Confounded? Welfare Reform and the Earned Income Tax Credit in the 1990s

Adam Looney
David Eccles School of Business, University of Utah,
And The Brookings Institution

A large literature credits the Earned Income Tax Credit (EITC) with increasing labor supply and earnings, and reducing welfare use among single mothers in the 1990s, and this evidence has since played a central role in designing social policy. A longstanding concern is that these estimated effects of the EITC are confounded by contemporaneous changes in welfare policy. However, in the absence of formal tests for confoundedness, this concern is unresolved. This paper provides new tests that show that the effect of the 1993 EITC expansion in difference-in-difference (DD) estimators is spurious, and that the identifying assumptions of the DD estimator are falsified. The confounding omitted variables—proxies for exposure to welfare reform—which cause the identifying assumptions to fail, are available in the data but were not selected by researchers. Machine learning approaches to covariate selection (Chernozhukov et al 2018) readily select these covariates, and specifications that control for them suggest there is no evidence that the EITC increased employment over the 1990s.

The Earned Income Tax Credit (EITC) is credited with increasing labor supply and reducing welfare use among single mothers in the 1990s, contributing to a view that the increases in employment and declines in welfare use in that era reflected the pull of "carrots" rather than the "sticks" of punitive welfare policies. More generally, "A substantial share of the evidence regarding the EITC's labor supply effects derives from the 1993 EITC expansion" as Nichols and Rothstein (2015) write. The analysis of the 1993 expansion, in particular, is the basis for much of the evidence on the effects of the EITC on labor supply, earnings, income, poverty, self-employment, and human capital accumulation that continues to inform policymaking today. Moreover, the changes in welfare use and labor supply among single mothers in the 1990s was far greater and more enduring than changes surrounding all other EITC expansions.

An enduring concern in the analysis of the 1993 EITC expansion is that it is confounded by contemporaneous changes in cash welfare associated with "welfare reform" (see, e.g. Mead (2014, Kleven 2024). While this threat to identification is widely known, it is not clear whether or to what extent it biases the estimated effect of the EITC; the standard assessment of the "parallel trends" identifying assumption in difference-in-difference (DD) estimators—whether the pre-reform trends in treatment and control groups are parallel—"look good" in Kleven's (2024) words, and, as an assumption, it is not possible to verify whether it would have been true after the reform. And thus, the debate about whether the estimated effects of the EITC is biased is unresolved.

To answer the question of whether the effect of the EITC is confounded, I provide two tests that illustrate that the canonical difference-in-difference (DD) estimators of the effect of the EITC in the published literature are indeed biased. The first test interprets confoundedness as bias induced by an omitted variable, and uses placebo tests of the DD estimator to illustrate the bias and its magnitude. In the literature, the DD estimate is formed from the comparison of a treatment group (e.g. single mothers with two or more children, for whom the EITC was increased the most) and one or more control groups (e.g. mothers of one child or childless women), before and after the EITC was expanded. In the placebo tests, I form "treatment" and "control" groups from within equally-EITC-treated groups, and re-estimate the DD estimators as if they had been differentially treated by the 1993 EITC expansion. By construction, the resulting

treatment effects are expected to be zero (ignoring sampling error) because there is no actual difference in treatment between the placebo treatment and control groups.

Instead, these placebo tests produce large, spurious "effects" of the EITC anytime two groups with different pre-reform rates of welfare use are compared, in direct proportion to the difference in their pre-reform welfare use, suggesting that exposure to welfare use is a omitted variable that biases the EITC treatment effect. In fact, the relationship between pre-reform welfare use and the spurious placebo treatment effects is identical to the relationship between pre-reform welfare use and the actual DD treatment effects estimated in the literature, suggesting the bias is significant.

In the second test, I interpret confoundedness as arising from imbalances between the treatment and control group. Starting with the simplest 2x2 DD, in which the DD estimator is a comparison of mean outcomes of the treatment and control group before and after reform, I use classic decomposition methods to express the estimator as the sum of the causal effect of the average treatment effect on the treated plus a bias term that reflects bias from differences in the composition and treatment and control group. Intuitively, if different subgroups of the population are trending at different rates (as observed e.g. exclusively within in the control group), then if the weighting of those subgroups differs between treatment and control group, then the two groups would not be expected to trend in parallel even in the absence of any treatment effect, falsifying the identifying assumptions.

