participation-adjusted CTEWS model. This version includes an adjustment that ensures that the actual treatment sample and the synthetic treatment sample have the same fraction of observations with non-zero earnings. Under this model, the Jobs First program is allowed to have two types of effects within each subgroup-quarter cell. First, the program is allowed to affect the share of women with positive earnings. Second, the program is allowed to affect the mean level of earnings

[26] In fact, Heckman et al. (1997) point out that when both the control and treatment-group distributions have non-negligible mass points at zero and non-negligible shares of observations with nonzero outcome values, the data will always reject that treatment effects are constant.

among those who work.

We implement the participation-adjusted CTEWS model as follows. First, we calculate the conditional mean level of earnings (mean earnings conditional on positive earnings) among women in each experimental program group. We use the term "conditional mean difference" to refer to the result of subtracting the control group's conditional mean earnings from the (actual) treatment group's conditional mean earnings. We calculate a separate conditional mean difference for each subgroup-quarter cell.[27] For each control-group woman whose actual earnings are positive, we construct her synthetic treatment-group earnings by adding the appropriate conditional mean difference to her actual earnings. For control group women whose earnings equal zero, we assign a synthetic treatment-group earnings value of zero. As above, we then calculate sample quantiles of the synthetic earnings distribution. In doing so, though, we reweight observations so that the weighted share of women with zero synthetic treatment-group earnings equals the share of zeros in the actual treatment group.[28] We then calculate synthetic QTE by subtracting the actual control-group sample quantiles from these weighted synthetic sample quantiles.

This procedure imposes two key results in each subgroup-quarter cell. First, the weighted share of observations with zero earnings is equal in the actual and synthetic treatment group distributions. Second, among women with positive earnings, the conditional mean of earnings is equal in the actual and synthetic treatment group distributions. Therefore, any systematic differences in the actual and synthetic QTE must result from distributional differences distinct from either the share of women with positive earnings or the average level of earnings among those who are employed.

Given that no women would be caused by Jobs First assignment to reduce earnings from a positive level to zero, we show in the appendix that the participation-adjusted CTEWS model is equivalent to a model in which

1. there is a set of women whose potential outcome involves zero earnings regardless of program assignment;

2. there is a set of "extensive-margin women" women whose potential outcome is zero under control-group assignment and have nonzero potential earnings when assigned to the treatment group;

[27] As we discuss below, a given conditional mean difference need not be a consistent estimate of a well-defined average treatment effect.

[28] We discuss the weights in Appendix Section C.

3. there is a set of "intensive-margin" women whose potential outcomes are nonzero under both assignments;

4. the distribution of extensive-margin women's potential earnings under treatment-group assignment is the same as the distribution of intensive-margin women's potential earnings under treatment-group assignment; and

5. within each subgroup-quarter cell, all intensive-margin women have the same treatment effect.

The first three of these properties would hold with many different treatment effect models. The fourth and fifth properties jointly allow the algorithm discussed above to consistently estimate the synthetic treatment-group earnings distribution. The fourth property allows us to estimate extensive-margin women's earnings distribution under treatment-group assignment by using the conditional earnings distribution for women actually observed with positive earnings in the treatment group. If this property did not hold, then we would not be able to estimate extensive-margin women's earnings distribution under treatment-group assignment without more information. The fifth property implies that the conditional mean difference is sufficient to characterize the difference in the conditional earnings distributions, given positive earnings, for the two program groups.

As a general matter, we are skeptical that even this participation-adjusted CTEWS model is consistent with labor supply theory. This skepticism is based primarily on the fact that one would expect the treatment-group assignment distributions of both offered wages and desired labor supply to differ among extensive- and intensive-margin women. Put differently, it seems unlikely that women induced to work by Jobs First have the same offered wage distribution as women who would work under either AFDC or Jobs First assignment. Nonetheless, the participation-adjusted CTEWS model is much more realistic than the simpler CTEWS models that do not address the substantial shares of women with zero earnings.

We present results for these participation-adjusted synthetic QTE in Figure 7. Results based on education subgroups appear in the top left panel; results based on age of youngest child subgroups appear in the top right panel; results based on subgroups defined using earnings in the seventh quarter before random assignment appear in the bottom left panel; and those based on subgroups using the share of pre-random assignment quarters with earnings are in the bottom right panel. Overall, these figures show a much closer resemblance between the synthetic and actual QTE than do those in Figures 5–6.

The negative synthetic QTE at the very bottom of the distributions occur because there are some

women in the control group who have very low, positive earnings and are members of subgroups with negative conditional mean treatment effects. As a result, these women's synthetic treatment group earnings are negative, so they wind up at the very bottom of the synthetic treatment group earnings distribution. With the exception of this effect, both the actual and synthetic QTE equal to zero for nearly all of first 48 quantiles in all panels of Figure 7, a result that follows from the participation adjustment.

Over quantiles 50–80 or so, the synthetic QTE do a reasonably good job of replicating the shape of the actual QTE, though they fail to achieve the amplitude of the actual QTE. However, in every case the synthetic QTE fail to fully replicate the negative QTE at the top of the earnings distribution. Because this feature of the actual QTE is clearly predicted by labor supply theory, this result is a potentially serious mark against even the most flexible mean impact estimator we consider.

In sum, adapting to distributional research the most common approach in the literature—estimating demographic subgroup-specific mean impacts—reveals almost none of underlying heterogeneity shown in the full sample QTE. Even allowing mean impacts to vary over time and adding the rich earnings and welfare history variables as additional subgroup variables does little. The most flexible CTEWS model (time varying impacts, participation adjustment) does considerably better. It is worth noting, though, that even this model implies behavioral assumptions that are inconsistent with (simple) static labor supply theory. Of course, merely looking at the figures does not tell us whether the distributions are more different than sampling variation would allow under the null hypothesis of distributional equivalence.

4.4 Testing for differences between actual and synthetic QTE

Next we turn to a discussion of the testing, in which we use a test statistic suggested by Chernozhukov & Fernandez-Val (2005, henceforth, CFV), which is similar to the Kolmogorov-Smirnov statistic. This statistic is the scaled maximum over all $q \in \{1, 2, \ldots, 99\}$ of the absolute value of the weighted difference between the quantiles of the synthetic treatment group and the actual treatment group. To be precise, let $\widehat{y}_{q1}$ be the sample q^{th} quantile of the actual treatment group earnings distribution, and let $\widehat{\widehat{y}}_{q1}$ be the sample q^{th} quantile of the synthetic treatment group earnings distribution. Finally, let $\widehat{V}(q)$ be the estimated variance of the difference in these sample quantiles. The test statistic is

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

CFV develop a bootstrap procedure to estimate the distribution of $\widehat{S}$ under the null hypothesis that the two distributions are equal. We provide details of this procedure in Appendix Section C; the key statistic emerging from this bootstrap is a level-α critical value for the statistic $\widehat{S}$.

