What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments

Marianne P. Bitler
RAND Corporation and IZA

Jonah B. Gelbach
University of Maryland

Hilary W. Hoynes
University of California, Davis and NBER*

First version: January 2003
This version: November 3, 2003

Abstract

Labor supply theory predicts systematic heterogeneity in the impact of recent welfare reforms on earnings, transfers, and income. Yet most welfare reform research focuses on mean impacts. We investigate the importance of heterogeneity using random-assignment data from Connecticut's Jobs First waiver, which features key elements of post-1996 welfare programs. Estimated quantile treatment effects exhibit the substantial heterogeneity predicted by labor supply theory. Thus mean impacts miss a great deal. Looking separately at dropouts and other women does not improve the performance of mean impacts. Evaluating Jobs First relative to AFDC using a class of social welfare functions, we find that Jobs First's performance depends on the degree of inequality aversion, the relative valuation of earnings and transfers, and whether one accounts for Jobs First's greater costs. We conclude that welfare reform's effects are likely both more varied and more extensive than has been recognized.

*Correspondence to Hoynes at UC Davis, Department of Economics, 1152 Social Sciences and Humanities Building, One Shields Avenue, Davis, CA 95616-8578, phone (530) 752-3226, fax (530) 752-9382, or hwhoynes@ucdavis.edu; Gelbach at gelbach@glue.umd.edu; or Bitler at bitler@rand.org. Bitler gratefully acknowledges the financial support of the National Institute of Child Health and Human Development and the National Institute on Aging. This work has not been formally reviewed or edited. The views and conclusions are those of the authors and do not necessarily represent those of the RAND Corporation. We are very grateful to MDRC for providing the public access to the experimental data used here. 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 would also like to thank Mary Daly, Guido Imbens, Lorien Rice, and Jeff Smith for helpful conversations, as well as seminar participants from Berkeley, Chicago–Harris School, Cornell, Davis, GW, Johns Hopkins, the IRP summer workshop, Maryland, the NBER, PPIC, Syracuse, and RAND.

1 Introduction

Several years have now passed since the elimination of Aid to Families with Dependent Children (AFDC), the principal U.S. cash assistance program for six decades. In 1996, enactment of the Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA) required all 50 states to replace AFDC with a Temporary Assistance for Needy Families (TANF) program. State TANF programs differ from AFDC in many fundamental ways. Key examples include lifetime limits on program participation, enhanced work incentives through expanded earnings disregards, stringent work requirements, and financial sanctions for failure to comply with these requirements. A critical element in evaluating this dramatic policy change is measuring the impact of TANF on earnings and income. In this paper, we focus on heterogeneity of the effects of these reforms. This should be a key issue for evaluating welfare reform, because labor supply theory makes very strong predictions concerning heterogeneity in both the sign and magnitude of labor supply responses to recent reforms. As a consequence, mean impacts will tend to average together positive and negative labor supply responses, possibly obscuring the extent of welfare reform's effects.

Despite the theoretical indeterminacy of mean impacts, most existing research on welfare reform and earnings takes that approach.[1] Several studies use nonexperimental data from the Current Population Survey (CPS) to examine the impact of PRWORA, and the state waivers that preceded it, on income. Evidence from these studies is mixed. For example, Moffitt (1999) finds no impact of waivers on family income, while Grogger (Forthcoming) finds that welfare reform increased mean income for female heads of household. Further, experimental studies examining state reforms implemented before 1996 via TANF-like waivers suggest that generous increases in the earnings disregards are important for generating mean income gains. However, these gains disappear after time limits (Bloom & Michalopoulos (2001), Grogger et al. (2002)).

In this paper, we shift attention away from mean impacts. Instead we allow for heterogeneous impacts of welfare reform by estimating quantile treatment effects (QTE). We use public-use data files from the Manpower Demonstration and Research Corporation's (MDRC) experimental evaluation of Connecticut's Jobs First waiver. The evaluation was conducted over a five-year period

[1] The welfare reform literature that has developed in the last several years is enormous. We confine our discussion of this literature to a few papers having particular relevance to our study of income and heterogeneity. For comprehensive summaries of this research, see the excellent reviews by Blank (2002), Moffitt (2002), and Grogger, Karoly & Klerman (2002).

beginning in 1996. Experimental data also are available for waiver demonstrations in other states. We focus on Jobs First because of its radical increase in the earnings disregard and its very short (21-month) time limit. Along these lines, Jobs First may be viewed as something of a supercharged version of many states' TANF programs. Thus it provides an excellent opportunity to study the impacts of these key reforms.

Our choice to use experimental data and methods is not incidental. As discussed in Blank (2002) and formalized in Bitler, Gelbach & Hoynes (2003, (Papers and Proceedings)), identifying the impact of TANF using nonexperimental methods is difficult given that (*i*) TANF was implemented in all states within a very short period, and (*ii*) the implementation took place during the strongest economic expansion in decades. Since the relevance of many of our findings lies in the ability of our techniques to detect heterogeneous treatment effects, we believe it is critical that our results not depend importantly on nuisance issues related to selection bias. To this end, experimental data provides a setting where identification is clear and essentially incontrovertible.

Our emphasis on heterogeneity is motivated partly because policymakers and researchers care directly about it. For example, in a recent Joint Center for Poverty Research newsletter entitled "What Policymakers Want to Know," Cabrera & Evans (2000) ask "What is the variability of response to welfare reform among families?... Typically, research findings are reported in terms of the average response of the welfare reform population with respect to some behavior or status of interest. This focus diverts attention from the subgroup of families that might be struggling, even when most are not." Such concerns have led a small number of authors to consider distributional concerns. Schoeni & Blank (2003 (Papers and Proceedings)) compare the full distribution of the income-to-needs ratio before and after TANF, finding increases at all but the very lowest percentiles. However, as the authors note, their simple before-and-after methods cannot distinguish impacts of TANF from the effects of strong labor markets. With the exception of these results and those in a few other sources,[2] the most common approach to addressing distributional concerns is to estimate mean impacts for subgroups of the population (defined using education, race, and welfare and

[2]Schoeni & Blank (2000) compare the 20th and 50th percentiles of the CPS family income distribution before and after implementation of TANF. They find negative (but insignificant) impacts of TANF on the 20th percentile, and positive and significant impacts on the 50th percentile for a sample of women with less than a high school education. Some of the MDRC waiver evaluations (e.g., Bloom, Scrivener, Michalopoulos, Morris, Hendra, Adams-Ciardullo & Walter (2002) and Bloom, Kemple, Morris, Scrivener, Verma & Hendra (2000)) include estimates comparing the fraction of treatment and control group members with income in broadly defined categories. This approach, which is essentially a tabular form of histogram plots, is similar in spirit to the approach we take.

employment history) thought to be particularly at risk for welfare dependence.[3] Michalopoulos & Schwartz (2000) review 20 randomized experiments, concluding that "Although the programs did not increase [mean] income for most subgroups they also did not decrease [mean] income for most subgroups" (p. ES-10). Grogger et al. (2002) summarize both nonexperimental and experimental evidence concerning mean impacts as follows: "the effects of reform do not generally appear to be concentrated among any particular group of recipients" (p. 231).

The focus in the literature on mean impacts contrasts with the very strong predictions that labor supply theory makes concerning welfare reform and heterogeneity. Consider, for example, Figure 1, which shows a stylized budget constraint in income-leisure space before and after Jobs First (whose characteristics we describe in detail below). Jobs First dramatically increased the disregard for calculating welfare benefits. In the pre-reform AFDC program, benefits were reduced dollar-for-dollar with increased earnings, leading to the horizontal portion of the budget set (which corresponds to a 100% tax rate).[4] Under Jobs First, recipients retain their *entire* benefit payment (a 0% tax rate) for earnings up to the poverty line.

Labor supply theory makes clear predictions about the impacts of this reform. Women who would choose to participate in welfare and not work when they face AFDC rules will increase earnings, provided their wages exceed a threshold level. By contrast, some women who would not participate under AFDC rules will decrease earnings (to the poverty line or below) to become eligible for Jobs First. Thus, mean impacts will average together positive and negative treatment effects, obscuring the full range of effects welfare reform has had. Because recent welfare reforms yield such clear theoretical predictions regarding heterogeneous treatment effects, they provide ideal terrain for exploiting QTE methodology. To be sure, quantile treatment effects have been used in previous experimental evaluations. Examples of their use in evaluating the Job Training and Partnership Act include Heckman, Smith & Clements (1997) and Firpo (2003), while Friedlander & Robins (1997) estimate QTE in evaluating employment training in earlier welfare reform experiments. However, the source of heterogeneous treatment effects in these cases is difficult to identify, since they mostly

[3]For example, Schoeni & Blank (2000) find that welfare reforms led to increases (insignificant in the case of TANF) in mean family income for female dropouts in the CPS. Using similar data, Bennett, Lu & Song (2002) find that TANF is associated with reductions in the income-to-needs ratio for poor children who live with a single parent having less than a high school education.

[4]As we discuss in section 2.3 below, the effective benefit reduction rate under AFDC may be much less punitive than 100%; what matters here is that it is significantly more than 0%.

involve changes to training or job search assistance. Unlike such black-box reforms, theoretical predictions are clear in the present context.

Our empirical findings may be summarized with two broad conclusions. First, we find evidence of substantial heterogeneity in response to welfare reform. Second, the heterogeneity is broadly consistent with the predictions of labor supply theory. Contrary to much recent discussion among policymakers and researchers, under plausible assumptions our results suggest the possibility that welfare reforms reduced income for a nontrivial fraction of treatment group members, especially after time limits take effect. An important remaining methodological question is whether the essential features of our empirical findings could have been revealed using mean impact analysis on judiciously chosen subgroups. In the one important case we consider here, we find striking evidence that the answer is a resounding no: we find evidence that *intra-group* variation in quantile treatment effects greatly exceeds the *inter-group* variation in mean impacts.

Using mean impacts—which correspond to assuming risk/inequality neutrality—we find that under some circumstances, Jobs First would pass a typical cost-benefit test. But when a reform has heterogeneous effects, mean impacts are not sufficiently informative for an overall evaluation if policymakers are inequality-averse. Thus, we also provide an evaluation of Jobs First's effects on the income distribution using a familiar class of social welfare functions that allow for inequality aversion. For lower levels of inequality aversion, Jobs First still is found to be beneficial on net, but this conclusion is reversed for relatively high levels of inequality aversion. The evaluation conclusion depends importantly on how one accounts for Jobs First's administrative costs, and also to some extent on whether one accounts for the Earned Income Tax Credit (EITC).

The remainder of the paper is organized as follows. In section 2, we discuss the Jobs First program and its expected impacts on labor supply, showing that both time limits and disregards clearly were substantially implemented. In section 3, we present mean treatment effects for comparison purposes. We then report quantile treatment effects in 4. In section 5, we provide a social welfare function analysis to draw welfare conclusions. We conclude in section 6.

2 Jobs First

In this section we summarize Jobs First's programmatic features and their expected impacts, as well as the public-use experimental data provided by the Manpower Demonstration and Research Corporation (MDRC), which conducted the official Jobs First evaluation. In the next subsection, we provide an overview of the Jobs First assignment regime and public-use data. In subsections 2.2 and 2.3, we discuss the two most important policy changes in Jobs First: time limits and the expanded earnings disregard. We then discuss the remaining features of Jobs First in subsection 2.4. We conclude this section in subsection 2.5 by discussing Jobs First's labor supply implications.

2.1 Jobs First assignment and data

Table 1 summarizes the main features of Jobs First; further details beyond the discussion here are available in MDRC's final report on the evaluation [Bloom et al. (2002), henceforth "the final report"]. The table also includes a summary of the pre-existing AFDC program for comparison. The Jobs First waiver contained each of the key elements in PRWORA: time limits, changes to earnings disregards, work requirements, and financial sanctions. Under federal law, evaluations were required of states that implemented waiver programs. The Jobs First evaluation comprised all cases that were either ongoing or opened in the New Haven and Manchester welfare offices during the random assignment period, which took place between April 1996 and February 1997. The evaluation continued through the end of December 2000, after which point no further data were collected. MDRC's evaluation and public-use samples include data on a total of 4,803 cases. Of these, 2,396 were assigned to Jobs First and 2,407 to AFDC. Quarterly earnings data and monthly data on welfare and Food Stamps income are available for the two years preceding program assignment and for at least 4 years after assignment.[5] Demographic data—including information on educational attainment, age, race and ethnicity, marital status, and work history of the sample

[5]There are 16 quarterly observations on Connecticut earnings after random assignment for every sample member, with the exception of 30 people who entered the sample in January or February of 1997. Earnings data come from Connecticut's Unemployment Insurance (UI) system, so earnings not covered by the UI system are missed; fortunately the vast majority of employment is covered by the UI system. Data on Food Stamps and welfare payments come from Connecticut's Eligibility Management System (EMS), which warehouses information about welfare use. To preserve confidentiality, MDRC rounded several key variables before releasing the public-use data (e.g., they rounded quarterly earnings data to the nearest $100 and AFDC and Food Stamps payments to the nearest $50). For cases with true amounts between 0 and the lowest reported nonzero value (either $50 or $100), true values are rounded up, so that there are no false zeroes in the data.

member—are collected at an interview prior to random assignment.[6] During the evaluation, the rest of Connecticut's caseload was moved to Jobs First; only the control group continued under the AFDC rules. Table 2 provides a number of summary statistics for the Jobs First population, as well as for the national AFDC caseload in 1994.[7] The Jobs First sample mirrors the characteristics of the national sample, with exceptions reflecting the somewhat more disadvantaged recipients in one of the evaluation sites (New Haven). This is reflected by somewhat greater fractions of never married, Hispanic, and less-educated recipients compared to the national caseload average

2.2 The time limit

Jobs First's 21-month time limit is currently the shortest in the U.S. (U.S. House of Representatives Committee on Ways and Means (2000)). About 29% of the treatment group reached the time limit in the first 21 months of the evaluation period, and more than half reached the time limit within four years after random assignment (Bloom et al. (2002)). However, under certain circumstances, Jobs First caseworkers were empowered to provide both indefinite exemptions (as described in footnote 13) from the time limit and to provide 6-month extensions. According to the final report, in the spring of 1998, 26% of the statewide (not just the Jobs First) caseload was exempt from the time limit. This number rose to 49% by March 2001, though this appears to be the result largely of progressive exits from the caseload by more able (and time-limited) recipients. Extensions were granted to a non-exempt woman if her family income was below the applicable maximum benefit payment and she had made a good-faith effort to find and retain employment.[8] If no good-faith determination was made, then an extension was still possible if "there were circumstances beyond the recipient's control that prevent[ed] her from working."[9]

In light of these statistics, it is critical to show that the time limit policy has *de facto* relevance.

[6]MDRC also conducted a survey on a subset of the sample about three years after random assignment. These data, of which we only make slight use below, have been used by others to analyze impacts on other measures of family and child well-being.

[7]The estimates for the national caseload are constructed using March 1995 CPS data. The sample includes all women aged 16–54 who have an own child in the household and whose family was reported to receive positive AFDC income in the prior calendar year.

[8]Determination of good faith appears to have been somewhat complicated, often involving "extensive investigation, including talking with former employers, but staff reported that it often remains unclear why recipients left a job, reduced hours, and so on."

[9]Details regarding exemptions and extensions are derived primarily from Chapter 3 of the final report (which discusses implementation of the time limit) and from the overview in Chapter 1.

Figure 2 plots the monthly hazard rate for leaving welfare among women in our data who are still in their first welfare participation spell (we discuss sampling issues below). The relatively smooth series (dashed line) is the hazard rate for the control group, which faced the AFDC rules and, thus, no time limit. The more jagged series (solid line) is the treatment effect on the hazard rate, which is computed as the simple difference in hazards for the Jobs First and AFDC groups. There are three salient features to this graph. First, there is an enormous spike in the treatment effect on the hazard at exactly month 22, the first month when the statutory time limit binds. Second, there are progressively smaller spikes at months 28, 34, and 46. These are months when 6-month extensions would expire for women who receive them. There is also a spike at month 39, one month before a third extension would expire (it is unclear why this spike is a month off). These spikes clearly imply that the time limit policy was enforced for at least some women, which is a key result. A third prominent feature of Figure 2 occurs in the first month following random assignment. The month-1 control group hazard is very large, showing that a fifth of AFDC-assigned women either leave welfare almost immediately or have their applications declined. For this month, the treatment effect on the hazard is significantly negative (around -0.08). Thus, while Jobs First women are still very likely to leave welfare very soon after assignment, they are significantly less likely than AFDC women to do so. Our discussion below of the Jobs First disregard expansion will shed some light on this finding.