To examine this, I compare individuals with higher predicted pre-reform rates of welfare use within each treatment and control group. First, this analysis shows that individuals with higher predicted pre-reform rates of welfare use experience greater increases in employment in all groups. Second, individuals with higher pre-reform rates of welfare use are over represented in the treatment group. This suggests that differences in the composition of the two groups bias the DD estimator. In the decomposition of in the 2x2 DD, all of the estimated effect is "explained" by the not-parallel-trends bias, and none by the treatment effect on the treated. Finally, to illustrate the bias from differences in composition, I draw on the graphical comparison of trends between treated and untreated groups used to motivate the DD estimator—in this case, the relative increase in employment of treated groups—to illustrate that the changes over time are

entirely the result of differences in the composition of treatment and control groups rather than differences in outcomes of similar individuals.

Taken together, these results suggest that the estimated labor supply effects of the 1993 EITC expansion are biased by changes in behavior caused by the sweeping changes in welfare policy that culminated in "welfare reform" in 1996. Ethnographic histories of the welfare reform era (e.g. DeParle 2004 or Weaver 2000) point to the role of "caseload reduction" or "changes in the 'culture' of welfare receipt, including the behaviors and attitudes of recipients and caseworkers" (Levine and Whitmore 1997) as the most salient and important cause of the reduction in welfare use and rise in employment among single mothers in the 1990s. Since there were large differences across individuals in their use of welfare and thus exposure welfare reform, welfare reform provides a compelling alternative hypothesis to explain changes in labor supply across groups.

What is curious is that despite widespread concern of welfare reform's potential confounding effects, and the fact that the biasing omitted variables are contained in the data used by researchers, those variables were not selected as either covariates or to rebalance treatment and control groups. In short, the confounding variables were not selected by accident, oversight, or indiscretion, which is not surprising given than covariate selection is often ad-hoc, subjective, and debatable.

As an alternative to this sometimes ad-hoc approach to covariate selection, modern machine learning tools provide a more systematic approach (provided, of course, the models are allowed to choose among a sufficiently rich set covariates). As an example, I implement the double-debiased machine learning estimator of Chernozhukov et al 2018, which selects covariates from among a large set of potential variables and group by time effects (to capture time-varying shocks to specific groups). Using these models, which balance treatment and control groups on selected variables as well as controlling for those variables directly, I replicate several canonical DD papers and find little role for the EITC in the changes in behavior in the 1990s.

This paper contributes to the understanding of the labor supply effect of the EITC. The 1990s experience is central to this evidence. More recently, Kleven (2022) provides a reappraisal of the EITC literature and argues that the extensive margin effect is small based on a reanalysis of all historical EITC expansions. He too argues the evidence generated by the 1993 EITC expansion is confounded by the effect of welfare reform, and raises several "puzzles" like the "fanning-out" of employment effects by family size (which, in turn, is related to rates of pre-reform welfare use). However, it is not clear whether or how much these factors bias the estimated effect of the EITC, or how to address the bias. This paper directly examines the magnitude of the bias and addresses the failures of the identifying assumptions used in the literature (and by Kleven himself), and provides guidance on which model is correct to the exclusion of others.