Table 4 presents the test statistics with critical values in parentheses for tests of size 0.05. Each row presents the test statistics and critical value for three different null hypotheses. Column 1 presents test statistics for the null hypothesis that constrains subgroup-specific treatment effects to be constant across all seven quarters; this column thus involves tests of the null that models of the type generating the left panels of Figure 5 are correct. Column 2 present test statistics for the null hypothesis that allows treatment effects to vary within both subgroup and quarter; this column thus involves tests of the null that models of the type generating the right panels of Figure 5 and all of Figure 6 are correct. In column 3, we generalize the null hypothesis to allow for a participation adjustment, as in Figure 7.

The statistical tests in Table 4 generally reject the null hypothesis that the visual differences in the figures above are due only to sampling variation. When we do not adjust for Jobs First participation effects, every subgroup's synthetic treatment group distribution is statistically different from the actual treatment group distribution at the 5% level. We note that this result holds even when we define subgroups based on the interaction of any two of the subgroups used above (not shown in tables).

The column 3 results show that even with the participation adjustment, the synthetic and actual treatment group earnings distributions differ significantly when we define subgroups using the demographic variables, and welfare history. On the other hand, we cannot reject equality of the synthetic and actual treatment group earnings distributions when we adjust for participation effects and define subgroups using either predicted wages or earnings history (whether measured using earnings levels in the seventh quarter before random assignment or using the share of pre-random assignment quarters with non-zero earnings). This result is in line with the visual similarity between the actual and synthetic QTE for these subgroups.

5 Conclusion

Program evaluators are often interested in exploring treatment effect heterogeneity. In practice, this is most commonly done by exploring differences across subgroups in mean impacts (e.g., Angrist (2004)). Subgroup choices are typically made based on either economic theory or previous empirical literature. In our own previous research, Bitler et al. (2006), we argue that focusing on mean impacts obscures considerable treatment effect heterogeneity. We illustrate this heterogeneity using quantile treatment effects (QTE). An interesting question concerns whether our previous results arise because of fundamental within-group treatment effect heterogeneity, or whether these results are the simply due to variation across subgroups in constant treatment effects. We dub the latter hypothesis the CTEWS, or constant treatment effects within subgroups, model of program impacts.

In this paper, we use data from Connecticut's Jobs First welfare reform experiment and take several approaches to assessing variations on the CTEWS model. First, we present Jobs First mean impacts across subgroups likely to proxy for wage opportunities (e.g., education and prior earnings) or preferences or fixed costs of work (e.g., number and age of children, marital history). We find that mean effects are not so different across subgroups defined by education, age or number of children, or other demographic measures. This result suggests that variation in mean impacts across categories within these subgroups is unlikely to explain much distributional heterogeneity. On the other hand, there are significant differences in the mean impacts of reform across subgroups defined by welfare and earnings history over the seven quarters before random assignment, as well as wages we predict via a Heckman selection model. These findings suggest that variables more closely associated with earnings capacity and the probability of welfare participation may be useful in explaining distributional heterogeneity.

Second, we estimate QTE within various subgroups. This approach offers a simple test of the null hypothesis that heterogeneity in QTE for the pooled sample can be explained via CTEWS models. If there is no within-group treatment effect heterogeneity, then quantile treatment effect profiles should be flat within subgroups. Our subgroup-specific results reject the CTEWS model resoundingly: we find considerable heterogeneity in QTE within subgroups. Interestingly, the QTE are surprisingly similar within demographic-group categories. For example, the QTE for women with no children under 5 are nearly identical to those for women whose youngest child is not kindergarten-aged. By contrast, QTE within groups defined by earnings history look quite different (as do those within education group to a lesser extent). Again, this result suggests that

variables related to earnings capacity and the likelihood of welfare use may be able to explain more distributional heterogeneity than can conventional demographic variables.

Third, we examine the distribution of various subgroup characteristics at different percentiles in the pooled earnings distribution. We find that for both the treatment and control groups, the within-percentile distribution of demographic subgroups is very similar across earnings percentiles. While this result might seem surprising, it is the mirror image of the subgroup-specific QTE results discussed above: if subgroups are actually randomly allocated across the earnings distribution, then subgroup-specific QTE and pooled QTE must be the same.

Finally, we directly test the appropriateness of CTEWS models by using subgroup-specific mean treatment effects to impose various CTEWS models on the control group earnings distribution. If the CTEWS model is correct, then the synthetic pooled earnings distribution that results should equal the actual pooled treatment group's earnings distribution. Since QTE are calculated by subtracting control group quantiles from treatment group quantiles, it is convenient to assess the CTEWS model by comparing actual and synthetic pooled QTE. Given the large share of quarterly observations with zero earnings, it is not surprising that the synthetic and actual QTE differ substantially in both economic and statistical terms unless we adjust for participation effects. Even when we adjust for participation, however, the synthetic and actual QTE are significantly different for subgroups defined using demographic variables like education, age of youngest child, and marital status. Synthetic earnings distributions based on predicted wages and on earnings and welfare history do somewhat better visually, and we cannot reject that the underlying synthetic and actual population earnings distributions are equal. The conclusions are largely unchanged when we use two way interactions of the subgroups; the only ones where we fail to reject that the synthetic and actual population distributions are the same are when one of the included subgroups uses earnings history before random assignment.

We draw three key conclusions from this work. First, even within subgroups that are commonly supposed to be quite diverse, within-group treatment effect heterogeneity is large by comparison to across-group heterogeneity. This finding is particularly true for subgroups based on conventional demographic variables. Second, members of commonly used demographic groups are relatively evenly dispersed across the earnings distributions. This result implies that within-group heterogeneity must play an important role in explaining treatment effect heterogeneity in the pooled population. Third, in trying to replicate the actual QTE, it is critical to adjust for the Jobs First's program's participation effects. Fourth, the groups most useful for replicating pooled-sample heterogeneity

are those based on predicted wages and earnings prior to random assignment. The variables used to create these groups are sometimes unavailable, in both experimental and non-experimental settings.

Taken together, these results suggest that even estimating subgroup-specific mean effects for a wide range of subgroups may not reveal all important treatment effect heterogeneity. We note finally that there is nothing special about our use of data from a welfare reform experiment. The methods used here can be applied in many program-evaluation settings to evaluate the importance of fundamental treatment effect heterogeneity.

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.

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

Bitler, M. P., Gelbach, J. B. & Hoynes, H. W. (2008), 'Distributional impacts of the Self-Sufficiency Project', *Journal of Public Economics* **92**(3–4), 748–765.

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.

Card, D. & Hyslop, D. (2005), 'Estimating the effects of a time-limited earnings subsidy for welfare-leavers', *Econometrica* **73**(6), 1723–1770.

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. (2008), 'Nonparametric tests for treatment effect heterogeneity', *Review of Economics and Statistics* **90**(3), 389–406.

Crump, R., Hotz, V. J., Imbens, G. & Mitnik, O. (2009), 'Dealing with limited overlap in estimation of average treatment effects', *Biometrika* **96**(1), 187–199.

Djebbari, H. & Smith, J. (2008), 'Heterogenous program impacts of the PROGRESA program', *Journal of Econometrics* **145**(1–2), 64–80.

Firpo, S. (2007), 'Efficient semiparametric estimation of quantile treatment effects', *Econometrica* **75**(1), 259–276.

Friedlander, D. & Robins, P. K. (1997), 'The distributional impacts of social programs', *Evaluation Review* **21**(5), 531–553.