Figure 2 concerns only the first spell, and hazards can be difficult to interpret after early months, since the risk set shrinks over time. Consider then Figure 3, which plots monthly welfare participation rates for women regardless of spell. The smooth series (dashed line) is the welfare participation rate by month for the AFDC group, while the other series (solid line) is again the treatment effect, calculated as the simple difference in the participation rate for the Jobs First and AFDC groups. This figure has four key features. First, the treatment effect of Jobs First on welfare participation is actually positive throughout the pre-time limit period. Second, there is a large drop in the treatment effect at exactly the month when time limits can first bind. Third, the treatment effect on welfare participation is negative after this point. Fourth, the figure also suggests the time limit was not binding for everyone. At month 22, the control group welfare participation rate was about 50%. If time limits were universally binding, we would have expected the drop in the treatment effect between months 21 and 22 to be much larger than the figure shows. This is of

course just another way of saying that exemptions and extensions were provided, as suggested by Figure 2. In any case, the two figures provide compelling evidence that the time limit policy was binding for a substantial number of women. This is the important fact for our purposes.

2.3 The expanded earnings disregard

As discussed above, Jobs First's disregard policy is quite simple: every dollar of earnings below the federal poverty line (FPL) is disregarded for purposes of benefit determination. This policy is very generous by comparison to AFDC's. The statutory AFDC policy was to disregard the first $120 of monthly earnings during a woman's first 12 months on aid, and $90 thereafter. In the first four months, benefits were reduced by two dollars for every three dollars earned (or otherwise received), and starting with the fifth month, benefits were reduced dollar-for-dollar, so that the long-run statutory implicit tax rate on earnings above the disregard was 100%.[10] In practice, there are two good reasons to think that the effective AFDC tax rate was lower than 100%, even after the fourth month.

First, a number of work-related expenses (e.g., transportation and child care costs) are supposed to be disregarded for benefit determination purposes.[11] Second, it appears from the final report that eligibility redetermination for AFDC recipients is done annually, rather than monthly. There can be a long lag between the month when an AFDC participant earns income and the date when benefits are reduced. In addition, the high 40% subsidy rate in the phase-in region of the EITC further reduces the effective tax rate faced by AFDC recipients.

To illustrate the dramatic difference in the treatment of earnings under AFDC and Jobs First, Figure 4 provides plots of local nonparametric (LOWESS) regression results for the observed relationship between quarterly earnings (the horizontal axis) and quarterly transfer income (defined

[10]The Jobs First expanded disregard also affects Food Stamp eligibility and benefits. Under AFDC rules, eligibility for AFDC conferred categorical eligibility for Food Stamps. Increasing the earnings disregard will, in general, lead to an increase in eligibility for welfare and an increase in Food Stamp eligibility. However, losing eligibility for welfare benefits (e.g., through time limits) need not eliminate Food Stamp eligibility, since one could still satisfy the Food Stamps need standards. In addition, Jobs First Food Stamps rules mirrored cash assistance rules, with Food Stamps benefits determined after disregarding all earnings up to the poverty line.

[11]Expense deductions lower the observed tax on earnings. But it is less clear whether incentives faced by a woman who has no work expenses and faces a 100% tax rate are any different from those faced by a woman who pays C in work expenses and then has C dollars disregarded for benefit determination. The key question is whether women derive *per se* utility from riding the bus to work or from sending children to daycare. If so, then the second kind of woman is better off than the first. If not, the presence of disregards simply prevents the net return to work from being negative.

as cash welfare plus Food Stamps).[12] In these figures, every person-quarter is treated as a distinct observation. Panel (a) of the figure is for the first seven post-assignment quarters, before time limits can bind for anyone. The dotted line is for the AFDC group, while the solid line is for the Jobs First group. The results show that for quarterly earnings below about $2,500, the slope for Jobs First members is essentially 0. For AFDC members and for Jobs First members with higher earnings, the slope implies that benefits fall about one dollar for every three earned. The picture is generally similar (though shifted downward) for quarters 8–16, presented in Panel (c).

One complication in interpreting these figures is that while data on transfer income are available monthly, earnings data are available only quarterly. One way to get a clearer picture is to consider only cases that have nonzero transfer income in all 3 months of a given quarter. Panels (b) and (d) replicate (a) and (c) with this selection criterion. The picture in Panel (b) is remarkable: for the Jobs First group, the associated benefit reduction rate is almost exactly zero across the entire earnings distribution, while the AFDC group's slope is again approximately -1/3. It seems clear that the effective tax rate on earnings is substantially below 1 for the AFDC group; this finding, which has been made in other contexts (e.g., see McKinnish, Sanders & Smith (1999) and Fraker, Moffitt & Wolf (1985)) is an interesting result in itself.

For our purposes, the key finding from these figures is that as implemented, Jobs First significantly reduced the effective tax rate on earnings. This finding has important implications for analyzing the labor supply implications of Jobs First, which we do below in Section 2.5.

2.4 Other changes in Jobs First

Jobs First changed other features of welfare in Connecticut including job search assistance, work requirements, sanctions, more generous child support pass-through, more generous asset limits, child care and medical insurance expansions, and family caps. These changes are less important in the current context either because they were relatively minor policy changes, or because they were not enforced stringently. As will be seen below, these changes all have essentially uniform predictions for labor supply.

Formal employment assistance under Jobs First was relatively limited. For example, the final

[12]The lowess regressions were estimated using a bandwidth parameter that includes 10% of the sample in each local regression.

report explains that in the first two-and-half years of the program, contracted providers supplied only "roughly two weeks of classroom instruction in job-seeking and job-holding skills, followed by several more weeks of monitored job search." Moreover, the final report makes clear that monitoring of compliance with employment mandates was very weak, partly due to low payments by the state to contractors meant to track and promote employment. Partly as a consequence, financial sanctions for failure to comply with the mandates were rarely levied. The report states that between 8–13% of Jobs First participants were ever sanctioned, by comparison to 5% of the AFDC-eligible control group. Unfortunately, the weak monitoring of employment patterns means that no reliable data on noncompliance are available. Moreover, under some circumstances caseworkers are empowered to grant Jobs First participants exemptions from work mandates and/or time limits.[13] In fact, over the life of the Jobs First evaluation, 30% of the sample received an exemption in at least one month. Thus, it is impossible to estimate a reliable noncompliance rate. What is clear is that, at least during the evaluation period, the Jobs First work requirements could generally be ignored by program participants with limited risk of sanction.

Jobs First also increased the effective pass-through of child support payments from $50, though only to $100.[14] While Jobs First allowed families to have up to $3,000 in assets and more car equity than under AFDC, the final report suggests that women assigned to Jobs First had no more savings than did women in the control group. The final report also suggests that the practical differences between child care assistance provided by AFDC and Jobs First were limited. With respect to health insurance, Jobs First provided transitional medical assistance via Medicaid for one year longer than the AFDC program. However, Connecticut expanded other health insurance policies during the evaluation period (e.g., Medicaid, SCHIP), so the difference may not be practically important. Finally, the partial family cap reduced by about half the incremental benefit paid after the birth of a child conceived while the woman was receiving welfare. However, the incremental welfare payment was only about $100 to begin with, and the final report notes that no differences

[13]Those circumstances include physical or mental incapacitation, responsibility to care for a disabled relative, having a child aged younger than 1, and being deemed unemployable due to limited work history and human capital. Having an exemption means that a woman has no work requirement and that welfare participation does not count toward the time limit as long as the exemption is in effect.

[14]Under AFDC, only $50 of each monthly child support payment actually accrued to the mother, with the state keeping the balance. Under Jobs First, families received all child support collected on their behalf, but only the first $100 was disregarded in determining their welfare payment. According to the final report, Jobs First families reported having received slightly higher child support payments (by about $30 a month).

in childbearing were observed across program assignment.

2.5 Expected impacts on earnings, transfers, and income

2.5.1 The disregard

Figure 1 uses stylized budget constraints to illustrate how the expanded disregard affects women's labor-leisure tradeoff. Consider first women who would locate at the corner of the AFDC budget set, working 0 hours and receiving the maximum benefit. For these women, Jobs First raises the effective wage from 0 to w, leading to unambiguous increases in employment rates, hours, earnings, and income. However, transfer income will be unchanged for these women, since they are already receiving the maximum payment. Of course, not all women will enter the labor market—even with the more generous disregard, the after-tax wage still may not exceed the woman's reservation level.

Next consider the impact of the Jobs First disregard on women who would ultimately have hours of labor supply that exceed the AFDC breakeven hours level. Presumably such women are those whose offered wages are temporarily low, due to frictional unemployment, marital dissolution, or some other negative shock, but then return to a long run value that makes it optimal to increase labor supply.[15] Such women may be affected by Jobs First's disregard expansion, which adds a new, much higher kink to the budget constraint where earnings equal the federal poverty line and benefits equal the maximum benefit. The usual prediction from such a massive shifting out of the budget set is that some women will reduce hours below the breakeven level in order to gain eligibility for transfers. Thus for these high-wage women, we expect both hours and earnings to fall, while transfers should rise (from zero to the maximum benefit). The total impact on income depends on whether the earnings reduction outweighs the increase in transfer payments.

For a final group of women, long run wages may be so high that in the absence of a negative shock they would never participate in either AFDC or Jobs First. For such women, who will tend to be at the top of the earnings distribution under either welfare system, the treatment effect on all three of our outcome variables will be zero.

In sum, the predicted effects on the earnings distribution of the disregard expansion are hetero-

[15] While a static model with a fixed offered wage—like that represented by Figure 1—cannot capture such dynamically varying labor supply, it is helpful in guiding our understanding of labor supply choices within periods of time when the wage is fixed.

geneous: no change at the bottom, increases in the middle, decreases at high earnings, and perhaps no change at the very top. While these predictions are well-known from earlier discussions of the role of AFDC benefit reduction rates, the Jobs First disregard expansion is larger than any previous change. And as mentioned in the introduction, mean impacts could mask sizable, opposite-signed impacts across the distribution.

2.5.2 The time limit

Time limit-induced elimination of welfare eligibility reduces welfare participation and transfer income, and when it occurs it should also lead to increases in labor supply and earnings. Participants may also "bank" their eligibility by reducing welfare and increasing labor supply even before the time limit binds, as discussed in Grogger & Michalopoulos (2003). Because earnings are predicted to rise while transfers are predicted to fall, theoretical predictions concerning total income and time limits are ambiguous.

2.5.3 Other program changes

Mandatory work activities should also lead to increased earnings and reduced transfers, as discussed in Moffitt (2002) and Besley & Coate (1998). Eliminating the option of receiving welfare and working fewer than a set number of hours will cause recipients who remain on aid to increase labor supply and earnings. Such work requirements reduce the utility associated with welfare receipt, so we also expect reduced welfare participation and increases in work. Sanctions act to financially penalize welfare participants for not complying with work requirements. Since they impose new costs on (some) recipients, sanctions should reduce welfare participation and benefits, while increasing work. The net impact on income is uncertain for both work requirements and sanctions.

3 Mean treatment effects

In this section, we report mean treatment effects. Before proceeding, we must address a complication concerning the quality of random assignment. Table 3 reports our estimates of several pre-treatment statistics: mean quarterly levels of earnings, cash welfare, and Food Stamps, as well

as the fraction of pre-treatment quarters in which each of these variables was nonzero.[16] This table shows that before treatment, the Jobs First group had significantly lower earnings and greater cash welfare use than did the AFDC group. The final report notes this but does not provide any explanation for how this occurred. To deal with this problem, all of MDRC's reported treatment effects in the final report are the estimated coefficients on a Jobs First treatment dummy in OLS regressions that include pre-treatment data on earnings, cash welfare, and Food Stamps for the four quarters preceding random assignment.

While this is a common way of adjusting for pre-treatment differences, it is now well known that a more theoretically appropriate approach is to use inverse-propensity score weighting. This is the approach we take. Briefly, we estimate the probability that a person is in the treatment group using predicted values from a logit model in which the treatment dummy is related to the following variables: quarterly earnings in each of the 8 pre-assignment quarters, quarterly AFDC and Food Stamps payments in each of the 7 pre-assignment quarters, and dummies indicating whether each of these 22 variables is nonzero. Denote the estimated propensity score for person i $\hat{p}_i$ and the treatment dummy T_i. Then the estimated inverse-propensity score weight for person i is

$$\hat{\omega}_i \equiv \frac{T_i}{\hat{p}_i} + \frac{1 - T_i}{1 - \hat{p}_i}, \tag{1}$$

with these weights then used in the standard fashion for all estimators employed below.[17,18] In practice, adjusting for pre-treatment differences leads to few changes in the estimated mean impacts and does not change the overall picture.

[16]For earnings, data are available for the 8 quarters preceding random assignment; for cash welfare and Food Stamps, data for all observations are available only for the 7 quarters preceding random assignment.

[17]The literature on propensity scores and mean treatment effects, which is large and growing, began with Rosenbaum & Rubin (1983). Recent papers focusing on mean treatment effects include Heckman, Ichimura & Todd (1998) and Hirano, Imbens & Ridder (2003). Firpo (2003) has shown that the same approach corrects for bias in estimation of quantiles of the counterfactual treated and control distributions, with the simple differences of adjusted quantiles then serving as estimates of the quantile treatment effects. The weights given in (1) uncover treatment effects for the entire population represented by the experimental population. Alternative weights could be used to estimate the effects of treatment on the treated, but our objective is to estimate the effects of Jobs First under the assumption of generalizability.

[18]Two estimation issues arise when using inverse-propensity score weighting. The first is that theoretical results generally require nonparametric estimation of the propensity score, while our estimates are parametric. Second, the variance of estimated treatment effects (mean or quantile) will depend partly on the variance of the estimated propensity score (and its covariance with treatment and with the unexplained part of the outcome of interest). We address this issue with the bluntest instrument possible, by simply bootstrapping all of our estimates.

3.1 Mean treatment effects

The first column of Table 4 reports estimated mean levels among the Jobs First group for several variables, over the entire 16-quarter post-treatment period. The first three rows concern average quarterly values of total income, earnings, and total transfers (where total income is defined as the sum of earnings and total transfers). The second column provides means for the AFDC group over the same period, and the third column provides the resulting mean impacts. These results show that over the four years following random assignment, the impact of Jobs First on average total income was $135, compared to an estimated baseline quarterly income of $2,612 for the control group. Thus, the mean impact of Jobs First on income was about 5%. About two-thirds of this impact is due to an (insignificant) increase in earnings, with the other third due to a significant increase in transfers.

The bottom three rows provide means and impacts for binary variables indicating whether the person had positive levels of income, earnings, and transfers. For example, the value of 0.852 for "Any income" among the treatment group means that among women assigned to Jobs First, 85.2% of all person-quarters had a positive value for at least one of UI earnings, cash assistance, or Food Stamps.[19] The results show that the probability of having any earnings was 7 percentage points greater among the Jobs First group than the control group, an effect of 14 percent relative to the control group baseline. The probability of having any income or any transfers is essentially identical across treatment status.

The findings from the previous section suggest that in the first 21 months—before time limits bind for anyone—behavior induced by Jobs First is very different from behavior during the final 27 months. Thus, we separately estimate mean treatment effects for the pre- and post-time limit periods. The second set of columns includes only the first 7 quarters of data, while the third set includes only the last 9 quarters. The results suggest that average earnings increased 7 percent in the pre-time limit period and (an insignificant) 6 percent in the post-time limit period (the impact is greater in the later period, but average earnings for the control group are also significantly greater). The fraction of Jobs First person-quarters with any earnings also rises in both periods,

[19]This also means that about 15% of person-quarters had no value in any quarter for any of these variables. This could mean that 15% of persons never have any income, that everyone has positive income for all but 15% of quarters, or something in between. We return to this issue in the next subsection.

by 17 percent in the pre-time limit period and by 12.3% in the post-time limit.

The mean impacts for transfers are starkly different in the early and later periods. During the first 7 quarters, Jobs First members received $217—or 16.3%—more transfer income than did control group women. During the later period, Jobs First members received $98—or 12 percent—less in transfers. The same pattern is clear for the fraction of person-quarters with positive transfers.

The net result of these changes in earnings and transfers is that Jobs First increased mean total income significantly—in both economic and statistical terms—in the pre-time limit period. Nearly three-fourths of this increase is due to increased transfer income, rather than earnings. By contrast, in the post-time limit period, mean income was virtually identical across treatment status. This is the result of nearly equal increases in mean income and reductions in mean transfers. Nonetheless, the post-time limit employment rate (the fraction of quarters with any earnings) is considerably greater for the Jobs First group. This means that, conditional on working, average earnings are lower among women caused by Jobs First to work in the last 9 quarters.