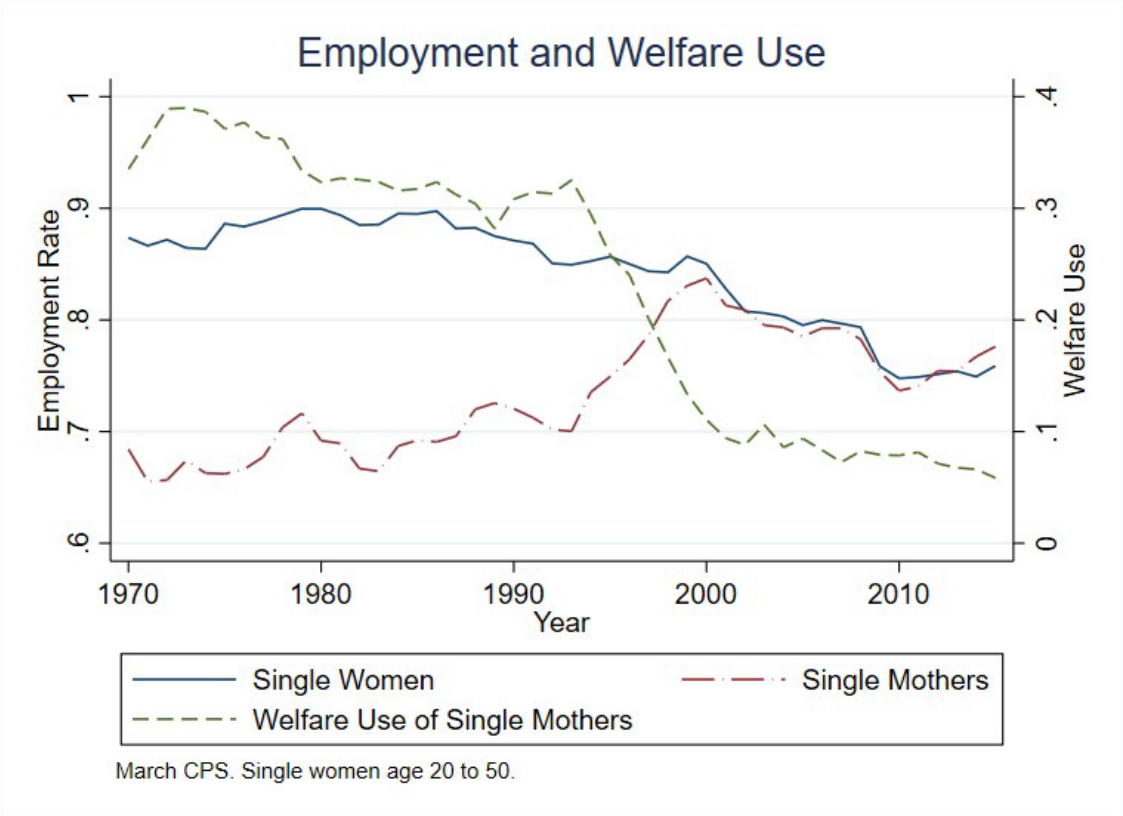
This paper proceeds as follows: The first section provides background on changes in welfare use and employment in the 1990s and the econometric evidence regarding the effects of the 1993 EITC expansion. Section two describes placebo tests of the EITC DD estimator, how they can expose omitted variables bias, and that that the canonical DD estimator produces spurious results whenever groups with different pre-reform exposure to welfare use are compared. The third section provides an alternative interpretation of confoundedness as bias arising from differences in the composition of treatment and control groups. This section illustrates that the "parallel trends" identifying assumption is falsified in the data. Having identified covariates in the data that bias existing measures, the next section uses machine learning methods to identify and address confounding factors in alternative specifications. These specifications suggest little role for the 1993 EITC in changing the labor supply of single mothers in the 1990s, and instead attribute most of the change to time effects proportionate to pre-reform welfare use—changes associated with welfare reform. Section five concludes with a discussion of the implications of this analysis for understanding the welfare reform era and contemplating contemporary debates over social policy.

1) Background: Changes in Welfare Use and Employment in the 1990s

Figure 1 shows that prior to the mid-1990s the employment rate of single women exceeded that of single mothers by an average of about 15 percentage points. That gap was entirely eliminated

over the latter half of the 1990s. At the same time, the number of women receiving welfare fell by more than two thirds.

Figure 1



A large literature suggests the 1993 EITC expansion was an important cause of this surge in employment and decline in welfare. Nichols and Rothstein's recent survey concludes that "There is an overwhelming consensus in the literature that the EITC raises single mothers' labor force participation."

The primary evidence of the effect of the EITC is derived from DD estimators, which compare mothers who were eligible for expanded EITC benefits (e.g. mothers of 2+ children) to less treated individuals (mothers of one child, who received a smaller EITC increase) to untreated (or even less treated) individuals (such as single women eligible for only a small EITC). In this strategy, the basic estimating equation has the following form:

$$y_{it} = \beta_0 + \beta_1 after_t + \beta_2 treat_i + \beta_3 (treat_i * after_t) + X\beta + \gamma Z + \varepsilon_{it} \qquad (1)$$

(1)

Where y_{it} is the outcome of interest (e.g. employment, earnings, or welfare use); $treat_i$ is an indicator (or vector of indicators) for whether the individual is in a treatment group—in this case, an indicator variable for number of EITC-eligible children: 0, 1, or 2+); $after_t$ is an indicator for time pre-or-post treatment (or, more typically, a vector of individual year effects); and X_i is a vector of individual characteristics. In addition to individual characteristics, some specifications include controls for certain time-varying policies, like welfare reform, state welfare waivers, state economic conditions, or state generosity of welfare programs either directly (e.g. as a welfare reform dummy variable indicating that welfare reform has been implemented in a state in a given year) or by including state effects or state-by-year effects.