Gelbach, J. B. (2005), Inference for sample quantiles with discrete data. Available at `http://glue.umd.edu/~gelbach/papers/working-papers.html`.

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.

Hall, P. (1988), 'Theoretical comparison of bootstrap confidence intervals', *The Annals of Statistics* **16**(3), 927–953.

Heckman, J. J., Smith, J. & Clements, N. (1997), 'Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts', *Review of Economic Studies* **64**, 487–535.

Heckman, J. J. & Vytlacil, E. J. (1999), 'Local instrumental variables and latent variable models for identifying and bounding treatment effects', *Proceedings of the National Academies of Science* **96**, 4730–4734.

Heckman, J. J. & Vytlacil, E. J. (2001), Local instrumental variables, *in* J. Heckman & E. Leamer, eds, 'Nonlinear Statistical Inference: Essays in Honor of Takesha Ameniya', North Holland, Amsterdam.

Hotz, V. J., Imbens, G. & Klerman, J. (2006), 'Evaluating the differential effects of alternative welfare-to-work training components: A re-analysis of the California GAIN program', *Journal of Labor Economics* **24**(3), 521–566.

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

Koenker, R. & Bassett, G. (1978), 'Regression quantiles', *Econometrica* **46**, 33–50.

Koenker, R. & Bilias, Y. (2001), 'Quantile regression for duration data: A reappraisal of the Pennsylvania reemployment bonus experiments', *Empirical Economics* **26**(1), 199–220.

Machado, J. & Mata, J. (2005), 'Counterfactual decompositions of changes in wage distributions using quantile regression', *Journal of Applied Econometrics* **20**, 445–465.

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

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge University Press, New York.

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

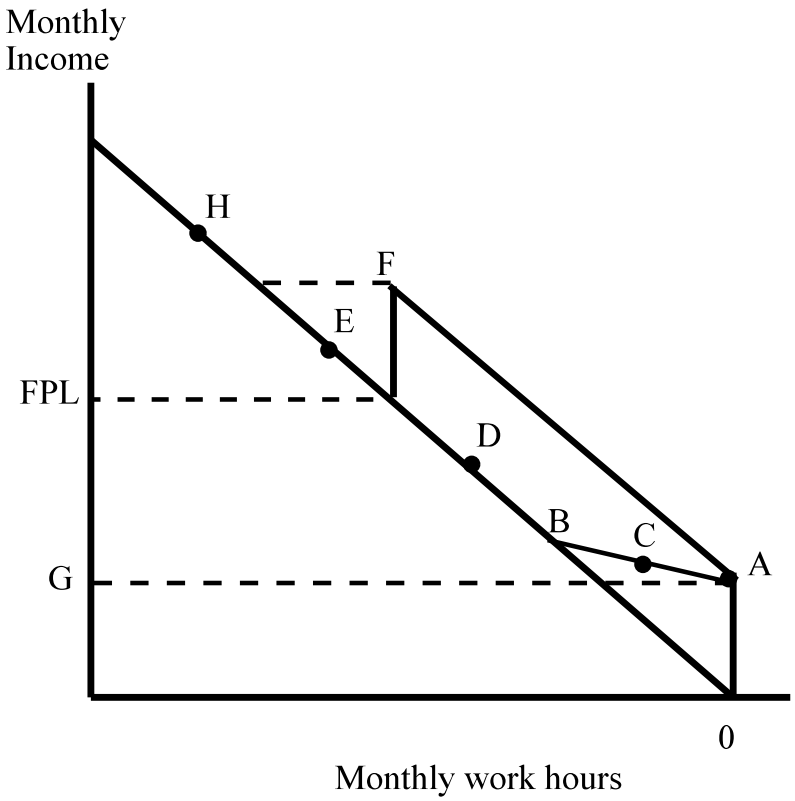
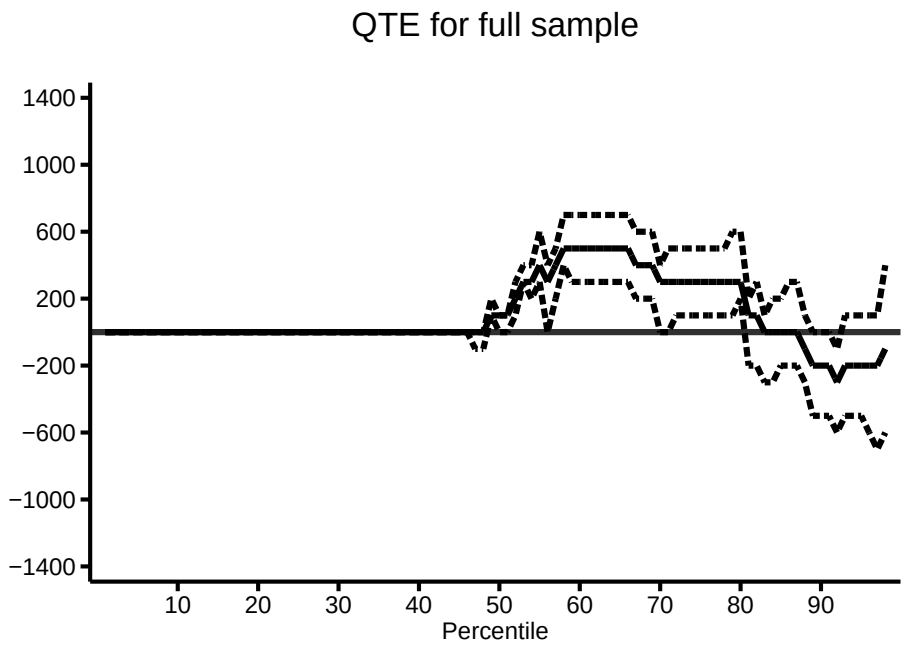
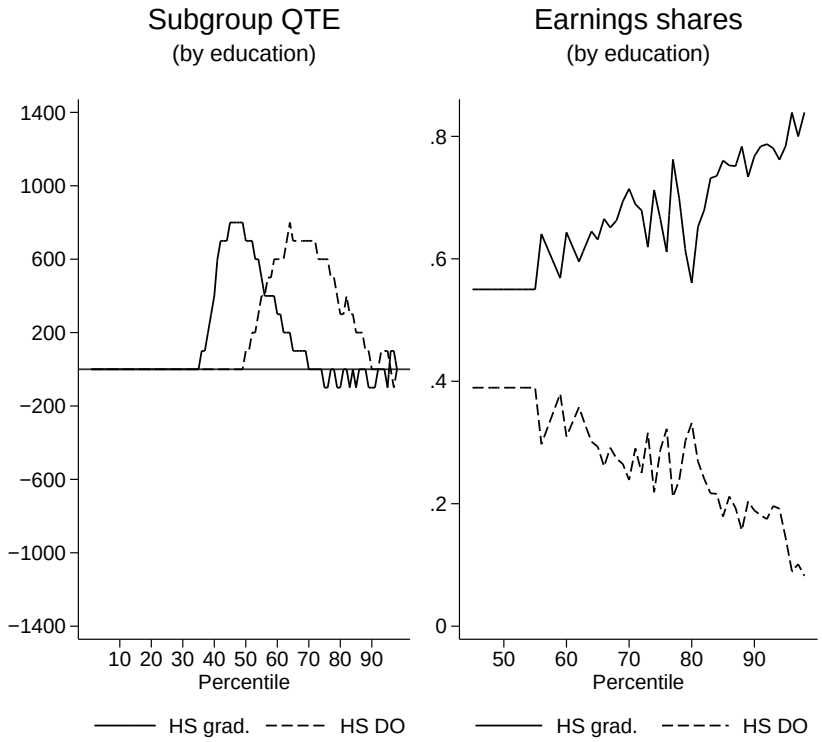