4 Quantile treatment effects

We now turn to quantile treatment effects. The first subsection considers some key methodological issues necessary for interpreting the quantile treatment effects results. In subsection 4.2, we present the main results: quantile treatment effects for 98 centiles in graphical form, using all experimental participants.[20] We also investigate whether our results are likely to be driven by migration out of Connecticut or by withdrawal from both the labor market and welfare following marriage or increased child support. In subsection 4.3, we discuss whether the heterogeneity we find could be satisfactorily uncovered by looking separately at high school dropouts and non-dropouts, a widely used approach in the literature on welfare reform. In general, the answer is a clear no. We summarize the QTE findings in subsection 4.4.

[20] We computed the QTE at the 99^{th} quantile but do not include it in the figures below because its variance is frequently large enough to distort the scale of the figures. The extreme variance at high quantiles for unbounded distributions is well known; we do not have the same problem at the bottom of the distributions because they are all bounded below by zero.

4.1 Methodological issues

For the moment, ignore the need to adjust for propensity score differences. The quantile treatment effect for quantile q may be estimated very simply as the difference across treatment status in the two outcome quantiles. For instance, if we take the sample median for the treatment group and subtract from it the sample median for the control group, we have the quantile treatment effect at the .5 quantile. Other quantile treatment effects are estimated analogously.

One important methodological distinction must be made: that between quantile treatment effects and quantiles of the treatment effect distribution. To understand the distinction, it will be helpful to briefly introduce a model of causal effects. Let $T_i = 1$ if observation i receives the treatment, and 0 otherwise. Let $Y_i(t)$ be i's counterfactual value of the outcome Y if i has $T_i = t$. The fundamental evaluation problem is that for any i, at most one element of the pair $(Y_i(0), Y_i(1))$ can ever be observed: we cannot observe someone who is simultaneously treated and not treated. Evaluation methodology thus focuses on inferences concerning various features of the joint distribution of $(Y(0), Y(1))$. In particular, the marginal distributions $F_0(y)$ and $F_1(y)$ are always observed, where $F_t(y) \equiv \Pr[Y_i(t) \leq y]$ for a randomly drawn i. These are marginal distributions because the counterfactual control outcome for i can never be observed when $T_i = 1$, so that implicitly we have to "integrate out" $Y(0)$ when considering F_1, and vice versa. One can also think of these marginal distributions as the conditional distributions of the observable outcomes $Y_i \equiv T_i Y_i(1) + (1 - T_i)Y_i(0)$, with the conditioning done on T_i. There is an enormous literature concerning the model described in the text (which is variously called the Roy Model, the Quandt Model, and the Rubin Causal Model) and the assumptions under which it is useful. See, for example, excellent papers by Heckman et al. (1997) or Imbens & Angrist (1994) for further details.

Quantile treatment effects are features of the marginal distributions $F_0(y)$ and $F_1(y)$. As usual, for treatment assignment t, the q^{th} quantile of distribution F_t is defined as $y_q(t) \equiv \inf\{y : F_t(y) \geq q\}$. The quantile treatment effect for quantile q is then $\Delta_q = y_q(1) - y_q(0)$; our above example concerning the QTE for the median involves setting $q = 0.5$. To account for inverse propensity score weighting, we define the empirical *cdf* as $\hat{F}_t(y) \equiv \sum_{i:\, Y_i(t) \leq y} \hat{\omega}_i / \sum_i \hat{\omega}_i$ and then proceed as before.

For observation i, the treatment effect is $\delta_i \equiv Y_i(1) - Y_i(0)$, and the cumulative distribution of treatment effects may be written as $G(d) \equiv \Pr[\delta_i \leq d]$ for randomly chosen i; quantiles of this

distribution satisfy $d_q \equiv \inf\{d : G(d) \geq q\}$. By contrast to quantile treatment effects, quantiles of the treatment effects distribution cannot be written as features of the marginal distributions. Rather, they require more detailed knowledge of the joint distribution. Under some conditions, the distribution of treatment effects is recoverable. A leading example assumes that the treatment effect is equal for all observations, in which case G is degenerate (and fully identified by the mean impact). However, the above discussion of labor supply impacts suggests that is not valid here. A second example is rank preservation. Under rank preservation, any person whose outcome in the counterfactual control distribution is the q^{th} quantile will also have an outcome that is the q^{th} quantile in the counterfactual treated distribution. It then follows that Δ_q and δ_q are equal, and since Δ_q is always identified by the difference of marginals at q, the cumulative distribution of treatment effects G is also identified by sorting the set of estimated Δ_q.

It may be that rank preservation holds for a large portion of the distribution. However, there will likely be parts of the distribution—such as at the bottom of the distribution where some respond to the Jobs First incentives and others do not—where rank preservation fails. As a consequence, quantile treatment effects must be understood as what they are: differences in the treated and control *distributions*, not the treatment effects for identifiable women in either distribution.

In the absense of rank preservation, certain important features of the joint distribution of $(Y(0), Y(1))$ are still identified. A simple example is the mean treatment effect; if this is the object of interest, the marginal distributions are just as informative as the joint distribution. Even with heterogeneous treatment effects, some important features of the joint distribution can be identified, depending on the estimated quantile treatment effects. For example:[21]

1. Fix a quantile q^*. The minimum treatment effect δ_q for all $q \geq q^*$ is no larger than the smallest quantile treatment effect Δ_q for $q \geq q^*$. Thus if any QTE is negative, at least one treatment effect is also negative.
2. The logical inversion also holds. Fix a quantile q^*. Then $\sup_G\{\delta : \text{control group rank is } q \leq q^*\} \geq \sup\{\Delta_q : q \leq q^*\}$. Thus if any QTE is positive, at least one treatment effect is also positive.

[21] For an illuminating discussion concerning the distinction between the distribution of treatment effects and quantile treatment effects, see Heckman et al. (1997). Some items in the list below are discussed there, while others are not but can be shown easily.

3. The variance of the distribution of treatment effects is at least as great as the variance of the quantile treatment effects.

4. If subgroups are defined with respect to characteristics that are either permanent or fixed over the period of study, then the above results hold within subgroup, which may yield further information (e.g., the maximum QTE may be greater in a subgroup than in the pooled sample).

These examples show that quantile treatment effects do provide considerable information about treatment effect heterogeneity. We have seen that, other things equal, expanded disregards should leave earnings unchanged for women with sufficiently low offered wages, increase earnings as we move up the wage distribution, and reduce earnings (due to entry effects) at the top of the wage distribution. Moreover, theory suggests that, provided women remain in Connecticut and do not get married, no element of Jobs First should cause anyone to reduce earnings from a positive level to zero.[22] This argument suggests that quantile treatment effects may be fruitfully used in tandem with labor supply theory to understand the effects of Jobs First. Lastly, we note that classical social welfare function analysis, as we discuss in section 5, requires only the marginal, fixed-treatment distributions.

The nature of Jobs First, and the results in section 2, suggest that in practice Jobs First is best thought of as two overlapping programs: one that increases the generosity of the welfare system for women who combine welfare and work, and a second that greatly restricts this generosity for women who initially stay on welfare and demonstrate nontrivial earnings capacity. This Jekyll-and-Hyde program design raises methodological questions in using nonlinear estimators, including quantile treatment effects. It seems clear that looking separately at the pre- and post-time limit periods is important. But it is less clear how to do this. One approach is to estimate quantile treatment effects for outcomes averaged over the relevant periods; for example, average quarterly earnings in the seven pre-time limit quarters. A second approach is to use the person-quarter as the unit of

[22] Welfare and unemployment insurance programs can be thought of as subsidizing job search, and expansions in their generosity are typically thought to raise reservation wages for working. This sort of effect would delay the onset of positive earnings, causing some treated women to have zero earnings in some months, when their counterfactual earnings given AFDC assignment would have been positive. However, Jobs First does not actually subsidize off-the-job search, since the expanded disregard affects disposable income only for women who have positive earnings. If anything, the expanded disregard (together with the time limit) should speed up exit from unemployment, which was certainly Connecticut's intention.

analysis.

There are arguments for each of these approaches. In general, one would like the unit of analysis to be the longest period through which consumption smoothing is possible. Thus, if women in the sample are able to smooth consumption across quarters over a two-year period, then averaged outcomes before and after time limits may be an appropriate measure of Jobs First's effects. However, it is easy to imagine that consumption smoothing will be difficult for welfare recipients: both the eligibility requirement of low assets and the fact that human capital for welfare recipients is typically low suggest that liquidity constraints likely are binding. An alternative argument for using average within-period outcomes is that it does not complicate standard error estimation, whereas using the person-quarter as the unit of analysis induces dependence in an otherwise-*iid* context. However, using inverse propensity score weighting complicates the estimation of standard errors anyway, a problem that we address by bootstrapping. Rather than take a stand on this issue, we simply report results computed each way; our results are qualitatively robust to these two approaches.

4.2 QTE results for the full sample

4.2.1 Earnings

Figure 5 introduces the QTE estimator. The top panel plots centiles of the earnings distribution using person-quarter observations among Jobs First and AFDC women in the first seven quarters following implementation of Jobs First (before time limits bind). The vertical difference between these lines at a given decile is an estimate of the reform's treatment effect on earnings at that quantile. We plot these QTEs in the bottom panel of Figure 5. For comparison purposes, the mean treatment effect is plotted as a horizontal (dashed) line, and the 0-line is provided for reference. Dotted lines represent 90% confidence intervals calculated using the empirical standard deviation of 250 bootstrap replications of the quantile treatment effects.[23] The variation in the impact across the quantiles of the distributions is unmistakably significant, both statistically and substantively.

This figure shows that for quarterly earnings in the pre-time limit period, the quantile treatment

[23]We use non-overlapping block bootstrap, with each person as a single block. This removes any within-person dependence in the estimates using the person-quarter as the unit of observation. The confidence intervals are based on a two-tailed test.

effect is identically zero for nearly half of all person-quarters. This result occurs because quarterly earnings are identically 0 for 48% of the pre-time limit treatment group person-quarters and 55% of the control group person-quarters. For quantiles 49–82, treatment group earnings are greater than control group earnings. Between quantiles 83–87, earnings are again equal (though non-zero). Finally, for quantiles 88–98, control group earnings exceed treatment group earnings. These results are exactly what basic labor supply theory, discussed above, predicts. It is useful to note that the range of the point estimates for the QTE is quite large: about [-$250 , $600], compared to a mean treatment effect of $93 with an estimated standard error of $57. As we noted above, no policy change in Jobs First should cause women to lower earnings from a positive amount to zero. Thus for all women who are assigned to Jobs First and have zero earnings, the treatment effect (as opposed to the QTE) is known to be zero.

For the remainder of the paper, we will present results like the bottom panel of Figure 5– plotting QTE estimates for each centile of the earnings, transfer, and total income distributions. The main results are reported in Figures6–8. In each figure, the graphs on the left hand side use the person-quarter as the unit of analysis, while those on the right-hand side use within-person quarterly averages. The top graphs are for the first 7 quarters, the middle ones for the last 9, and the bottom ones for all 16 quarters.[24] For comparison, Appendix figures 1–3 provide the inverse *cdf*s analogous to the top panel of Figure 5.

For the post-time limit period, earnings effects are broadly similar to those just discussed, though with a somewhat wider range. Considering all observations over the entire 16-quarter period also yields similar results.

The three figures using averaged quarterly earnings are all considerably smoother than those using person-quarters as the unit of analysis, as we expect when there is any within-person, over-time variation in earnings. Using averaged outcomes also reduces the range of estimated quantile treatment effects, especially at the top. However, this range is still very large relative to the mean treatment effects and their standard errors.

Finding that earnings quantile treatment effects are negative at higher quantiles suggests a program-entry (or more likely non-exit) response among women who would have had high earnings

[24]Thus, for the first seven quarters, quantile treatment effects are calculated using $7 \times 4,803 = 33,621$ observations. For the last nine quarters, there are $(9 \times 4,803) - 30 = 43,207$ observations.

and stayed off welfare were they subject to AFDC rules. Thus, the finding is consistent with the prediction that Jobs First's generous disregard will cause large negative income effects on hours worked for high-wage women. Indeed, according to data from the three-year survey discussed in the final report, hours worked in the month of the survey were lower among high-earnings Jobs First women than among high-earnings AFDC women.

In our view, the reduction in earnings at the top of the distribution caused by Jobs First is most likely due to the disregard expansion and its eligibility-inducing negative income effects. However, time limits and some of the other features of Jobs First will tend to lower the wage at which women are willing to exit welfare. Such accelerated search provides an alternative explanation fo the negative QTE results at higher quantiles of the earnings distribution. In fact, data from the three-year follow-up survey do suggest that employed Jobs First women have lower wages throughout much of the top half of the wage distribution.

But if lower reservation wages for exiting welfare were the only cause of reduced earnings at the top, then welfare participation rates at higher earnings levels should be lower among Jobs First than among AFDC women. To test this hypothesis, we first sort average earnings into 10 bins corresponding to deciles of the control group earnings distribution. We then calculate person-specific welfare participation rates and compare then across treatment status. If the search explanation is correct, then welfare participation rates should be lower among high-earnings Jobs First women; if the eligibility effect is correct, the opposite is true. In the pre-time limit period, we find that welfare participation rates are significantly greater among Jobs First women with earnings at least equal to the control group median. Moreover, this difference increases as we move to higher earnings bins. For example, for the three highest control group deciles, welfare participation rates are 12–15 percentage *points* greater in the Jobs First group. In the post-time limit period, this difference at higher earnings largely disappears. We would expect this result, since women who are induced to stay on Jobs First because of the generous disregard will generally have high earnings and thus be unlikely to get extensions. We thus conclude that in the pre-time limit period, negative quantile treatment effects at higher quantiles are likely due to disregard-induced negative income effects. In the post-time limit period, reduced work-no-welfare reservation wages are a more likely explanation.

4.2.2 Transfers

Figure 7 presents results for transfer income, defined as the sum of cash payments and the face value of Food Stamps. The most notable feature of these results is the radical difference in the treatment effects of Jobs First across the pre- and post-time limit period. Consider first the person-quarter results in the program's first seven quarters. The QTE is identically 0 for the bottom 20 quantiles, reflecting the fact that for 20% of person-quarters, both the treatment and control group have zero transfer income. For all person-quarter quantiles (except for two) above the 20th, transfer income in the pre-time limit period is greater among Jobs First women than among AFDC women. This finding greatly extends the result for mean treatment effects presented in section 3. Moreover, the range of quantile treatment effects in this period is very large, with the largest QTE reaching $700. As a basis of comparison, this is nearly three times the upper limit of the 90% confidence interval around the mean effect of $217; it is also nearly a third of the maximum quarterly value of Connecticut's combined AFDC-Food Stamps payment for a family of three. By comparison to AFDC, Jobs First in the pre-time limit period clearly is associated with a substantial upward shift of the transfer income inverse *cdf*.

The person-quarter graph for the post-time limits period is much different. The graph shows that for the lowest 48 quantiles, the Jobs First and AFDC transfer distributions are equal, with both showing zero transfer income at all these quantiles. However, at essentially all quantiles between 49–96, the Jobs First group receives less transfer income. The size of the reductions in transfer income can be quite large: the largest quarterly reduction is $550, and the reduction is at least $300 for all quantiles from 64–76. By stark contrast to the pre-time limit period, these results suggest that Jobs First in the post-time limit period is associated with a substantial *downward* shift of the transfer income inverse *cdf*. And again, these findings tell a much richer story than does the mean treatment effect of -$98, together with its standard error of $25.

Putting together the pre- and post-time limit periods in the bottom graph shows that the full, 16-quarter effects of Jobs First on the transfer income distribution are relatively limited. Aside from the $250 QTE at the 99th quantile, the largest QTE is only $150. The basic picture from the person-quarter approach is that, over the full period, the large positive quantile treatment effects in the pre-time limit period and the large negative effects in the post-time limit period more or less cancel each other out.

Using averaged quarterly transfers yields a generally similar story. As with earnings, the cross-quarter, within-person averaging reduces the variability of quantile treatment effects. For the pre- and post-time limit periods, averaging also reduces the range of the quantile treatment effects, but without changing the basic qualitative results. For the overall time period, however, results using averaging and person-quarter units appear somewhat different. First, the range of effects is actually greater using the averaged transfers. Second, the treatment effects on the distribution of average transfer income are clearly positive at lower quantiles and negative at higher quantiles.

4.2.3 Total income

We can now discuss the effects of Jobs First on total income. Total income as we observe it is the sum of earnings and transfers. Since a given increase in total income could be driven by either large reductions in transfers together with increases in earnings, or no change in transfers together with smaller increases in earnings, there need not be any particular relationship between quantile treatment effects for total income and for its components. Thus, it is worthwhile to consider quantile treatment effects for total income, which we present in Figure 8.