The effect of the treatment is captured by the coefficient β_3 which measures the difference between the treatment and control group after the implementation of the reform. In the EITC literature, for example, this may be a dummy variable for having 2+ kids (whose EITC benefits expanded the most) after the expansion of the EITC (for example, post-1993). In other parameterizations, $Treatment$ and $(Treatment * After)$ are replaced with a specific simulated dollar value of the EITC, corresponding to their expected values for a typical single mother in each treatment or control group, or the average amount of taxes paid for each of those groups. In either case, identification of the coefficient of interest is generated by the same basic treatment-vs-control, before-vs-after comparison (the DD), except that the scaling and interpretation of the coefficient differs (e.g. between the average effect vs average effect per dollar).

As illustrative examples, I replicate the estimating equations of four important papers in the literature that use variations of this empirical strategy to identify the effect of the EITC. I focus

on these papers because they are representative of the published evidence on the EITC, use different parameterizations of the EITC, examine the EITC's effects on different subpopulations, and use common and well-known data from the March CPS.

These papers are: (1) Strain and Whitmore Schanzenbach 2022, which provide estimates of the effect of EITC on labor supply where the EITC effect is summarized as a dummy variable. (2) Hoynes and Patel 2018, in which the EITC variable is a simulated tax amount. (3) Meyer and Rosenbaum 2001, in which changes in the EITC are included within a "taxes paid if working" variable. For each of these three studies, the authors report estimates for subpopulations like mothers with different levels of educational attainment (e.g. less than high school or high school only). (4) Michelmore and Pilkauskas (2021) examine the effect of the EITC on the labor force participation of mothers with different-aged youngest children, modeling the effect of the EITC on labor supply by child's age using a simulated EITC amount interacted with child's age group.[1]

For each of the papers above, I replicate each analysis in a representative DD regression over the period from 1991-2001 in the March CPS, the time period after earlier EITC expansions but surrounding the 1993 EITC expansion and welfare reform.[2] I follow, as closely as possible, the sample selection choices that best represent the literature: non-married women aged 20 to 50 (excluding childless women enrolled in school). I exclude single mothers whose youngest child is older than 18. I include covariates typical in the literature: a quartic in the mother's age, dummy variables for education level (less than high school, high school, some college, college graduate), race (white, Black, other), age of youngest child (0-2, 3-5, 6-12, 13+), and whether ever married. Rather than including controls for state-level policies and local economic characteristics, like the level of the minimum wage, generosity of the welfare system, or state unemployment rate (for which covariate estimates are not central to this analysis), I include

[1] The literature on the behavioral response to the EITC originates with Eissa and Liebman's 1996 analysis of the 1986 expansion, and includes a large literature examining earlier reforms (e.g. the 1975 expansion (Bastian 2020) and more recent expansions. Because these studies do not focus on the 1990s welfare reform era, I exclude them from this analysis.

[2] Note that the period of analysis in Meyer and Rosenbaum 2001 is from 1984-1996, and thus only partially overlaps with the period of analysis in this paper. Hoynes and Patel (2018) study 1991-1998. In their analysis of the 1993 expansion, Strain and Whitmore Schanzenbach (2022) examine 1989-1998. Michelmore and Pilkauskas (2021) study the period from 1990-2000.

state-by-year fixed effects. [Summary statistics for this well-known dataset will be included in the appendix.] In the analysis presented below, the dependent variable is employment last year (defined as having positive earnings), however the overall analysis is similar if the dependent variable were welfare use last year or annual earnings.