Figure 2: QTE for full sample for earnings



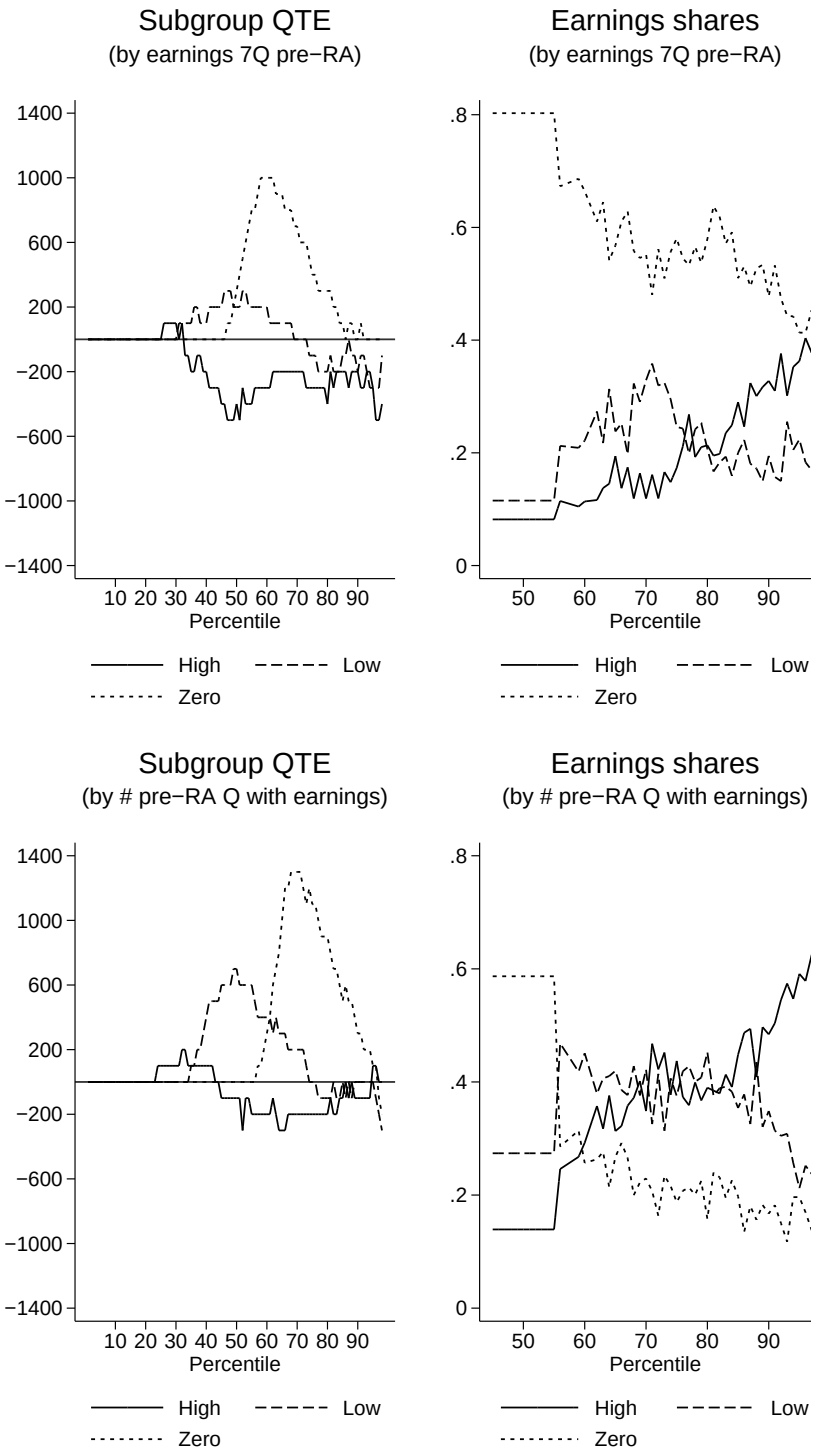
Notes: Figure shows quantile treatment effects on earnings for centiles 1–98. Solid line is QTE, dashed lines are pointwise 95% confidence intervals (calculated using the percentile method). QTE calculated using inverse propensity score weighting.

Figure 3: QTE and quantile-specific subgroup shares in the control group, using education to define subgroups



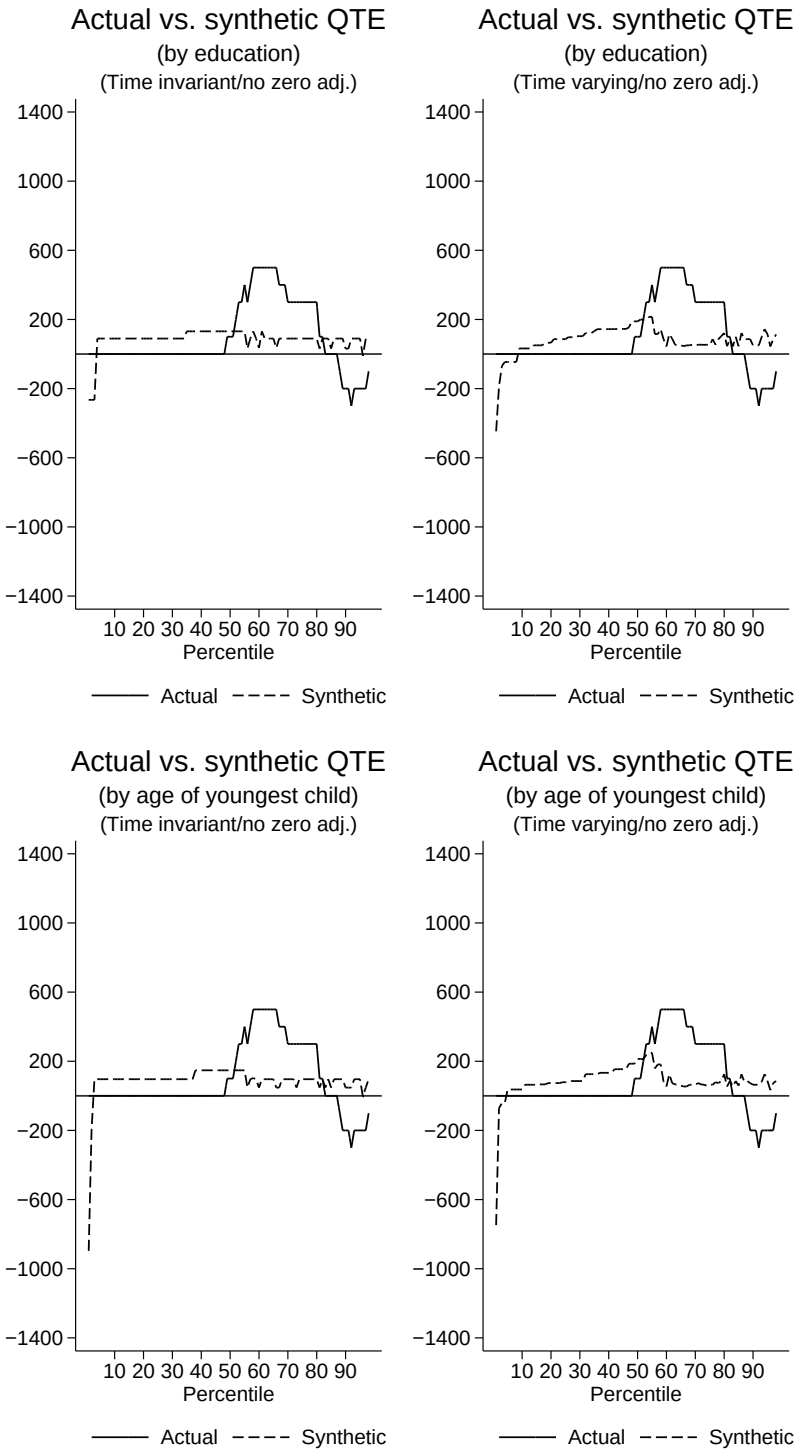
Notes: Left figure shows QTE by education subgroups. Right figure shows the share of women who are high school graduates and the share who are high school dropouts, for centiles 45–98 of the control group earnings distribution. Because earnings are zero for all centiles below 45, there is no variation in the group shares within these centiles, so we omit them. QTE calculated using inverse propensity score weighting.