Person-quarter results for the pre-time limit period again suggest a large degree of heterogeneity in quantile treatment effects. Quantile treatment effects range from 0 for the bottom 9 quantiles—where total income from administratively measurable sources is 0—to $800 at the top of the range. The mean treatment effect for this period is $296, with a standard error of $47, so again, the range of quantile treatment effects is much greater than the range provided by a confidence interval for the mean treatment effect. For the post-time limit period, the mean treatment effect of $10 is trivial, but the quantile treatment effect results clearly show that it would be wrong to think Jobs First has no impact for the entire distribution. Quantile treatment effects for total income are zero for the first 18 quantiles and are actually negative for about 25 quantiles. For some of these quantiles the reduction in quarterly total income is as large as $300—the magnitude of the mean treatment effect in the pre-time limit period.

Putting together the two periods, the person-quarter quantile treatment effects on total income are presented in the bottom-left graph of Figure 8. The graph shows that for the bottom 14 quantiles, both the Jobs First and AFDC distributions have 0 total income. The quantile treatment effects remain non-positive, and generally negative, until quantile 37, though the magnitudes of the

negative quantile treatment effects is not very large. Quantile treatment effects are positive, and in some cases relatively large, for quantiles 41–98. Compared to the mean treatment effect of $135 (standard error=54), this is again quite a large range. As with earnings and transfers, results using averaged quarterly total income as the dependent variable yield the same basic qualitative conclusions regarding Jobs First's impact.

Results for total income thus suggest that compared to AFDC, Jobs First exhibits reductions or no effects on income at lower quantiles, with quantile treatment effects at higher quantiles clearly being positive. This finding drives much of the analysis in the next section.

4.2.4 Robustness checks related to exits from administrative data

One concern in interpreting the above QTE results involves women who have zero total income in some quarters. For these women to survive, they must have some way to finance consumption other than UI-covered Connecticut earnings, cash assistance through either Jobs First or Connecticut's AFDC program, and Food Stamps. Such women could have some other source of earnings (UI-noncovered or under-the-table), they could have support (cash or in-kind) from family members or absent non-custodial parents, or they could have moved out of Connecticut. A substantial amount of discussion in the final report (mostly using the three-year followup survey) suggests that neither marriage nor migration rates were systematically affected by welfare policies and that child support payments were only slightly impacted. That is not enough for our purposes, however, because it is always possible (for example) that high-earnings women were systematically caused by Jobs First to stay in Connecticut, while low-earnings women systematically moved out. In this situation, the overall migration rate would be unaffected, but the zeroes in our total income data would be driven partly by missing data.

To deal with this issue, we take the following conservative approach. For each woman with zero total income in some quarter, we find the last chronological quarter in which she had nonzero total income. We then act as if she moved out of the state at the beginning of that quarter, excluding all subsequent quarters (all of which have zero earnings, transfers, and total income) for that woman from the analysis. When we do this, slightly more than a fifth of the sample of person-quarters is dropped. While this is a nontrivial share, there is virtually no variation across treatment status in the overall probability of "exiting" the administrative data. Furthermore, at each quarter in

the followup period, there are no statistically significant differences in the probability of exiting the sample between the treatment and control group.[25] Nonetheless, we recalculated the quantile treatment effects excluding our synthetic "movers". Not surprisingly, the results estimated on this sample of nonmovers are qualitatively virtually identical to the figures presented above.

To further explore the robustness of our results to potential movers, we also re-calculated our estimates for the specifications using averaged outcomes as the dependent variables. To do this, we drop all women who ever "move" (rather than just dropping quarterly post-"move" observations, as just described). Again, this alteration makes no noticeable difference in our results.

4.3 Quantile treatment effects by subgroup

So far, we have included all women for whom we have usable data. One might argue that this stacks the deck in favor of finding significant treatment effect heterogeneity. We think the heterogeneity demonstrated above would still be important in this case. But it is interesting to examine whether mean treatment effects together with judicious choice of observable subgroups would allow us to capture most or all of the heterogeneity demonstrated above. To examine this question, we follow a common approach in the welfare reform literature, considering separately high school dropouts and women with at least a high school diploma.[26] Non-dropouts are often used as a comparison group: given non-dropouts' lower welfare participation rates, reforms are often thought to affect them less than dropouts. We note that to be part of the experiment, all women in our sample had to apply for welfare, so this argument is less clearcut than is typically the case. Nonetheless, this is a logical way to consider the subgroups question.[27]

The top row of Figure 9 provides graphs of quantile treatment effects in quarters 1–7 (person-

[25] The highest absolute-value t-statistic is 1.42 in a regression of indicators for "moving out" each quarter on treatment status.

[26] Dropout status is collected at a baseline intake survey around the time of random assignment. This information is then provided by MDRC in the same public-use file as the administrative records. This variable is missing for a small number of observations, which we exclude from the subgroup analysis.

[27] Various parts of the final report (especially Appendix I) contain analyses of a wide array of subgroups. However, the focus is on groups labeled "most disadvantaged" and "least disadvantaged". The most disadvantaged are those who "had received cash assistance for at least 22 of the 24 months prior to random assignment, had not worked in the year prior to random assignment, and did not have a high school diploma or GED certificate" (see footnote 12 on page 22 of the final report). The least disadvantaged are those for whom none of these three conditions hold. While the information on prior employment and welfare use is valuable and would probably be used by researchers when available, most public-use datasets (e.g., the CPS) do not have such detailed data. For comparability, we use only pre-treatment dropout status to cut the data. Analysis using the MDRC categorizations shows that this choice does not affect the conclusion we reach here.

quarter as unit of analysis) among dropouts for each of earnings, transfers, and total income. The bottom row provides analogous graphs for nondropouts. Figure 10 presents a similar collection of graphs for the last nine quarters after treatment.

We offer several observations regarding these results. First, the differences in mean treatment effects across dropout status are basically trivial, never exceeding $100 per quarter. Second, the heterogeneity in quantile treatment effects within dropout status appears to be no less than the heterogeneity when we pool observations. The profiles of the graphed quantile treatment effects are sometimes shifted in one direction or another, but the basic fact of heterogeneity is pervasive, and the basic shapes of the quantile treatment effects are similar. We note that the apparent compression of the earnings and total income graphs among dropouts is just an artifact of scale distortion due to the high variance in QTE estimates for dropouts; if we plot only the point estimates, the shapes are very similar.[28] These graphs provide abundant evidence that if we used the most common mean impacts-based subgroup approach to dealing with potential treatment effect heterogeneity, we would miss virtually the entire story.[29] The conclusion that there is little heterogeneity in treatment effects has in fact been drawn in the mean impacts literature, as shown by this quote from Grogger et al. (2002, p. 231) summarizing both nonexperimental and experimental evidence: "the effects of reform do not generally appear to be concentrated among any particular group of recipients."

4.4 Summary of QTE results

The results presented in this section establish several clear conclusions. First, mean treatment effects miss a lot: estimated quantile treatment effects show a great deal of heterogeneity. Second, the results for earnings are clearly consistent with predictions from labor supply theory that effects at the bottom should be zero or positive, while effects at the top should be negative. It is important that the true treatment effect on earnings appears to be zero for a large fraction of person-quarters.

[28] We also computed estimates using averaged values of the outcomes. As above, the profile of these estimated quantile treatment effects was smoother, but the basic story is entirely unchanged.

[29] One interesting difference in the earnings quantile treatment effects across dropout status is that dropouts do not exhibit negative earnings impacts for higher quantiles. This is likely due to the fact that the upper quantiles of the dropout quarterly earnings distribution occur at very low levels. For example, among Jobs First women, the first 7 quarters, the 95th quantile is just $3,700, barely the federal poverty line for a family of three. By contrast, the 95th quantile among nondropouts is $5,300, well above the point where Jobs First eligibility would be lost for a family of three. This suggests that the disregard expansion's income effect is less likely to affect dropouts than non-dropouts.

Third, the effects of Jobs First are very different in the pre- and post-time limit period, especially with respect to the transfers distribution.

Fourth, it is not unreasonable to believe that Jobs First led to substantial increases in income for a large group of women. On the other hand, it had at best no impact, and perhaps a negative one, on another sizable group of women. This finding is at odds with that of Schoeni & Blank (2003 (Papers and Proceedings)), discussed in our introduction. Moreover, we find that most of the shift in the income distribution occurs at above-median quantiles.

Fifth, our results are unaffected by dropping observations from women who may have moved out of state or otherwise left the public assistance system while having no earnings (e.g., gotten married). Sixth, focusing on differences in mean treatment effects between dropouts and non-dropouts—perhaps the most common comparison-group approach—is virtually useless in uncovering the treatment effect heterogeneity we demonstrate.

5 Assessing Jobs First using alternative social welfare functions

Perhaps the most common approach taken to evaluating whether programs like Jobs First are worthwhile overall is mean cost-benefit analysis. For example, in the final report, MDRC computes average income gains for Jobs First women relative to AFDC women, and then compares these gains to the additional costs to the government of running Jobs First rather than AFDC. But since Jobs First's treatment effects vary considerably, mean cost-benefit analysis may miss important distributional effects.

To evaluate the program, we take a classical social welfare function approach. Ignoring experimental assignment for the moment, let person i's transfer income in quarter t be given by τ_{it}, and let earnings be w_{it}. Then, accounting for payroll taxes of 7.65% (and ignoring other taxes), i's disposable income in t is $y_{it} \equiv \tau_{it} + 0.9235 w_{it}$.[30] The standard approach is to define social welfare over y_{it}. However, moving women from welfare to work has been an explicit goal of virtually every welfare reform program undertaken in the last decade. To allow for this evident preference among policymakers for earnings rather than transfer income, we define social welfare

[30] We do not observe family size in the public-use data, which makes it more difficult to impute state and federal income taxes. MDRC includes an imputed EITC, which we do use below. Note that payroll taxes are thus treated as if their incidence is entirely on workers and valued at zero over the period of study.

over $y^*_{it}(\alpha) \equiv 0.01 + y_{it} + \alpha w_{it}$, so that α measures the additional weight placed on a dollar of earnings compared to a dollar of transfer income. We add 0.01 because in cases with a high degree of inequality aversion, the social welfare function is unboundedly negative if anyone has zero consumption. Letting $z = (\tau, w)$ be the observed vector of disposable income for all N persons over all T periods, our class of social welfare functions is

$$S(z; \rho, \alpha) \equiv \frac{1}{NT} \sum_{i,t} \begin{cases} \frac{y^*_{it}(\alpha)^{1-\rho}}{1-\rho} & \text{if } \rho \neq 1 \\ \ln y^*_{it}(\alpha) & \text{otherwise} \end{cases} \tag{2}$$

We then compute $S_J = S(z; \rho, \alpha | J)$ and $S_A = S(z; \rho, \alpha | A)$, where J and A respectively denote Jobs First and AFDC assignment. To evaluate Jobs First in terms of the social welfare function, we simply observe the sign of $\Delta S \equiv S_J - S_A$. The parameter ρ is the standard coefficient of relative risk aversion. This functional form nicely nests the mean impacts case, which occurs when $\rho = 0$. For now, we ignore administrative and operating costs of Jobs First, but we incorporate them below.

An advantage of this social welfare function approach is that it requires only the observed data on the marginal distributions. In other words, when a social welfare function is the evaluation criterion, "names don't matter": two policies that increase the social welfare function by 100 are viewed equally, even if one reduces utility for all but one person (whose utility rises significantly) and the other raises everyone's utility equally. This approach is not new (*e.g.*, see Atkinson (1970)) and has recently been advocated by Abadie, Angrist & Imbens (2002). An alternative view is that distributions of treatment effects (or some functionals of them) are themselves necessary inputs to the evaluation of programs. For example, Heckman et al. (1997, p. 488) write that "Appeal to a mythical social welfare function begs fundamental questions of political economy. The distribution of the benefits (and costs) from a programme determines the support for a programme if voters are self-interested or if they are altruistic." According to this view, "names matter".[31] In general, we agree that the distribution of treatment effects will be important for determining political viability

[31] A third approach is that taken by Dehejia (Forthcoming), who treats program evaluation as a decision-theoretic problem. He uses data from the Alameda County portion of California's job search- and training-based GAIN experiment, conducted in the late 1980s, to assess the role of heterogeneity in guiding optimal program assignment in the post-evaluation period. This approach solves the fundamental evaluation problem described above by assuming it away, i.e., by assuming that the functional form of the joint distribution $(Y(0), Y(1))$ can be estimated parametrically. Angrist & Dehejia (2001) also allow for individual and social welfare considerations in the presence of risk and inequality aversion.

of a reform. But in the present context, we do not think this shortcoming is serious, since members of the welfare population appear to have little political clout, either individually or as a group.

One must still decide whether the results of our evaluation can be generalized to some steady state, or whether they are confined to the particular population and time period represented here. The experimental sample was drawn in such a way as to represent the New Haven and Manchester welfare populations during the period between April 1996 and February 1997. The fact that both existing recipients and new applicants are included is helpful, since it means that our sample represents the true population over the intake period.

On the other hand, the underlying process generating new potential welfare recipients (due to job losses, divorces, and out-of-wedlock births) and welfare exits (due to new job matches, marriages, and aging-out of children) may not be dynamically stable. In this case, our results will not generally be informative about states of the world with different in- and out-flow rates. In addition, generalizability requires that there be no state dependence in welfare use. Our results show that Jobs First clearly changes the dynamic profile of welfare use, and true state dependence would interact with these changes. Thus the observed short-run program dynamics would not accurately represent steady-state behavior.[32] Lastly, we would need to assume that a four-year horizon is sufficient to evaluate a program. If (as is the case) some women spend many years on welfare, and if those women are an important part of the welfare population, then a four-year demonstration simply does not provide support for conclusions regarding the long-run effects of the program.

Unfortunately, there is not much we can do about any of these concerns, since the demonstration program is over. It is worth noting, however, that the same problem besets traditional mean-impact cost-benefit analysis: if dynamics are changed importantly by the reform, then no short-run evaluation will be fully informative about steady-state program effects. This weakness is related to experiments, not to our approach to using their data.[33] With all these caveats in mind, we believe the social welfare function exercise is worth doing. As above, we must also decide whether to use person-quarters as the unit of analysis or to use averaged values, a choice that boils down

[32]For a discussion of the econometric problems endemic to studies of welfare use and state dependence, see Chay, Hoynes & Hyslop (1999). These authors find significant evidence of true state dependence using administrative data from California.

[33]For an excellent discussion of the pros and cons of social experiments, see Heckman & Smith (1995).

to one's beliefs concerning the feasibility of consumption smoothing. Under either approach, if we make the assumptions sufficient to regard our results as representing a steady-state, we are entitled to argue that a given person in our sample represents all observably and unobservably similar people in a steady state, so our population represents the relevant steady-state population. If we think no consumption smoothing is possible, we can simply treat each person-quarter as a separate observation. Different quarters for a given person would then appropriately represent steady-state people of different cohorts. On the other hand, if we think that full consumption smoothing is possible, it would be more appropriate to use average income over the entire post-assignment period, since long run averages better represent the consumption resources available. As above, we handle this issue by reporting results calculated each way.[34]

We begin in Table 5 by focusing on person-quarters as the unit of analysis. The first row reports S_A and $\Delta S = S_J - S_A$ under the assumption that $\alpha = 0$, so that a dollar of transfer income has the same social value as a dollar of earnings. As noted, $\rho = 0$ yields the simple mean-impact cost-benefit analysis.[35] The first column shows that average quarterly income for the AFDC group is just under \$2,493, with the treatment effect being \$127. For this social welfare function, Jobs First passes the cost-benefit test. As we move to the right in the first row of this table, ρ increases, and the social welfare function places greater weight on lower-income women. Thus in percentage terms, the program gain with $\rho = 1/2$ (2.0, compared to a control group baseline of 87.8) is less than half that for the means case. When we get to log utilities, the program effect is actually slightly (though not significantly) negative. Lastly, when ρ reaches 2, Jobs First is associated with a statistically significant decline in social welfare of between 3–4 percent.

The second row of Table 5 increases α from 0 to 0.9235. Thus a dollar of after-payroll tax earnings is double-weighted in this specification. Since average earnings increase slightly with Jobs First, the $\rho = 0$ treatment effect increases. Moreover, we know that except for very high quantiles, where earnings fell, Jobs First either didn't affect the earnings distribution or raised it; since higher quantiles get relatively less weight when $\rho > 0$, the treatment effect rises relative to the $\alpha = 0$ case

[34]There is a large literature on consumption smoothing in the general population. With respect to people for whom transfer payments are an important source of income, Gruber (2000) finds little crowd-out of AFDC payments to women who transition into divorce, while Stephens (n.d.) finds that daily consumption is significantly sensitive to receipt of Social Security payments. Both of these findings suggest that full consumption smoothing is too strong an assumption.