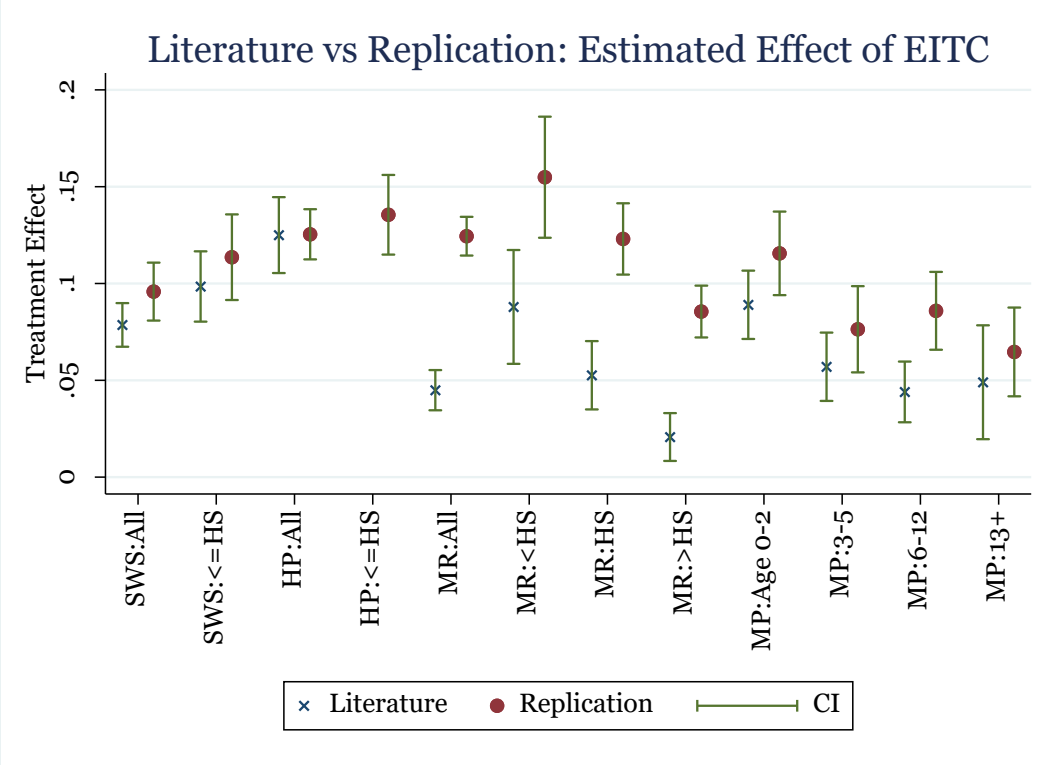
In the replication of Strain and Whitmore Schanzenbach (2022), the EITC expansion is represented as a dummy variable that is equal to zero before 1994, equal to one for mothers with one and two children after 1995; 0.92 for mothers of one child and 0.5 for mothers of two children in 1994 (reflecting the fact that the EITC expansion was partially phased in that year); and equal to one for mothers of one child and 0.78 for mothers of two children in 1995. In the replication of Hoynes and Patel (2018), the EITC expansion is represented as a simulated value of the EITC. In the replication of Meyer and Rosenbaum (2001) it is the level of taxes if working. In these cases, the value of the EITC varies by year and number of children (0, 1, or 2+ children). When examining the differential effect of the EITC on labor supply by child's age (Michelmore and Pilkauskas 2021), I specify the simulated EITC treatment variable interacted with the group dummies for age of youngest child. In all cases, the identification is at the tax year and family size level (or year, family size, and age of youngest child level).

Figure 2 summarizes the results of this replication and compares the coefficient estimates to those in the published literature. For each specification, my replication of the DD estimator produces economically and statistically significant coefficient estimates. Despite somewhat different time periods, control variables, and treatment variables, the replication closely resembles published estimates.[3] In other words, these DD estimators strongly imply that the EITC had large effects on employment.

In each of these estimates, the common identifying assumption is that the outcomes of treatment and control individuals would have been the same in the absence of the EITC after controlling for time effects, group, and other characteristics of sample individuals and the applicable economic or policy environment.

[3] The exception is that the estimated effect of the "taxes if work" variable from Meyer & Rosenbaum is larger than estimated in their analysis because of the difference in the period of time they analyze. Meyer and & Rosenbaum exclude the post 1996 period in part because of concerns it is confounded by the effects of welfare reform.

Figure 2



Note: The figure replicates the central specifications of the effect of the EITC of Strain and Whitmore Schazenbach 2021 (SWS); Hoynes and Patel 2018 (HP); Meyer and Rosenbaum 2001 (MR); and Michelmore and Pilkauskas 2021 (MP) within a consistent March CPS sample from 1991-2001 within the sub-samples indicated on the x-axis, and compares the replicated estimates to published estimates. (HP does not provide an estimate of the effect of the EITC on a high-school-only population, which is why it is omitted.)

The key threat to this identifying assumption is the effect of contemporaneous changes to welfare policy ("welfare reform"). Aid to Families with Dependent Children (AFDC) was "ended as we know it" in the words of then-President Clinton, with the passage of the Personal Responsibility And Work Opportunity Reconciliation Act of 1996. Even before 1996, many states had sought and received waivers to change AFDC parameters to incentivize employment or sanction non-compliance with program rules. Ethnographies of the welfare reform era suggest a major role for the associated changes in bureaucratic processes and attitudes rather than employment subsidies like the EITC. The Rockefeller Institute's implementation study, for example, found that many local sites had reoriented their efforts toward diverting would-be recipients away from welfare and toward work; "local officials and workers view state officials as wanting to see, above all else, lower caseloads" (as quoted in Kent Weaver 2000). In areas with the most dramatic reductions in welfare use, like Wisconsin, where the number of program recipients declined from

226,197 in 1994 to 96,054 in 1997, ethnographers attributed the changes to work requirements enforced by new bureaucratic procedures (DeParle 2005).