Figure 4: QTE and quantile-specific subgroup shares in the control group, using earnings prior to random assignment to define subgroups



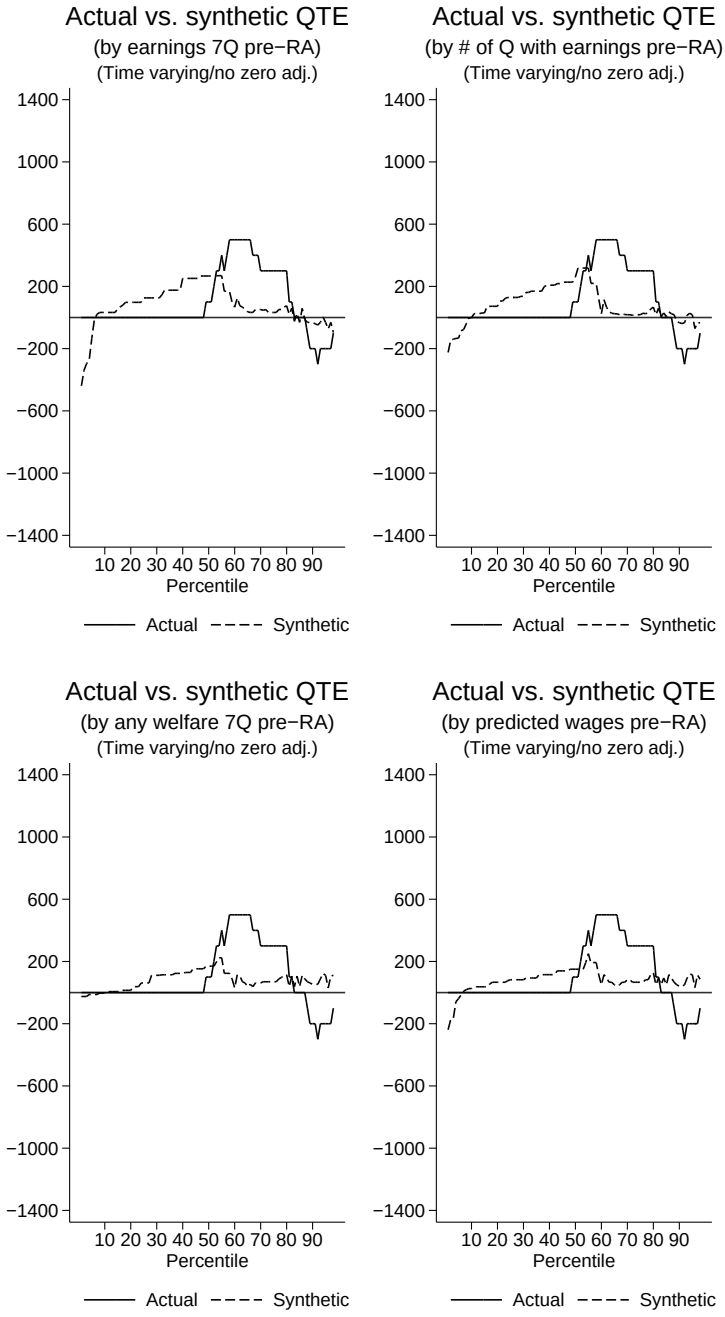
Notes: Top left figure shows QTE for those with high, low, and zero earnings over the seven quarters before random assignment. Top right figure shows these subgroups' sample shares at each of centiles 45–98 in the control group earnings distribution. Because earnings are zero for all centiles below 45, there is no variation in the group shares within these centiles, so we omit them. Bottom figures are analogous to top ones, except that subgroups are defined using the share of quarters with earnings before random assignment.QTE calculated using inverse propensity score weighting.

Figure 5: Actual and synthetic QTE with no participation adjustment,
subgroups based on education (top graphs) and age of youngest child (bottom graphs)