[35]Note that our use of inverse propensity score weights means our results are not directly comparable to those in the final report. Also, unlike MDRC, we ignore discounting and inflation-adjustment.

for $\rho \in \{1/2, 1\}$ and is unaffected for $\rho = 2$.

Up to this point, we have ignored the issue of Jobs First's operational costs. Over the full five-year study period, Jobs First actually cost \$2,252 per person more than AFDC to administer. This added cost was due to additional expenditures on case management, education and training, other employment-related and post-time limit support services, child care, and transportation services. Moreover, as Table 4 shows, quarterly transfer payments were greater under Jobs First than AFDC, by \$42. How to account for these differences in administrative and operating costs is somewhat tricky. It might be tempting to simply subtract some multiple of the per-quarter added costs from S_J, but it is unclear how to do this when we are changing the curvature of the social welfare function across parameterizations. Instead, we observe that the government could have retained AFDC and simply made lump-sum transfers to all women in the study equal to $2252/20 + 42 = 154.60$ per quarter. We thus add this amount to each AFDC-group woman's transfer income in each quarter, noting that this provides a lower-bound on the impact of accounting for differential costs (since there might be other better ways to spend the money).

The third and fourth rows of Table 5, in Panel B, report results that account for differences in costs. Jobs First is now associated with significant social welfare reductions in all but the mean-impacts case in which earnings are double-weighted. In the inequality-averse cases with $\rho \in \{1, 2\}$, the reduction in social welfare is quite large in relative terms. For the $\rho = 2$ case, it is clear that this is so because the \$154.60 transfer for AFDC group women with zero income—but not for similarly situated Jobs First women—dominates the rest of the distribution. But even for the $\rho = 1/2$ case, accounting for costs makes a clear difference.

Table 6 replicates this analysis using values of earnings and transfers that have been averaged over the entire 16-quarter period. These results are notable in two ways. First, with the exception of the mean-impacts cases, the baseline levels of the social welfare function are larger than the person-quarter baselines, though not always enormously so.[36] For instance, when $\rho = 1/2$, $\alpha = 0$, and we ignore costs, baseline social welfare is 87.8 using person-quarters but 93.9 using averaged income values. Thus, the social value of full consumption smoothing in this case is 6.9% of social

[36]The mean-impacts cases differ slightly because a small number of women have no income data for the 16th quarter after assignment (because these women entered the experiment at the very end of the intake period and were followed for only 15 quarters). These women wind up being dropped from the averaged specifications, whereas only their Q16 data are dropped from the person-quarter specifications.

welfare when no smoothing is possible. As we increase ρ and social welfare becomes more sensitive to consumption fluctuations, the distinction between using person-quarters and averaged values becomes more important. For the $\alpha = 0$, $\rho = 2$ case, control group baseline social welfare is 94% greater with full smoothing.

Availability of full smoothing also has a large effect on the evaluation of Jobs First. When full smoothing is done and costs are ignored, Jobs First has a small positive effect on social welfare even with $\rho = 1$, and essentially no impact when $\rho = 2$. When Jobs First costs are redistributed as a lump sum to AFDC group members, we see negative effects of Jobs First only in all but the cases that double-weight earnings and have $\rho \in \{0, 1/2\}$. We note that these effects are considerably smaller than in the person-quarter analysis, though they are still statistically significant.

These results suggest that Jobs First-driven changes in income variability were substantial, at least for women with lower average incomes. To investigate this question, we created the coefficient of variation in total income over the full 16-quarter period for each woman in the sample. We then computed (inverse propensity score-weighted) differences in this variable across treatment status. For women in the Jobs First group, the mean coefficient of variation was 0.024 greater than the AFDC group's mean of 0.673, a difference that was not statistically significant. However, this apparent similarity masks considerable heterogeneity across the sample. For AFDC women with total, four-year income below the AFDC group's median of $40,000, the baseline coefficient of variation was 0.937; for Jobs First women with four-year income below $40,000, the coefficient of variation was 0.119 greater than this (with a standard error of 0.038). Among women with income above the AFDC group's median, the coefficient of variation was only 0.385, with virtually no difference by treatment group. These facts show that one thing Jobs First did do was greatly increase dynamic income variability in the lower half of the income distribution. This effect will clearly exacerbate the role of any liquidity constraints.

A natural remaining question is how the availability of the EITC affects our results. MDRC's public-use file includes a variable measuring imputed EITC takeup and amounts. Takeup is a serious issue, since receiving the EITC requires filing a tax return. Indeed, MDRC's three-year survey found that among women with annual earnings below $5,000, fewer than 60% reported filing a tax return for the previous year. These women are all in the phase-in range, so all would be entitled to a refundable credit. By contrast, 93.5% of women with earnings above $15,000

filed, and most of these women would be in the phase-out range. For a mean-impact analysis, estimating takeup and the amount of the credit may not be too worrisome. However, things are more complicated with a nonlinear social welfare function, given the EITC's own highly non-linear structure. Nonetheless, we generated a version of Table 5 using this imputed EITC variable, and the results are qualitatively very similar to those reported here.

6 Conclusion

This paper yields several important findings. First, mean impacts miss a great deal of treatment effect heterogeneity. Our estimated quantile treatment effects demonstrate that systematic heterogeneity is the rule, not the exception, in evaluating the effects of Jobs First. Moreover, comparing mean impacts for dropouts and nondropouts—a reasonable approach—fails entirely to uncover this heterogeneity. In fact, the within-group treatment effect heterogeneity is as great as the heterogeneity using the pooled sample.

Second, the results are consistent with—and in some cases likely confirm—basic predictions of labor supply theory concerning the expanded Jobs First disregard. Third, Jobs First's impact on transfer income appears to depend critically on whether time limits have yet taken effect. Before time limits, Jobs First leads to considerable increases in transfer income; afterward, it leads to considerable decreases. Fourth, for the bottom 40–50% of the income distribution (not necessarily of particular people located there), Jobs First either has no impact on total income or reduces it. At higher quantiles, this result is reversed. Using several parameterizations of a social welfare function, we find that Jobs First may be beneficial or not, depending on inequality aversion and accounting for Jobs First's greater administrative and operating costs.

Taken as a whole, these findings paint a mixed picture of Jobs First. On the one hand, it does not appear to have caused mass immiseration as some critics of recent welfare reforms would have predicted.[37] Moreover, there is clear evidence of sizable—compared to typical mean impacts—earnings gains for at least some women. On the other hand, the Jobs First experience shows that the law of unintended consequences has not been repealed. When sizable income gains occurred, they appear to have come from in large part through increased transfer income—hardly a fulfillment

[37] At least, if it did cause massive income reductions for some women, others must have had similarly sized gains.

of the stated PRWORA objective to "end the dependence of needy parents on government benefits by promoting job preparation, work, and marriage" (U.S. House of Representatives Committee on Ways and Means (2000)). Our results suggest that if a program like Jobs First is to end dependence, it is unlikely to do so only through work-related gains: even the comparatively large treatment effects on the earnings distribution we estimate here are far from large enough to make most families self-sufficient.

One important implication of our results is that people with different social welfare functions in mind can differ on whether Jobs First was a success. Those who place great weight on earnings and are relatively unconcerned about inequality may reasonably conclude that the program worked well, especially if they believe that substantial consumption smoothing is feasible. More inequality-averse observers would conclude the opposite; the social welfare loss is especially great when consumption smoothing is not feasible and inequality aversion is high.

References

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. & Dehejia, R. (2001), When is ATE enough? Risk aversion and inequality aversion in evaluating training programs. Typescript.

Atkinson, A. B. (1970), 'On the measurement of inequality', *Journal of Economic Theory* **2**, 244–263.

Bennett, N., Lu, H.-H. & Song, Y. (2002), Welfare reform and changes in the economic well-being of children, Working Paper 9399, NBER.

Besley, T. & Coate, S. (1998), 'Workfare versus welfare: Incentive arguments for work requirements in poverty-alleviation programs', *American Economic Review* **82**(1), 249–61.

Bitler, M. P., Gelbach, J. B. & Hoynes, H. W. (2003, (Papers and Proceedings)), 'Some evidence on race, welfare reform and household income', *American Economic Review* **93**(2), 293–8.

Blank, R. M. (2002), Evaluating welfare reform in the United States, Working Paper 8983, NBER.

Bloom, D., Kemple, J. J., Morris, P., Scrivener, S., Verma, N. & Hendra, R. (2000), *The Family Transition Program: Final Report on Florida's Initial Time-Limited Welfare Program*, Manpower Demonstration Research Corporation, New York, NY.

Bloom, D. & Michalopoulos, C. (2001), How did welfare and work policies affect employment and income: A synthesis of research, Working paper, Manpower Demonstration Research Corporation.

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.

Cabrera, N. & Evans, V. J. (2000), 'Welfare reform and its consequences: What questions are left unanswered?', *Joint Center for Poverty Research Poverty Research News* **4**(6).

Chay, K., Hoynes, H. & Hyslop, D. (1999), 'A non-experimental analysis of 'true' state dependence in monthly welfare participation sequences', *American Statistical Association, 1999 Proceedings of the Business and Economic Statistics Section* pp. 9–17.

Dehejia, R. H. (Forthcoming), 'Program evaluation as a decision problem', *Journal of Econometrics* .

Firpo, S. (2003), Efficient semiparametric estimation of quantile treatment effects. Typescript, UC Berkeley Department of Economics.

Fraker, T., Moffitt, R. & Wolf, D. (1985), 'Effective tax rates and guarantees in the AFDC program, 1967-1982', *Journal of Human Resources* **20**(2), 251–63.

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

Grogger, J. (Forthcoming), 'The effect of time limits, the EITC, and other policy changes on welfare use, work, and income among female-headed families', *Review of Economics and Statistics* .

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

Grogger, J. & Michalopoulos, C. (2003), 'Welfare dynamics under time limits', *Journal of Political Economy* **111**(3), 530–54.

Gruber, J. (2000), 'Cash welfare as a consumption smoothing mechanism for divorced mothers', *Journal of Public Economics* **75**, 157–82.

Heckman, J., Ichimura, H. & Todd, P. (1998), 'Matching as an econometric evaluations estimator', *Econometrica* **64**, 605–54.

Heckman, J. J. & Smith, J. A. (1995), 'Assessing the case for social experiments', *Journal of Economic Perspectives* **9**(2), 85–110.

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.

Hirano, K., Imbens, G. W. & Ridder, G. (2003), 'Efficient estimation of average treatment effects using the estimated propensity score', *Econometrica* **71**(4), 1161–1189.

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

McKinnish, T., Sanders, S. & Smith, J. (1999), 'Estimates of effective guarantees and tax rates in the AFDC program for the post-OBRA period', *Journal of Human Resources* **34**(2), 312–45.

Michalopoulos, C. & Schwartz, C. (2000), What works best for whom: Impacts of 20 welfare-to-work programs by subgroup, Working paper, Manpower Demonstration Research Corporation, U.S. Department of Health and Human Services, and U.S. Department of Education.

Moffitt, R. (1999), The effect of pre-PRWORA waivers on welfare caseloads and female earnings, income and labor force behavior, *in* S. Danziger, ed., 'Economic Conditions and Welfare Reform', W. E. Upjohn Institute for Employment Research, Kalamazoo, MI, pp. 91–118.

Moffitt, R. (2002), Welfare programs and labor supply, Working Paper 9168, NBER.

Rosenbaum, P. & Rubin, D. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**, 41–55.

Schoeni, R. & Blank, R. (2003 (Papers and Proceedings)), 'Changes in the distribution of child well-being over the 1990s', *American Economic Review* **93**(2), 304–8.

Schoeni, R. F. & Blank, R. M. (2000), What has welfare reform accomplished? Impacts on welfare participation, employment, income, poverty, and family structure, Working Paper 7627, NBER.

Stephens, M. (n.d.), '3rd of tha month: Do social security recipients smooth consumption between checks?', *American Economic Review* **93**(1), 406–22.

U.S. House of Representatives Committee on Ways and Means (2000), *Background Materials and Data on Programs Within the Jurisdiction of the Committe on Ways and Means*, U.S. Government Printing Office, Washington.

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

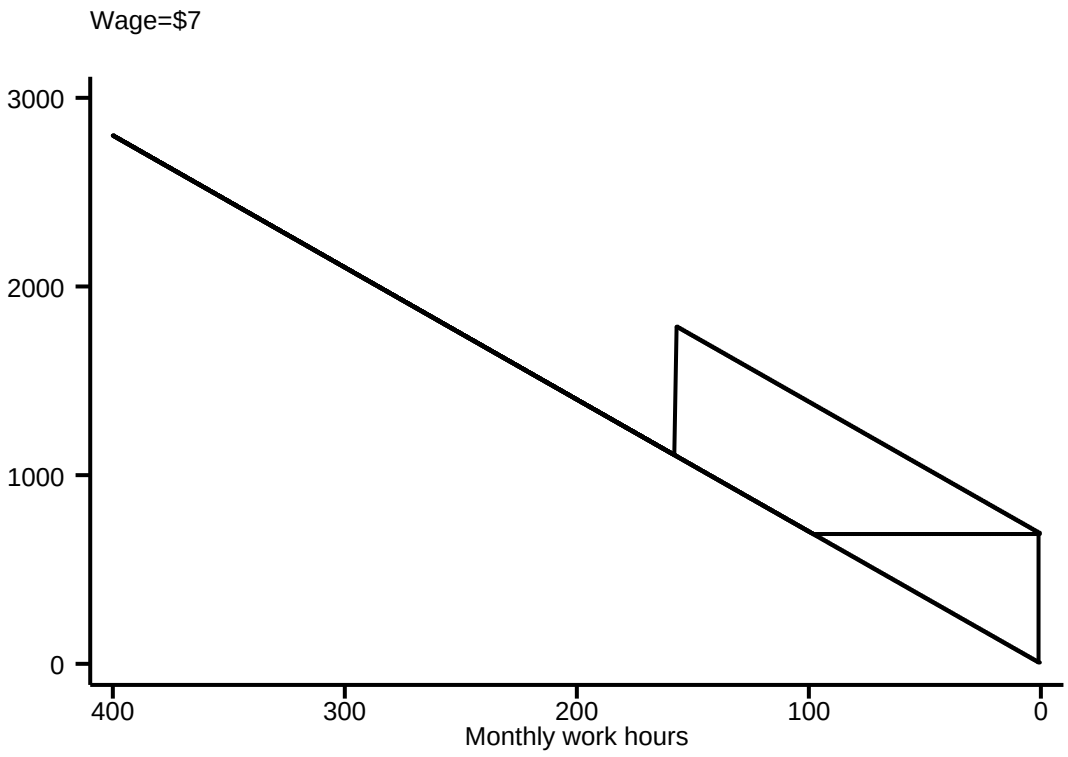


Figure 2: Jobs First monthly first-spell hazard rate: AFDC baseline and Jobs First–AFDC treatment effects

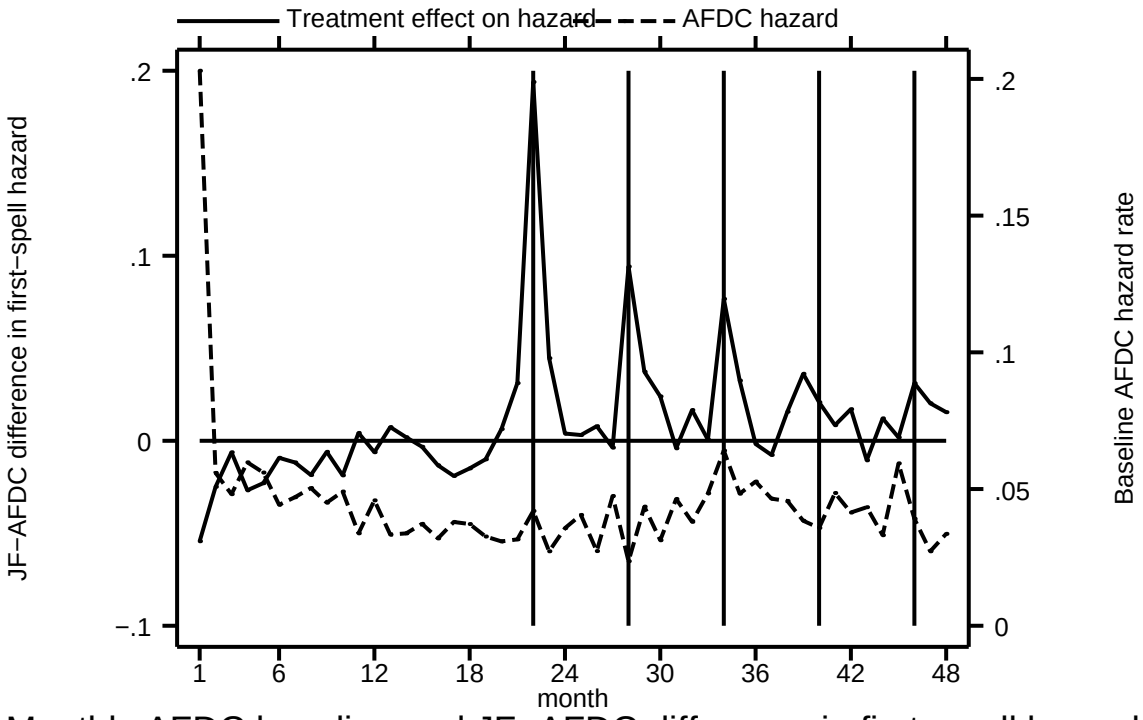


Figure 3: Monthly cash assistance receipt rates: AFDC baseline and Jobs First–AFDC treatment effects