The essential empirical content of these historical narratives is that exposure to bureaucratic changes and changing social norms was proportional to exposure to the welfare system; groups of women with higher rates of participation in the early 1990s were "treated" more by welfare reform than those who had little interaction with the welfare system. Hence, the econometric concern is that controlling for the average effect of welfare reform, e.g. with state-by-year effects or other state-level covariates that vary by year do not address the heterogeneity in exposure across groups within state and year.

Indeed, this is the substance of Kleven (2019)'s concern that the labor supply effects of the EITC in the 1990s are confounded. Pre-reform indicators of welfare use are stronger predictors of the increase in employment over the 1990s than EITC-qualifying number of children (Looney and Manoli 2014; Kleven 2019). While the potential confounding effect of welfare reform is longstanding, and Kleven's reassessment of the 1990s experience is rhetorically compelling, it has been unpersuasive. For instance, in their rejoinder to Kleven, Strain and Whitmore Schanzenbach (2021) re-estimate canonical DD estimators of the effect of the EITC in the 1990s (and around other reforms), make methodological choices that are standard in the literature, and again find large coefficient estimates. In effect, the literature's consensus was and remains that the standard tests of the identifying assumptions (e.g. parallel pre-reform trends) and demographic and state-level covariates are adequate to address the potential bias from contemporaneous welfare changes. Moreover, there is not clear econometric guidance how to adjudicate such disagreements.

2) Placebo Tests of the Difference in Difference Estimator

As their name implies, placebo tests are used to test whether an estimator produces a treatment effect even in the absence of actual treatment. I use placebo tests to assess whether the EITC DD estimator produces a spurious treatment effect, and, if so, to diagnose the source of the spurious effect.

Consider again the following canonical DD regression estimator in (1) modified to include a potential omitted variable *Z* which is correlated with *treatment*after* but is omitted from the estimated model:

$$y_{it} = \beta_0 + \beta_1 after_t + \beta_2 treat_i + \beta_3 (treat_i * after_t) + X\beta + \gamma Z + \varepsilon_{it}$$

Estimating equation (1) without including *Z* results is classic omitted variables bias:

$$\beta_{hat} = \beta_3 + \gamma \text{cov}(treatment * after, \text{Z}) / \text{var}(\text{treatment} * \text{after})$$

When the *treatment*after* effect is a dummy variable, then the bias term simplifies to $\gamma(\Delta Z)$ which is the effect of the omitted variable on *y* times the difference in the mean of the omitted variable between the two groups ΔZ.

$$\beta_{hat} = \beta_3 + \gamma(\Delta Z) \qquad (2)$$

Omitted variables bias is typically a concern in cases in which data on the omitted variable is unmeasured or otherwise unavailable to the researcher. The present case is unusual in that the potential confounding factor—exposure to welfare reform policies—is known and is measurable in the data. This suggests a two-step placebo test procedure: first, estimate the DD estimator in a circumstance when no treatment effect is expected—within groups of identically-treated individuals. Second, assess whether any placebo treatment effects are correlated with observed but omitted characteristics of the placebo treatment and control groups, like their predicted pre-reform propensity to receive welfare.

By estimating the model on placebo treatment and control groups whose true treatment is the same, there should be no treatment effect, and thus β_3 should be zero. If the estimates of β_{hat} are non-zero (beyond that expected from sampling error), then that suggests bias. Moreover, to the extent that the estimated placebo effects (β_{hat}) are correlated with characteristics of the sample that are omitted from the regression, that indicates the source of the bias.

The Placebo Method

I assign placebo treatment and control group status to groups formed from individuals that were treated equally by the EITC (e.g. mothers of one child), and then re-estimate the key test statistic (the DD estimate of the treatment effect) that estimate the "treatment" of one subgroup of mothers of one child to another "control" subgroup of mothers with one child. I estimate these placebo DD models *within* treatment and control groups to deliberately exclude any comparison between differently-treated individuals.