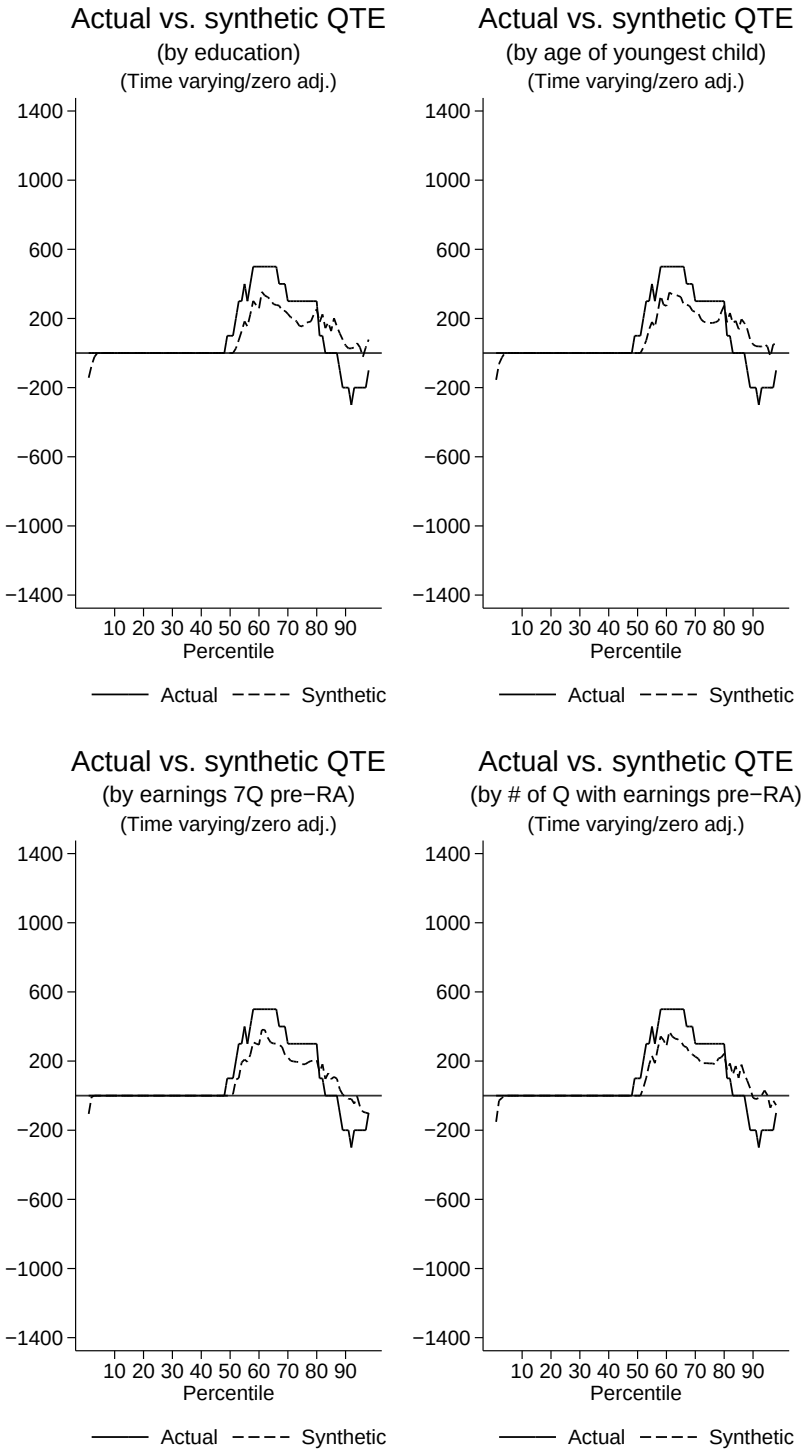
Notes: In each figure, the solid line plots the actual QTE, and the dashed line plots the synthetic QTE. In the top graphs, subgroups are based on education. In the bottom graphs, subgroups are based on age of youngest child. In the left figures, sythetic earnings are calculated under the constraint that subgroup-specific treatment effects are constant across quarter. In the right figures, we allow the subgroup-specific treatment effects to vary across quarters. We do not adjust for partipation effects in any of the graphs above. QTE calculated using inverse propensity score weighting.

Figure 6: Actual and synthetic QTE with no participation adjustment,
subgroups based on earnings before random assignment (top graphs) and predicted wages (bottom graphs)



Notes: In each figure, the solid line plots the actual QTE and the dashed line plots the synthetic QTE. In the top graphs, subgroups are based on earnings in the seventh quarter before random assignment. In the bottom graphs, subgroups are based on predicted wages. In the left figures, synthetic earnings are calculated under the constraint that subgroup-specific treatment effects are constant across quarters. In the right figures, we allow the subgroup-specific treatment effects to vary across quarters. We do not adjust for participation effects in any of the graphs above.

Figure 7: Actual and synthetic QTE with participation adjustment, subgroups based on demographics and earnings before random assignment



Notes: In each graph, the solid line plots the actual QTE and the dashed line plots the synthetic QTE. In all graphs, we allow for time-varying program effects on conditional mean earnings and also adjust for participation effects using the re-weighting method discussed in the text. In the top left graph, subgroups are based on education. In the top right graph, subgroups are based on age of youngest child. In the bottom left graph, subgroups are based on earnings in the seventh quarter before random assignment. In the bottom right graph, subgroups are based on the share of pre-random assignment quarters with positive earnings.

Table 1: Characteristics of experimental sample, pre-random assignment

	Levels (unweighted)		Differences	
	Jobs First	AFDC	Unweighted	Weighted
White	0.362	0.348	0.014	0.001
Black	0.368	0.371	-0.003	-0.000
Hispanic	0.207	0.216	-0.009	-0.001
HS dropout	0.331	0.313	0.018	-0.000
HS diploma/GED	0.550	0.566	-0.016	0.001
More than HS diploma	0.063	0.058	0.004	0.000
At least HS diploma/GED	0.613	0.624	-0.011	0.001
More than two children	0.227	0.206	0.021*	-0.000
At most two children	0.484	0.472	0.012	0.003
Youngest child 5 or younger	0.618	0.634	-0.017	-0.014
Youngest child 6 or older	0.350	0.331	0.019	0.014
Never married	0.624	0.631	-0.007	-0.000
Div./wid./sep./living apart	0.317	0.312	0.005	0.000
Div./wid./sep./married	0.330	0.324	0.006	0.000
Mother younger than 25	0.289	0.297	-0.007	-0.000
Mother aged 25–34	0.410	0.418	-0.007	0.000
Mother older than 34	0.301	0.286	0.015	0.000
Recipient (stock) sample	0.624	0.593	0.031**	-0.001
Any AFDC 7^{th} quarter pre-RA	0.548	0.528	0.020	-0.000
Cash welfare (AFDC) pre-RA	891	835	56**	-1
# of pre-RA quarters with any cash welfare	0.573	0.544	0.029**	-0.001
Earnings level, pre-RA	683	796	-113***	-1
Earnings 7^{th} quarter before random assignment				
None	0.700	0.674	0.026*	0.000
Low	0.149	0.165	-0.016	-0.012
High	0.151	0.161	-0.010	0.011
Earnings level	682	781	-99**	-2
Number of quarters with positive earnings pre-random assignment				
Zero	0.436	0.403	0.033**	-0.002
Low	0.316	0.316	-0.001	0.002
High	0.249	0.281	-0.033**	-0.000
Share of pre-RA quarters with any earnings	0.327	0.357	-0.030***	0.000
Predicted pre-random assignment wages				
Low	0.682	0.698	-0.010	-0.008
Medium	0.163	0.159	0.011	0.007
High	0.155	0.143	-0.000	0.001
Level of predicted wages	8.90	8.86	0.037	0.039

Notes: For difference estimates, ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels; standard deviations and standard errors are available on request. Predicted wages are constructed by using estimates for our Heckman selection model estimated on CPS data together with baseline demographic characteristics from the experimental data. Baseline data are missing for a small number of observations for some variables. Statistics for earnings and cash welfare are based on up to seven quarters of pre-random assignment data. First three columns present unweighted statistics, fourth column presents inverse propensity score weighted differences. Inference for differences of pre-random assignment variables that vary over time is based on standard errors that are robust to residuals correlated across time at the individual level.