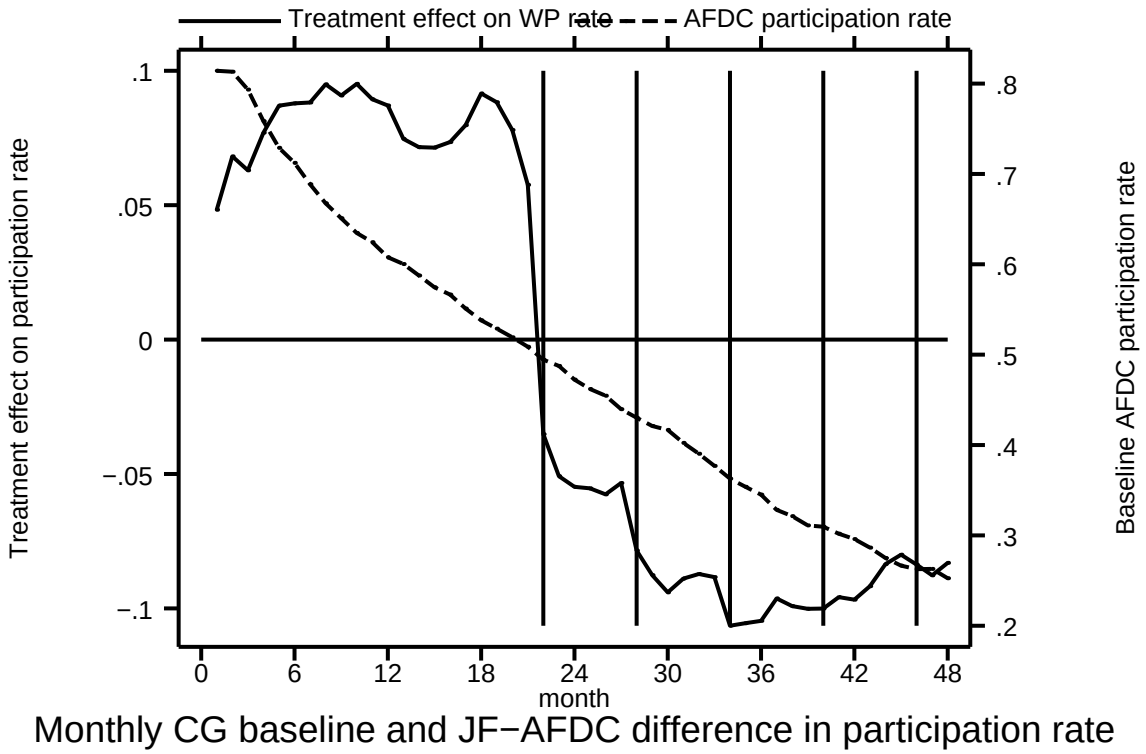
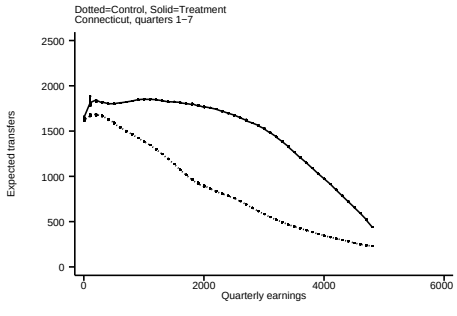
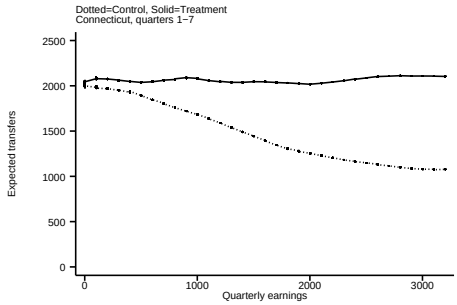


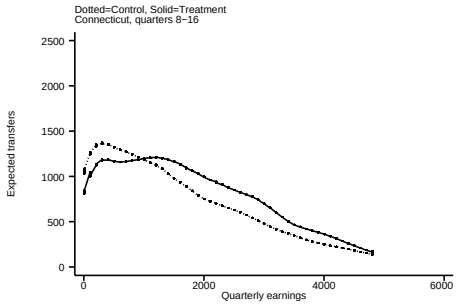
Figure 4: LOWESS regressions of quarterly transfers on quarterly earnings (unit of observation is the person-quarter)



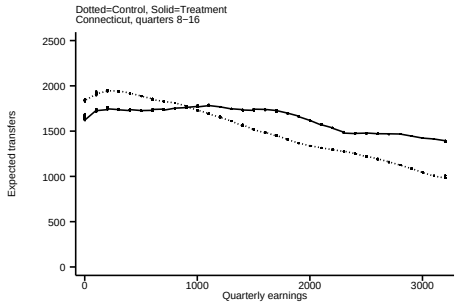
(a) Full sample, quarters 1–7



(b) Only person-quarters with welfare income all 3 months, quarters 1–7

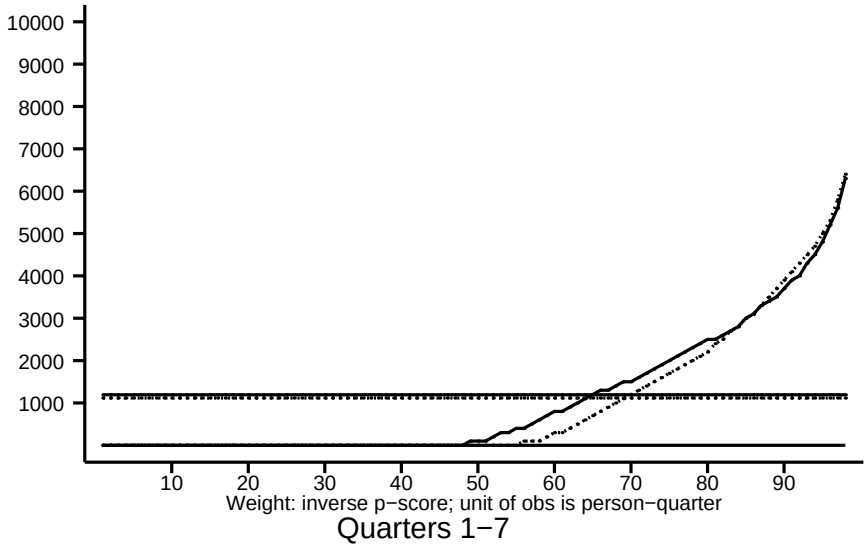


(c) Full sample, quarters 8–16



(d) Only person-quarters with welfare income all 3 months, quarters 8–16

Figure 5: Jobs First inverse cdfs and QTE estimates for earnings distribution, quarters 1–7



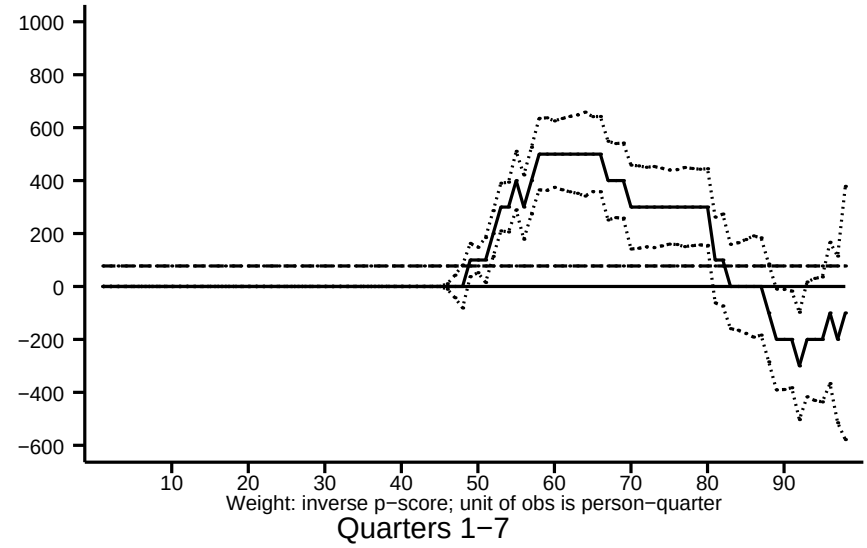
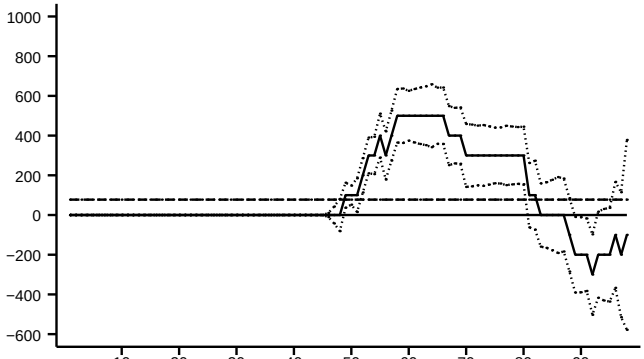
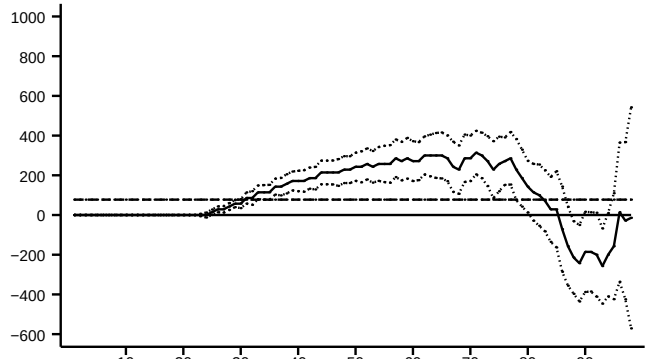


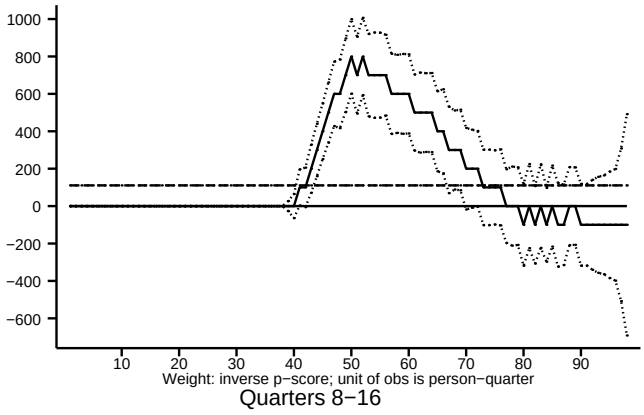
Figure 6: Quantile treatment effects on the distribution of earnings
(weighting done by inverse-propensity score, horizontal dashed line is mean treatment effect)

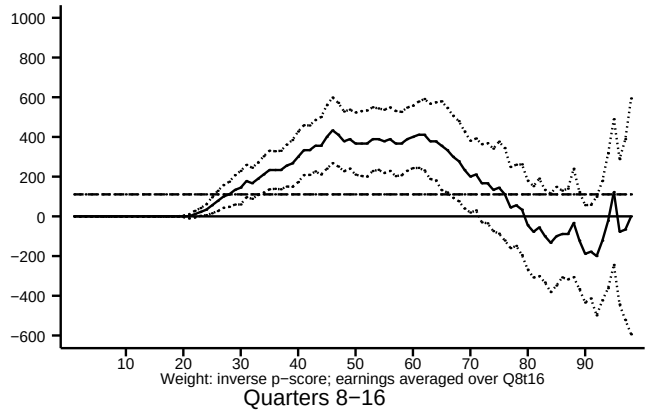
Person-quarter is unit of obs

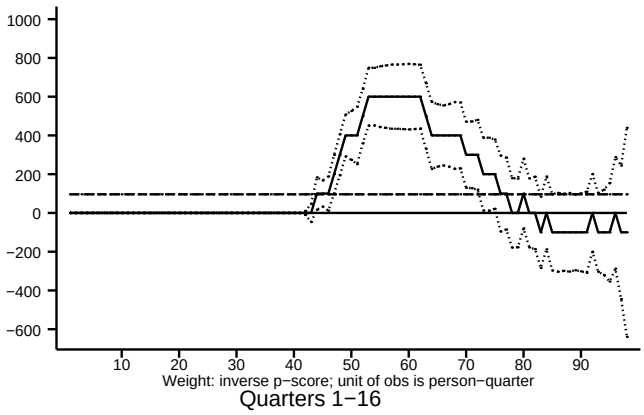
Averaged outcomes











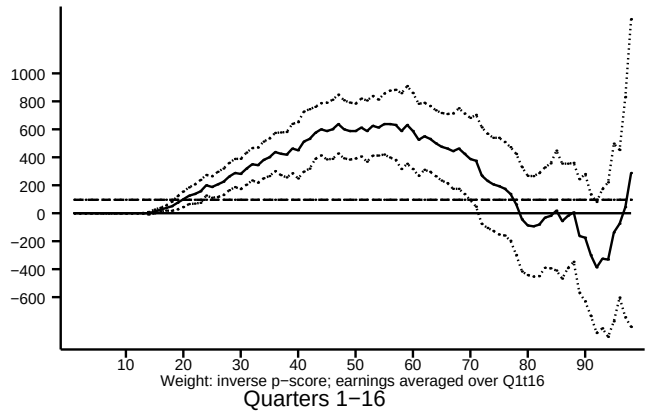
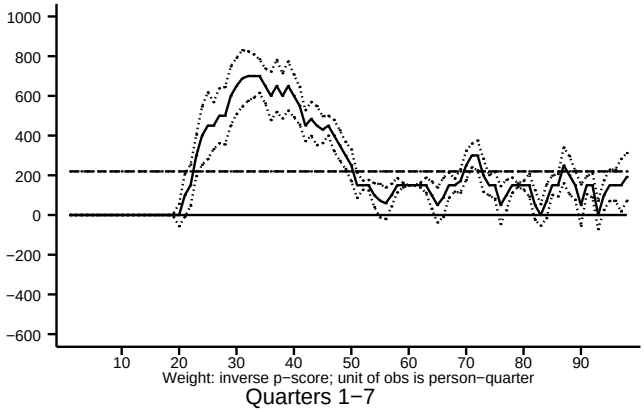
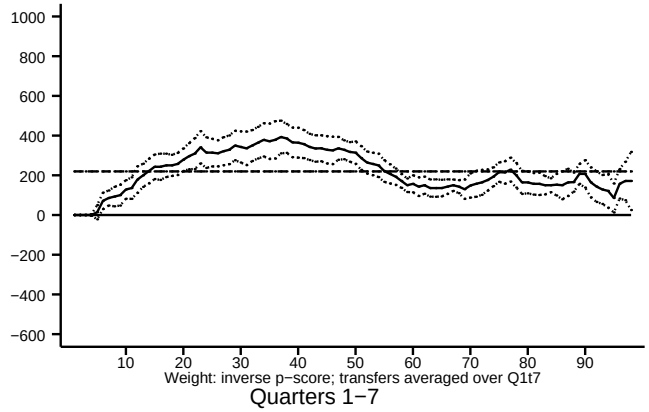


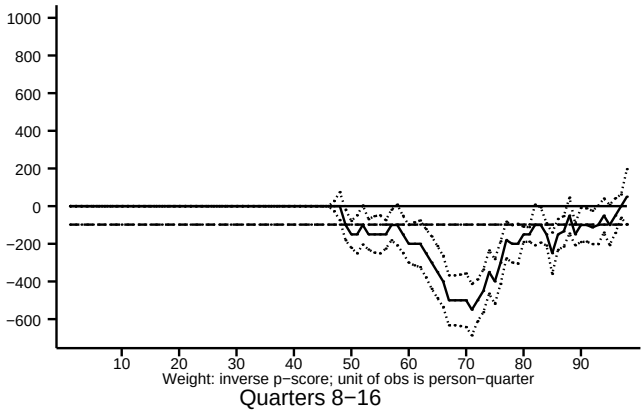
Figure 7: Quantile treatment effects on the distribution of transfers
(weighting done by inverse-propensity score, horizontal dashed line is mean treatment effect)