In these placebo tests, the EITC treatment effect is the EITC indicator variable constructed following Strain and Whitmore Schanzenbach (2022). In each specification I estimate the effect of the EITC treatment from a pairwise comparison between one randomly defined placebo treatment group and one placebo control group both drawn from the same actual treatment or control group. Following the structure of the policy (in which treatment assignment is based on a common demographic characteristic: number of children), I form placebo groups based on groups with similar pre-reform characteristics.

Because the posited confounding variable is related to welfare reform, in the analysis below, I form placebo groups based on demographic characteristics that predict similar pre-reform exposure to welfare use, nonemployment, or earnings. Specifically, I estimate the predicted likelihood of welfare use and nonemployment using a probit estimator on the same covariates described in the discussion of equation (1) including demographic characteristics, educational attainment, state of residence, and family composition, and also a predicted log wage using the same characteristics in linear regression. I estimate these relationships exclusively in data from 1991, 1992, and 1993, before the major expansion of the EITC and the implementation of significant welfare reform changes.[4]

I predict a pre-reform rate of welfare use, nonemployment, and wage for each individual in the sample, and form groups defined by deciles of each variable within each EITC treatment group (2+ children, one child, no children). Within each EITC treatment group, I assign one placebo

[4] The analysis is insensitive to choice of base year, because the predicted outcome is based on time invariant characteristics.

group to be treatment and another to be control, estimate the DD regression, and capture the coefficient estimate. I repeat this for all 90 potential permutations of the pairwise comparisons within each of the three EITC treatment groups. And I repeat the exercise three times, one for each set of deciles formed by predicted welfare use, nonemployment, and earnings. (There are many ways to form these placebo groups, such as simply forming cells by arbitrary combinations of demographic characteristics; having tried many approaches the outcome of the analysis is the same.)

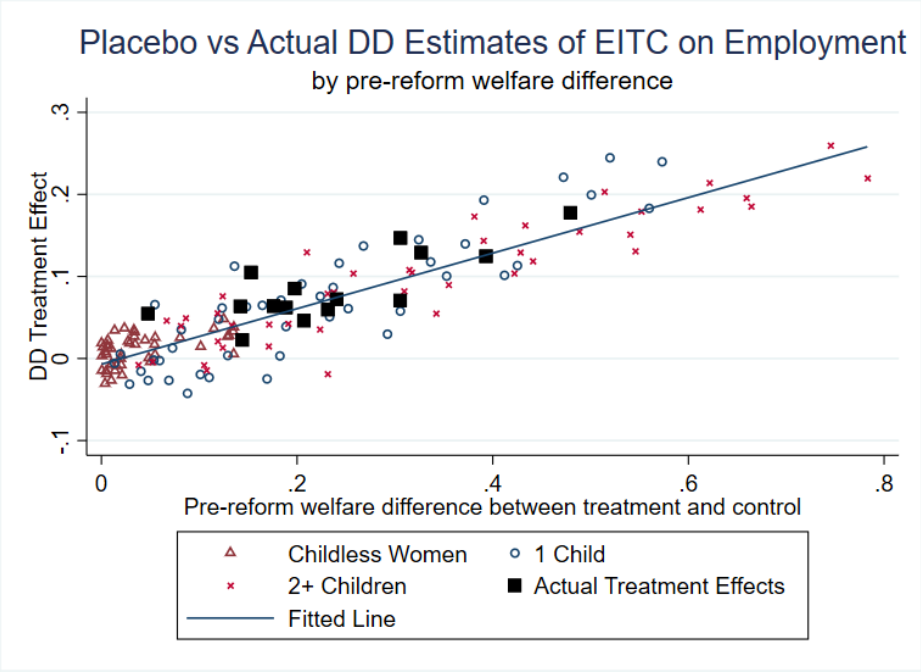
For each DD estimator, I also measure the difference in pre-reform welfare use and employment between the placebo treatment and control groups, to assess whether the placebo coefficient estimates vary systematically with (omitted) characteristics of the two groups related to these pre-reform characteristics.

For comparison to these placebo estimates, I estimate actual DD estimates based on the samples and comparisons in the four papers replicated above. Because these papers include more than two total treatment and control groups (e.g. mothers with 2+ children, one child, and childless women), I estimate all pairwise comparisons of treatment/control groups identified in these papers; the total EITC treatment effect is a weighted average of these pairwise comparisons (Goodman-Bacon 2021). Collectively, these include comparisons across women with two or more, one, or zero children; all women and women with different levels of educational attainment (less than high school, high school, more than high school); and mothers with different-aged youngest children (age 0-2, 3-5, 6-12, 13+).