Table 2: Mean differences in earnings by demographic subgroups

Subgroup	Mean difference	95% CI	Control group mean	N_C/share	N_T/share
All	82	[39, 124]	1113	16,849	16,772
By education of case head:					
No HS degree/GED	131	[77, 185]	646	0.31	0.33
At least HS/GED	90	[36, 143]	1323	0.62	0.61
F-statistic [p-value]	0.83	[0.3634]			
By whether youngest child is ≤ 5*:*					
Youngest child ≤ 5	96	[39, 144]	1043	0.63	0.62
Youngest child ≥ 6	148	[78, 218]	1150	0.33	0.35
F-statistic [p-value]	1.28	[0.2582]			
By number of children in case:					
2 or more	155	[99, 211]	1038	0.47	0.48
1 or pregnant	78	[22, 134]	1122	0.49	0.48
F-statistic [p-value]	3.13	[0.0770]			
By marital status of case head:					
Never married	86	[40, 132]	1037	0.63	0.62
Ever married	146	[71, 222]	1188	0.32	0.33
F-statistic [p-value]	1.72	[0.1896]			

Results in first column involve treatment-control differences in average quarterly earnings during the first seven quarters after random assignment. F-statistic and p-value in each panel are from tests that the mean treatment effects are the same across the relevant subgroups. Other entries in the first column are mean quarterly earnings for women assigned to Jobs First minus mean quarterly earnings for women assigned to AFDC, for the indicated subgroup. There are 4803 women in the sample and all means are estimated using inverse propensity score weighting. Columns 2—5 contain 95 percent CIs, the control group mean, and the number of observations in the treatment and control groups (top panel) or share of observations with that group (other panels) for the subgroup in that row. At the bottom of each panel, the F-statistic and p-value for the test that the mean treatment effects are the same across the subgroups are shown. Note that the F-statistics in the table exclude the values for a small number of observations missing some of the demographic characteristics (also missing from the tables). The relevant F-statistics [p-values] for the tests which include the missing data categories are 8.98 [0.0001] for education, 37.0 [0.0000] for age of youngest child being less than 5, 38.0 [0.0000] for the number of children in the case, and 14.4 [0.0000] for marital status of the case head.

Table 3: Mean differences in earnings by earnings/welfare history/predicted wage subgroup

Subgroup	Mean difference	95% CI	Control group mean	N_C/share	N_T/share
All	82	[39, 124]	1113	16,849	16,772
By whether earnings are zero 7^{th} quarter before RA:					
Yes	174	[135, 214]	751	0.67	0.70
No	-119	[-217, -22]	1906	0.33	0.30
F-statistic [p-value]	42.7	[0.0000]			
By level of earnings 7^{th} quarter before RA:					
Zero	174	[135, 214]	751	0.67	0.70
Low	43	[-48 , 134]	1325	0.16	0.15
High	-358	[-526, -190]	2536	0.16	0.15
F-statistic [p-value]	32.3	[0.0000]			
By share of quarters with any earnings before RA:					
Zero	206	[164, 249]	451	0.40	0.44
Low	96	[32 , 161]	1098	0.32	0.32
High	-134	[-245, -24]	2173	0.28	0.25
F-statistic [p-value]	20.6	[0.0000]			
By whether on AFDC 7^{th} quarter before RA:					
Yes	115	[60, 170]	954	0.53	0.55
No	43	[-21, 107]	1298	0.47	0.45
F-statistic [p-value]	2.83	[0.0927]			
By predicted wages before RA:					
Low	80	[36, 124]	1045	0.69	0.68
Medium	145	[37 , 254]	1218	0.15	0.16
High	7	[-144, 159]	1322	0.16	0.15
F-statistic [p-value]	5.64	[0.0007]			

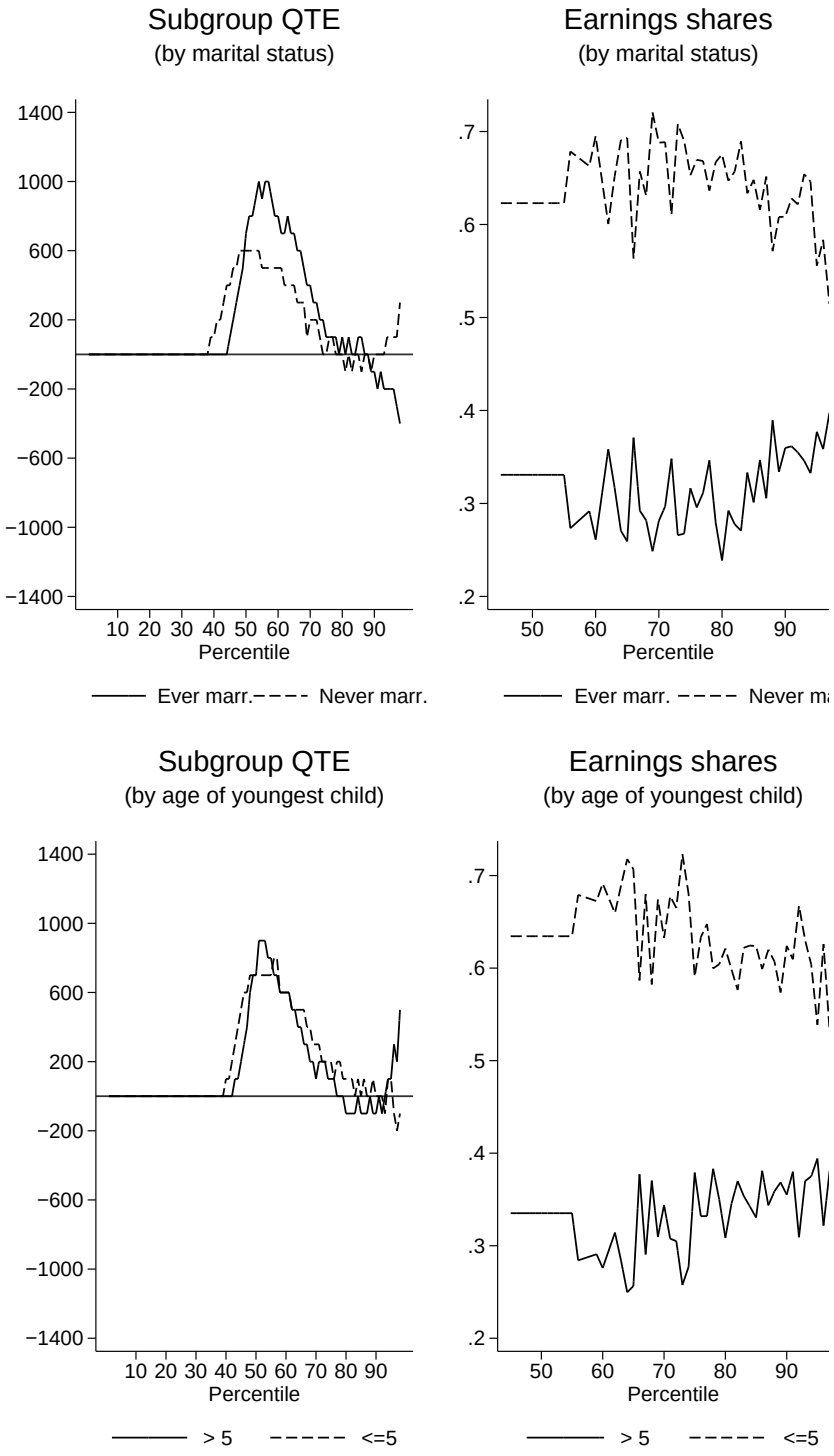
Results in first column involve treatment-control differences in average quarterly earnings during the first seven quarters after random assignment. F-statistic and p-value in each panel are from tests that the mean treatment effects are the same across the relevant subgroups. Other entries in the first column are mean quarterly earnings for women assigned to Jobs First minus mean quarterly earnings for women assigned to AFDC, for the indicated subgroup. There are 4803 women in the sample and all means are estimated using inverse propensity score weighting. Columns 2—5 contain 95 percent CIs, the control group mean, and the number of observations in the treatment and control groups (top panel) or share of observations with that group (other panels) for the subgroup in that row. At the bottom of each panel, the F-statistic and p-value for the test that the mean treatment effects are the same across the subgroups are shown.