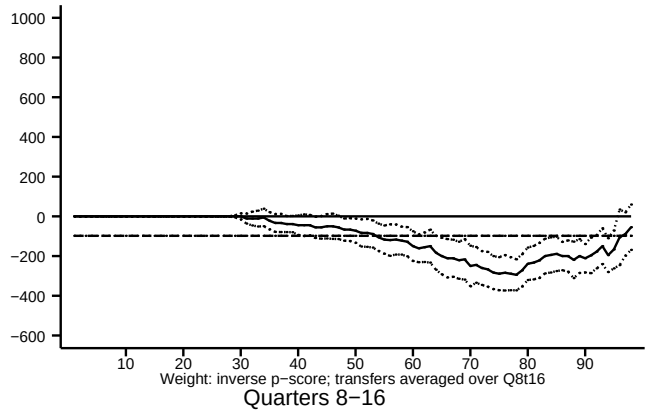
Person-quarter is unit of obs

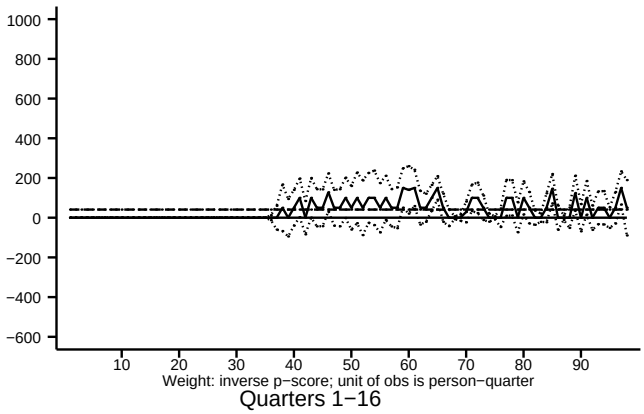
Averaged outcomes











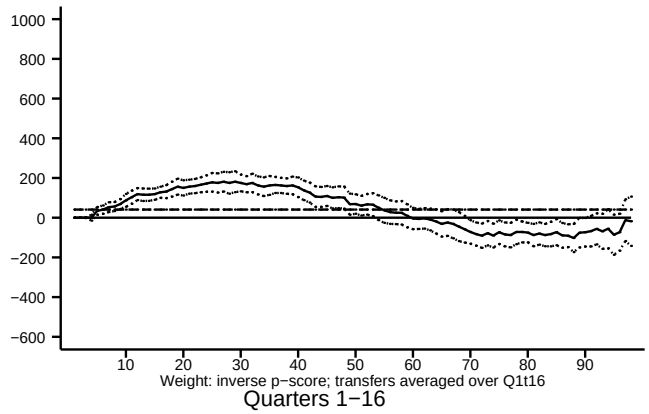
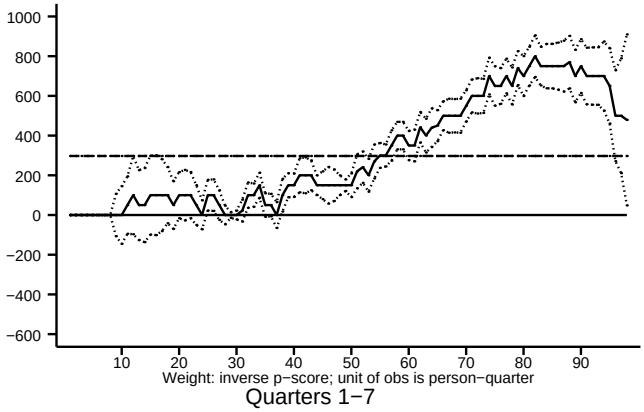
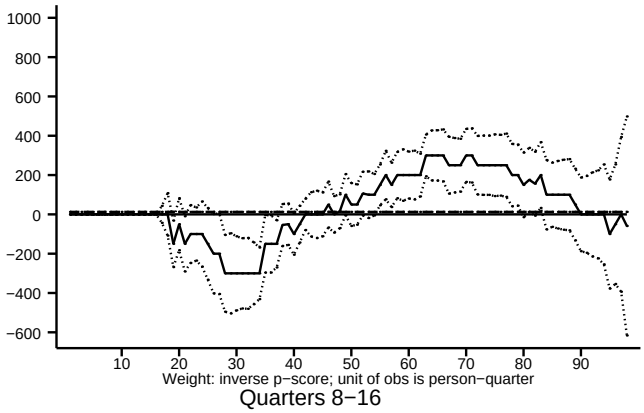


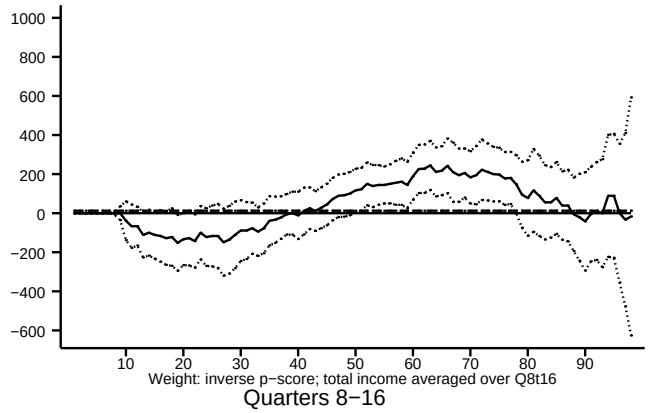
Figure 8: Quantile treatment effects on the distribution of income
(weighting done by inverse-propensity score, horizontal dashed line is mean treatment effect)

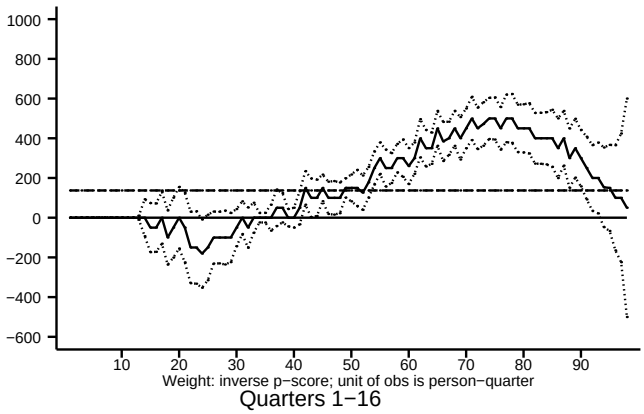
Person-quarter is unit of obs

Averaged outcomes









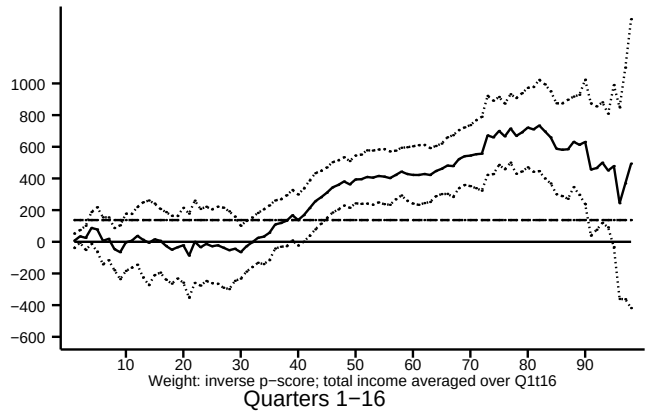


Figure 9: Quantile treatment effects by dropout status, Quarters 1–7
(person-quarter is unit of analysis; weighting done by inverse-propensity score, horizontal dashed line is mean treatment effect)

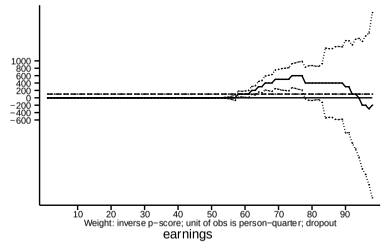
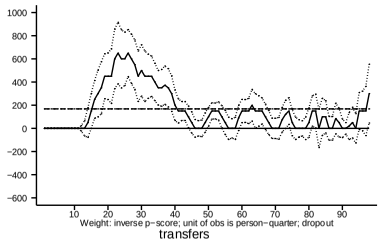
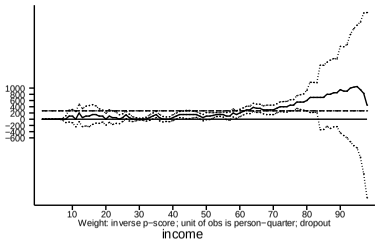
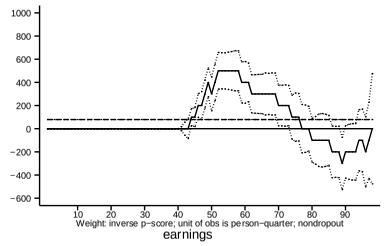
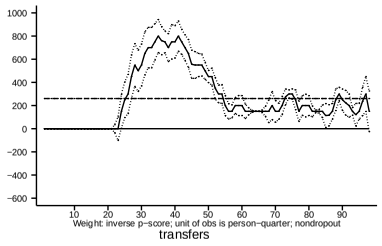
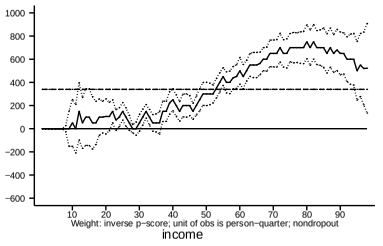
Subgroup	Earnings	Transfers	Income
Dropouts			
Nondropouts			

Figure 10: Quantile treatment effects by dropout status, Quarters 8–16
(person-quarter is unit of analysis; weighting done by inverse-propensity score, horizontal dashed line is mean treatment effect)

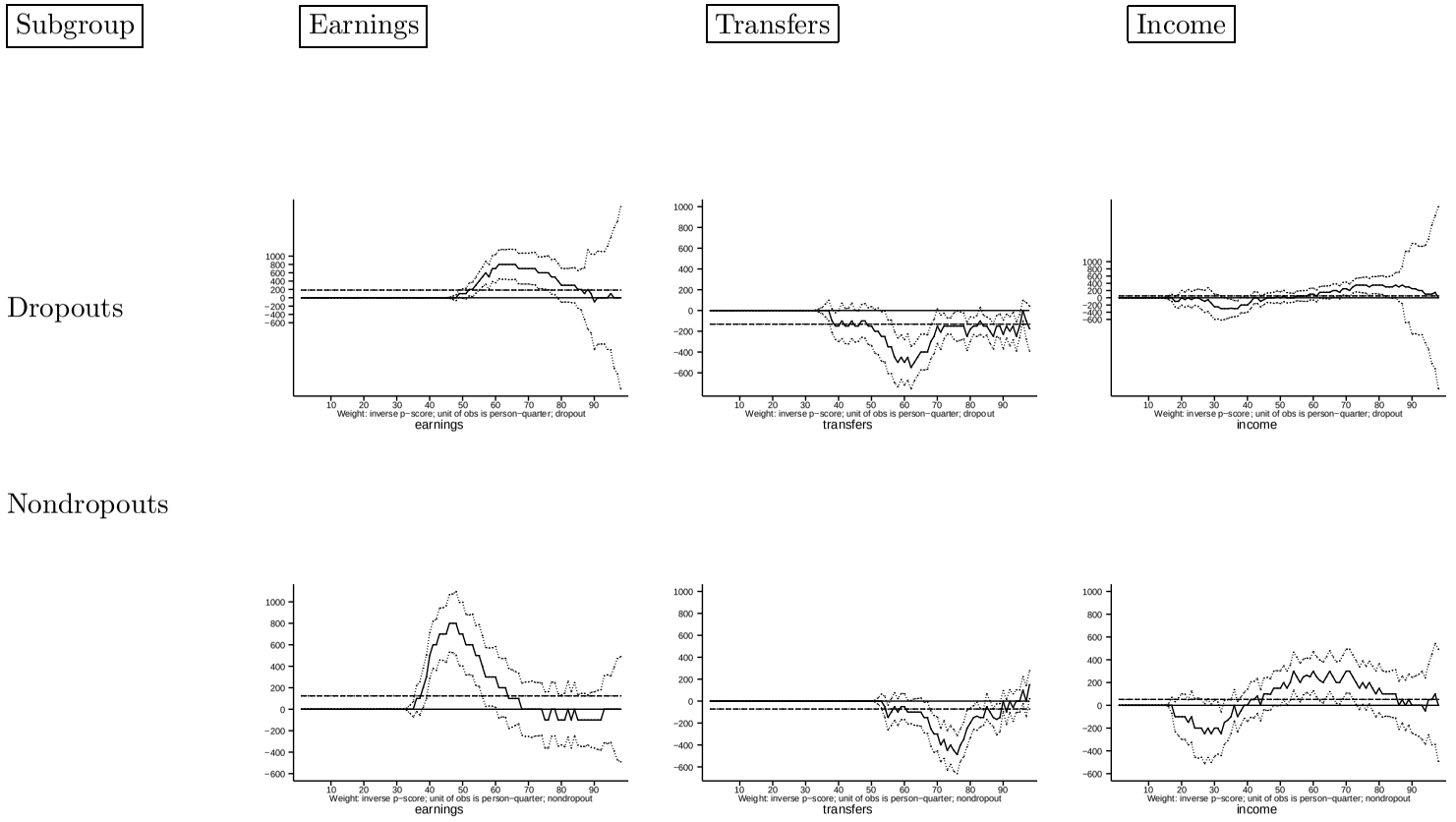


Table 1: Key differences in Jobs First and AFDC programs

	Jobs First	AFDC
Time limit	21 months (6-month extension if in compliance and non-transfer income less than maximum benefit)	*None*
Work Requirements	Mandatory work first, exempt if child < 1	Education/training, exempt if child< 2
Sanctions	1^{st} violation: 20% cut for 3 months 2^{nd} violation: 35% cut for 3 months 3^{rd} violation: grant cancelled for 3 months	*(Rarely enforced)* 1^{st}: adult removed from grant until compliant 2^{nd}: adult removed ≥ 3 months 3^{rd}: adult removed ≥ 6 months
Earnings Disregard	All earned income disregarded up to poverty line (policy also applied to food stamps)	Months 1–3: \$120+1/3 Months 4–12: \$120 Months > 12: \$90
Other policies	•Asset limit \$3000 •Partial family cap (50%) •2 years transitional Medicaid •Child care assistance •Child support pass-through	•Asset limit \$1000 •100-hour rule and work history requirement for 2-parent families •1-year transitional Medicaid

Sources: Bloom et al. (2002).

Table 2: Means for Jobs First and the 1994 national caseload

	Full sample		National
	Jobs First	AFDC	Caseload
Applicant (flow) sample	0.376	0.407	
Worked previous year	0.502	0.542	0.387
Any AFDC previous year	0.669	0.638	1.000
Race/ethnicity			
White	0.362	0.348	0.450
Black	0.368	0.371	0.367
Hispanic	0.207	0.216	0.131
Marital status			
Never married	0.654	0.661	0.444
Div/wid/sep/living apart	0.332	0.327	0.343
Education			
HS dropout	0.350	0.334	0.374
HS diploma/GED	0.583	0.604	0.377
More than HS diploma	0.063	0.058	0.249
N	2,396	2,407	765

Note: National caseload statistics were constructed using all females aged 16–54 in the 1995 March CPS who had an own child in the household and whose family was reported to have positive AFDC income for calendar year 1994. All national caseload statistics are computed using March supplementary weights. Standard deviations omitted because all variables are binary.

Table 3: Mean pre-treatment earnings, cash welfare, and food stamps statistics

	Jobs First	AFDC	Difference
Average quarterly level:			
Earnings	679	786	-107***
	(27)	(31)	(41)
Cash welfare	891	835	56**
	(16)	(16)	(23)
Food stamps	352	339	13
	(7)	(6)	(9)
N	2,396	2,407	4,803
Fraction of quarters with:			
Any earnings	0.300	0.326	-0.026*
	(0.009)	(0.010)	(0.013)
Any cash welfare	0.548	0.528	0.020
	(0.010)	(0.010)	(0.014)
Any food stamps	0.575	0.570	0.005
	(0.010)	(0.010)	(0.014)
N	2,396	2,407	4,803

Note: Standard errors in parentheses.
***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively (significance indicators provided only for difference estimates). For earnings, 8 quarters of pre-treatment data are used. For cash welfare and food stamps, only 7 quarters are available for all observations.