Table 4: Tests of whether QTEs deviate from those calculated by adding mean TE within subgroup to the control group distribution

	Subgroup mean adjustments:		
Subgroup	None	Time	Zeros & Time
Full sample	423**	502**	277**
	(245)	(233)	(270)
Education	485**	505**	299**
	(258)	(253)	(280)
Age of youngest child	488**	478**	331**
	(246)	(256)	(274)
Marital status	492**	493**	325**
	(260)	(269)	(280)
Earnings level 7^{th} Q pre-RA	469**	513**	229
	(242)	(248)	(271)
Share of quarters with earnings pre-RA	512**	520**	225
	(241)	(261)	(270)
Welfare receipt 7^{th} Q pre-RA	481**	505**	295**
	(240)	(253)	(270)
Predicted wages pre-RA	472**	440**	265
	(261)	(271)	(270)

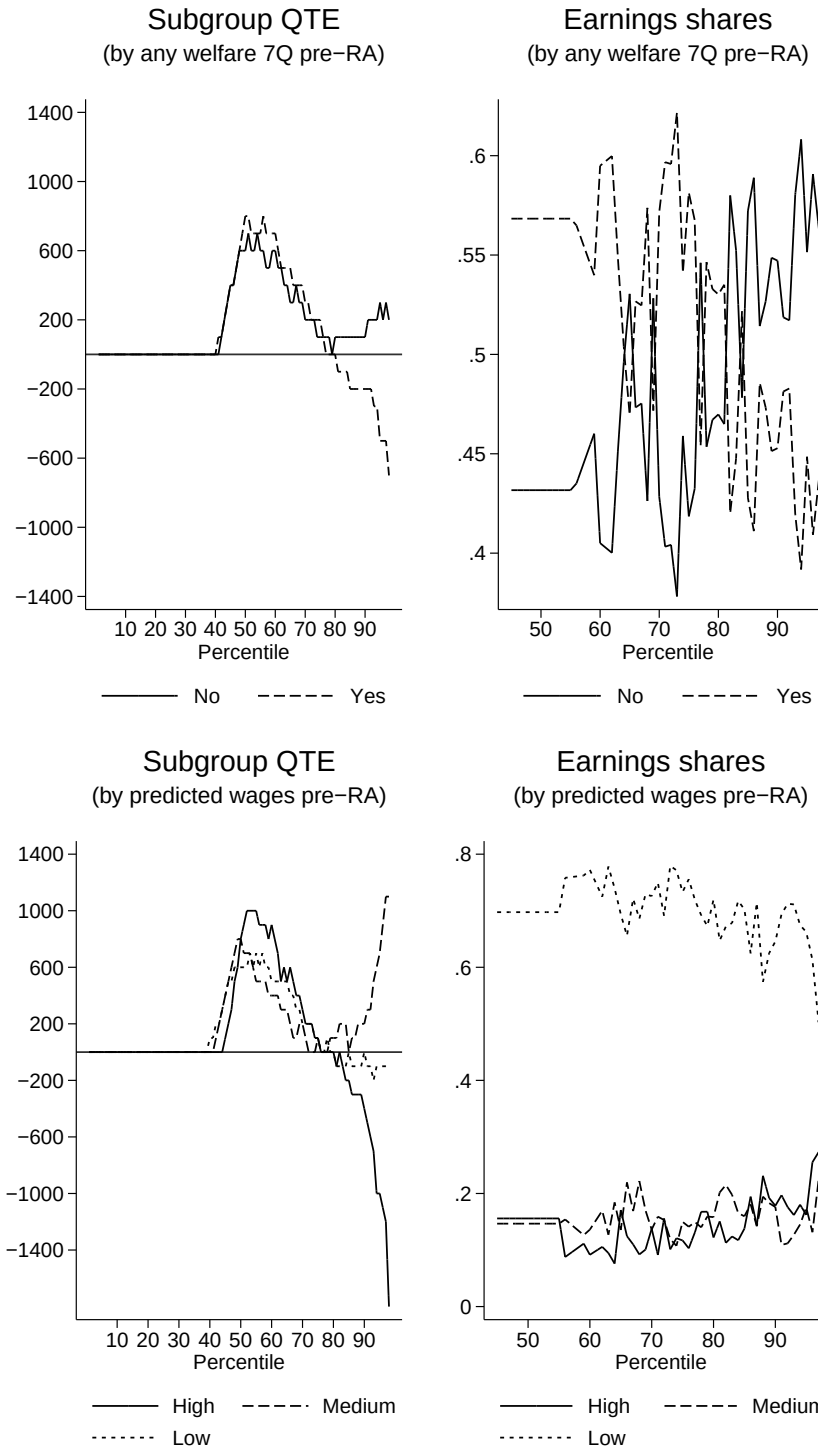
Notes: Top entry in each cell is $\widehat{S}$ test statistic of null that synthetic and actual QTE distributions are equal (modified K-S test). First-column synthetic earnings computed with the constraint that the subgroup-specific treatment effects are constant across time. Second-column synthetic earnings computed with subgroup-specific treatment effects allowed to vary across time. Neither of the first two columns adjusts for participation effects. Third column allows both time-varying conditional mean earnings effects and adjusts for participation effects. ** denotes statistical significance at the 5% level. Subgroup-specific means are estimated using propensity score weighting. For more details, see text.

Appendix Figure 1: QTE and quantile-specific subgroup shares in the control group, marital status and age of youngest child to define subgroups



Notes: Top left figure shows QTE for the never married and ever married. Top right figure shows these subgroups' sample shares at each of centiles 45–98 in the control group earnings distribution. Because earnings are zero for all centiles below 45, there is no variation in the group shares within these centiles, so we omit them. Bottom figures are analogous to top ones, except that subgroups are defined using age of youngest child.QTE calculated using inverse propensity score weighting.

Appendix Figure 2: QTE and quantile-specific subgroup shares in the control group, using AFDC participation and predicted wages prior to random assignment



Notes: Top left figure shows QTE for those with any AFDC receipt or no AFDC receipt 7 quarters before random assignment. Top right figure shows these subgroups' sample shares at each of centiles 45–98 in the control group earnings distribution. Because earnings are zero for all centiles below 45, there is no variation in the group shares within these centiles, so we omit them. Bottom figures are analogous to top ones, except that subgroups are defined using predicted wages prior to random assignment. QTE calculated using inverse propensity score weighting.