Table 4: Mean outcomes and impacts, weighted by inverse propensity score

	All quarters			Quarters 1–7			Quarters 8–16		
	Jobs First	AFDC	Difference	Jobs First	AFDC	Difference	Jobs First	AFDC	Difference
Average quarterly level:									
Income	2,747	2,612	135**	2,747	2,451	296***	2,748	2,738	10
	(36)	(40)	(54)	(32)	(33)	(47)	(44)	(50)	(66)
Earnings	1,659	1,566	93	1,195	1,116	79*	2,022	1,914	108
	(37)	(43)	(57)	(31)	(36)	(48)	(46)	(53)	(70)
Transfers	1,088	1,047	42*	1,552	1,335	217***	727	824	-98***
	(16)	(17)	(23)	(18)	(18)	(26)	(17)	(18)	(25)
N	2,381	2,392	4,773	2,396	2,407	4,803	2,381	2,392	4,773
Fraction of quarters with:									
Any Income	0.852	0.857	-0.005	0.865	0.862	0.003	0.809	0.820	-0.010
	(0.005)	(0.005)	(0.008)	(0.007)	(0.007)	(0.010)	(0.007)	(0.007)	(0.010)
Any Earnings	0.560	0.491	0.070***	0.569	0.486	0.083***	0.593	0.528	0.065***
	(0.007)	(0.008)	(0.011)	(0.010)	(0.010)	(0.014)	(0.008)	(0.008)	(0.012)
Any Transfers	0.626	0.622	0.004	0.684	0.644	0.040***	0.495	0.519	-0.024*
	(0.007)	(0.007)	(0.010)	(0.009)	(0.010)	(0.014)	(0.009)	(0.009)	(0.012)
N	2,381	2,392	4,773	2,396	2,407	4,803	2,381	2,392	4,773

Note: Standard errors (in parentheses) do not account for variance due to estimated nature of propensity score. Propensity score estimated using a logit model for treatment status. See text for covariates included in this model.

***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively (significance indicators provided only for impact estimates).

Table 5: Social welfare impacts under alternative choices of SWF under assumption of no consumption smoothing (*i.e.*, person-quarter is unit of analysis)

	$\rho = 0$		$\rho = 1/2$		$\rho = 1$		$\rho = 2$	
	S_A	ΔS	S_A	ΔS	S_A	ΔS	S_A	ΔS
A. Ignoring costs								
$\alpha = 0$	2,493	127	87.8	2.0	5.98	-0.02	-14.24	-0.52
	(12)	(16)	(0.2)	(0.3)	(0.02)	(0.03)	(0.18)	(0.25)
$\alpha = 0.9235$	3,937	216	106.3	3.4	6.25	-0.00	-14.24	-0.52
	(24)	(32)	(0.3)	(0.5)	(0.02)	(0.03)	(0.18)	(0.25)
B. Accounting for costs								
$\alpha = 0$	2,648	-27	94.2	-4.4	7.44	-1.47	-0.00	-14.76
	(12)	(16)	(0.2)	(0.3)	(0.01)	(0.02)	(0.00)	(0.18)
$\alpha = 0.9235$	4,092	62	112.4	-2.7	7.69	-1.44	-0.00	-14.76
	(24)	(32)	(0.3)	(0.5)	(0.01)	(0.02)	(0.00)	(0.18)

Note: Statistics calculated using inverse propensity score weighting; standard errors not corrected for use of estimated weights.
S_A is value of social welfare function for AFDC group; ΔS is social welfare function value for Jobs First group (S_J) minus this value.
Earnings are adjusted downward 7.65 percent to adjust for payroll taxes; state and federal income taxes are not accounted for.

Table 6: Social welfare impacts under alternative choices of SWF under assumption of full consumption smoothing (*i.e.*, income is averaged over all 16 quarters)

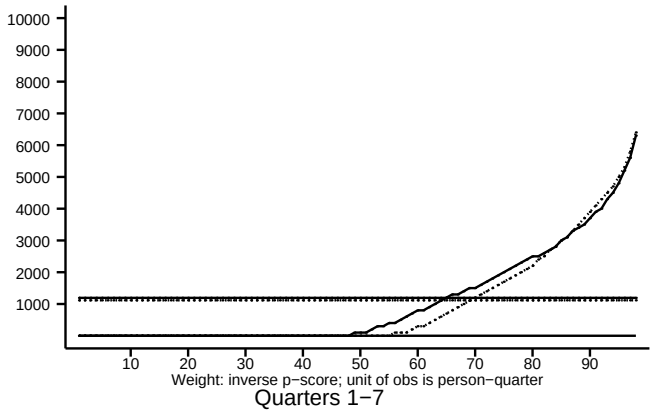
	$\rho = 0$		$\rho = 1/2$		$\rho = 1$		$\rho = 2$	
	S_A	ΔS	S_A	ΔS	S_A	ΔS	S_A	ΔS
A. Ignoring costs								
$\alpha = 0$	2,492	128	93.9	2.5	7.44	0.05	-0.94	-0.02
	(37)	(50)	(0.7)	(1.0)	(0.03)	(0.04)	(0.20)	(0.28)
$\alpha = 0.9235$	3,938	213	114.8	3.7	7.77	0.07	-0.94	-0.02
	(75)	(100)	(1.0)	(1.5)	(0.03)	(0.05)	(0.20)	(0.28)
B. Accounting for costs								
$\alpha = 0$	2,647	-27	97.9	-1.5	7.65	-0.16	-0.00	-0.96
	(37)	(50)	(0.7)	(1.0)	(0.02)	(0.03)	(0.00)	(0.20)
$\alpha = 0.9235$	4,093	59	118.3	0.2	7.97	-0.13	-0.00	-0.96
	(75)	(100)	(1.0)	(1.4)	(0.02)	(0.04)	(0.00)	(0.20)

Note: Statistics calculated using inverse propensity score weighting; standard errors not corrected for use of estimated weights.
S_A is value of social welfare function for AFDC group; ΔS is social welfare function value for Jobs First group (S_J) minus this value.
Earnings are adjusted downward 7.65 percent to adjust for payroll taxes; state and federal income taxes are not accounted for.

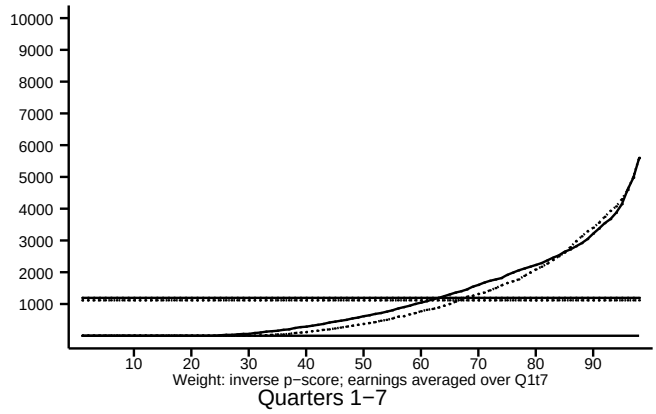
Appendix Figure 1: Inverse CDFs for earnings distributions (person-quarter is unit of observation, weighting done by inverse-propensity score, horizontal dashed line is mean treatment effect)

Person-quarter is unit of obs

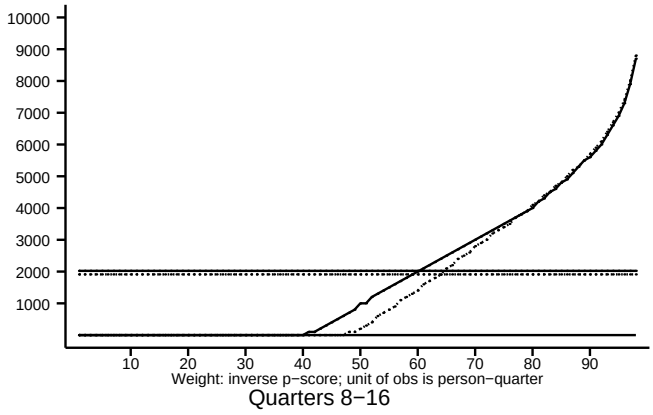
Averaged outcomes



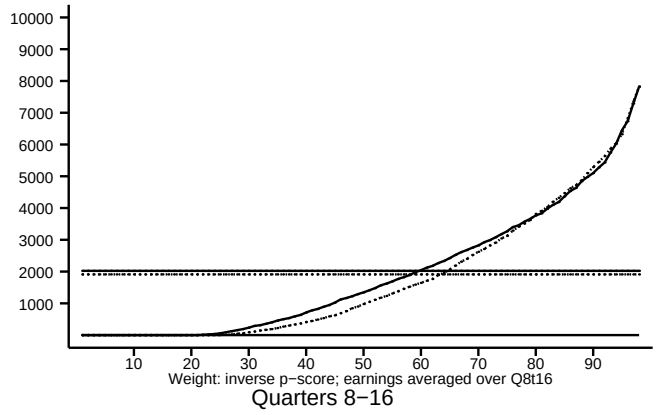
Quarters 1−7



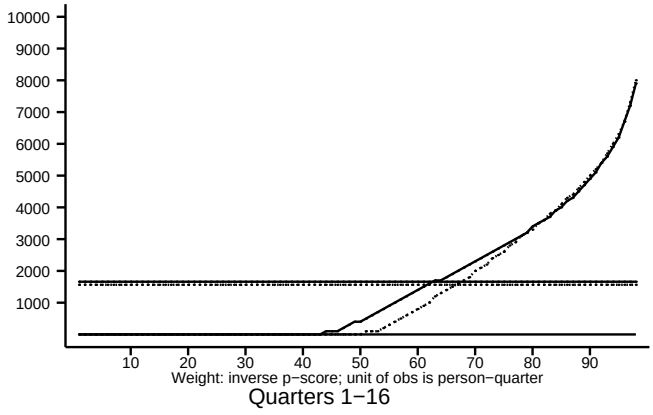
Quarters 1−7



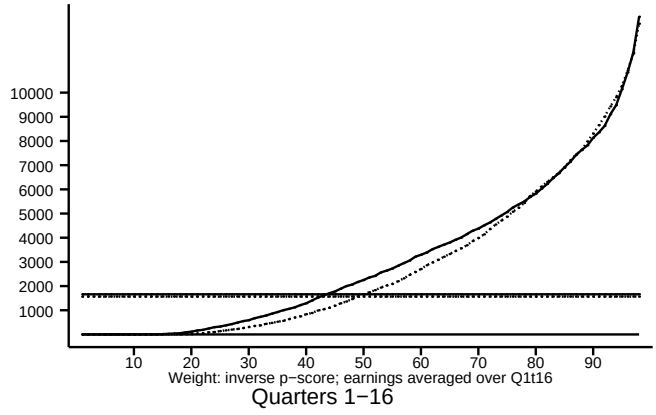
Quarters 8−16



Quarters 8−16



Quarters 1−16

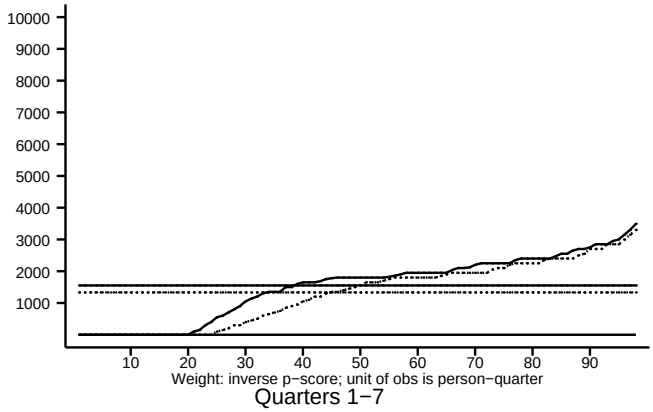


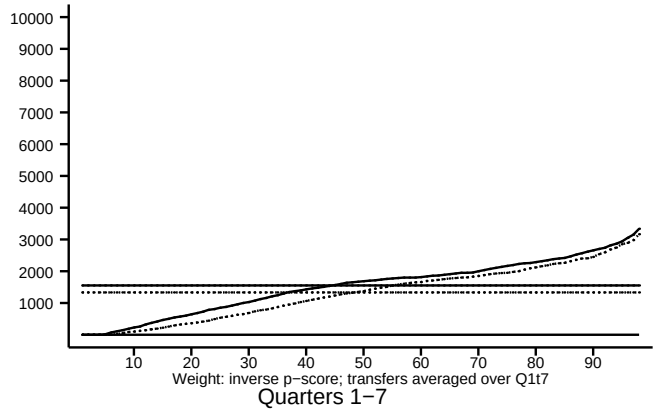
Quarters 1−16

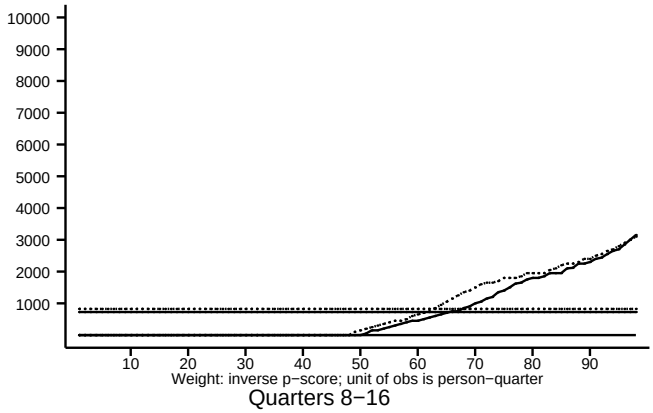
Appendix Figure 2: Inverse CDFs for transfers distributions (person-quarter is unit of observation, weighting done by inverse-propensity score, horizontal dashed line is mean treatment effect)

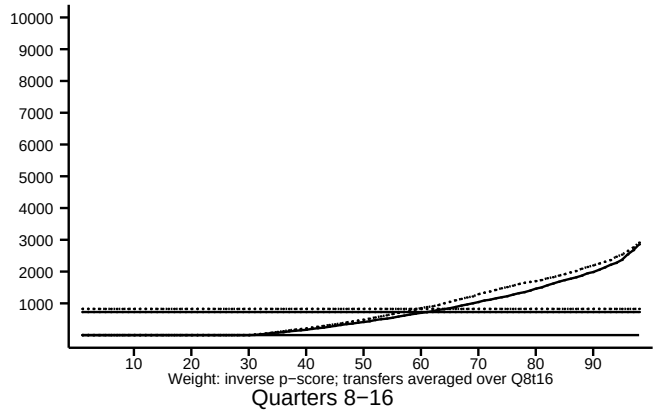
Person-quarter is unit of obs

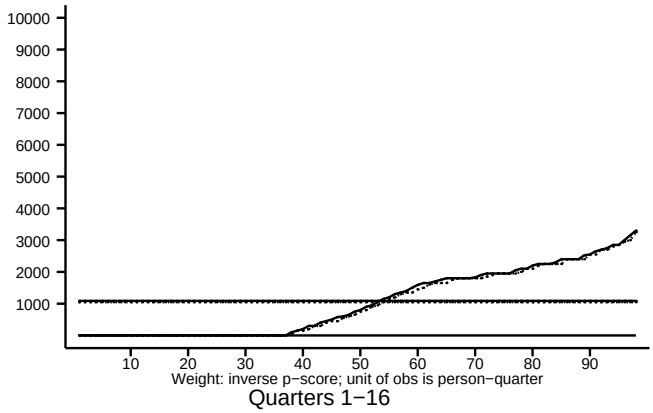
Averaged outcomes

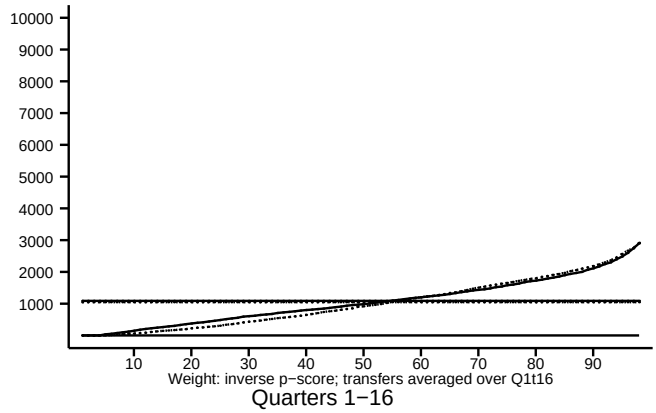












Appendix Figure 3: Inverse CDFs for income distributions (person-quarter is unit of observation, weighting done by inverse-propensity score, horizontal dashed line is mean treatment effect)

Person-quarter is unit of obs

Averaged outcomes