Results of the Placebo Tests

Figure 3 presents each placebo estimate as a function of the difference in the predicted pre-reform rates of welfare use between the treatment and control group. The figure only presents results from the permutations where the placebo treatment group's rate of welfare use exceeds that of the control group; the other permutations are simply a mirror image of this figure across the origin.

Figure 3



Note: The figure presents placebo and actual DD estimates arrayed by the gap in predicted pre-reform rates of welfare use between each treatment and control group for the permutations in which the treatment group's predicted welfare use exceeds that of the control group, and disaggregates the placebo effects based on whether they are estimated by pairwise comparisons of childless women, mothers with one child, and mothers of 2 or more children.

First, Figure 3 shows that the placebo effects are often large and economically important—suggesting that "treated" groups experienced large increases in employment compared to the "control" group, even though these individuals are treated equally by the EITC.

Second, the magnitude of the placebo effects is linearly increasing in the difference in the pre-reform rates of welfare of the two groups being compared. Anytime a group with a higher rate of welfare use is compared with a group with lower rates of welfare use, the placebo estimates are biased toward finding larger treatment effects. However, when the two groups being compared have similar pre-reform rates of welfare use, the estimated treatment effect is close to zero, as indicated by the intercept of the fitted line.

Figure 3 also disaggregates the placebo effects into those from comparisons within childless women, mothers of one child, and mothers of 2+ children. The figure shows that within each group, the relationship between the placebo effect and the gap in welfare use is the same.

Finally, the black squares, which represent the actual DD estimates (comparisons across groups whose actual EITC treatment was different) from the replicated literature, follow the same pattern. Knowing only the slope of the placebo-fitted line and the gap in pre-reform welfare use between actual treatment and control group, one can accurately predict what the actual DD coefficient will be. In fact, in a regression of the actual DD estimates on the values predicted by the fitted line in Figure 3, yields a coefficient of one and an intercept of zero.

In summary, the placebo tests suggest that the DD estimator is indeed biased by a confounding effect that is proportional to pre-reform exposure to welfare use.

3) How much of the estimated EITC treatment effect arises because of differences in the exposure to welfare reform across groups?

The above analysis has interpreted confoundedness as arising from an omitted variable. Another useful way to interpret confoundedness in this context is as arising because of imbalances in the composition of the treatment and control groups. Consider the following decomposition of the simplest 2x2 DD estimator, where *Y* indicates the outcome of interest (such as employment), superscripts *T, C* index treatment and control group, and the subscripts *0,1* index time before and after treatment.

$$DD = (Y_1^T - Y_0^T) - (Y_1^C - Y_0^C)$$

$$DD = \Delta Y^T - \Delta Y^C$$

Now decompose the mean outcomes of the treatment and control group into subgroups indexed by *s* where subgroups are formed, for example, by demographic characteristics or by deciles of the ex-ante propensity to use welfare, and the relative weight of each subgroup in the treatment group is given by ω_s^T. The DD estimator can be restated as:

$$DD = \sum \omega_s^T \Delta Y_s^T - \sum \omega_s^C \Delta Y_s^C$$

Adding and subtracting the change in the control group weighted by the treatment group weights ($\omega_s^T \Delta Y_s^C$) (as in a basic Kitawaga-Oxaca-Blinder decomposition) and rearranging gives:

$$DD = \sum \omega_s^T (\Delta Y_s^T - \Delta Y_s^C) + \sum (\omega_s^T - \omega_s^C) \Delta Y_s^C \quad (3)$$

The first component represents the weighted sum of the "treatment effects" calculated as the difference-in-difference in the outcome within each subgroup *s*, where the weights represent the fraction of the treatment group belonging to subgroup *s*.

The second component represents the quantity "explained" by differences in the composition of the treatment and control group holding fixed the coefficient estimates (the time trend) estimated within the control group. Note that in this decomposition, this latter term contains no information on the post-reform outcome of the treatment group.

If the latter component in (3) is not zero, however, then even in the absence of any treatment effect, the counterfactual outcomes of the two groups would not be parallel. Intuitively, when treatment and control groups differ in their composition, the examination of whether trends of subgroups within group diverge instead of evolve in parallel can provide an indication of whether the "parallel trends" identifying assumption holds.

Specifically, under the identifying assumptions of the DD estimator, this second component should *always* sum to zero, however subgroups are defined.[5] Given the decomposition, the component can be zero in two intuitive scenarios: if (1) the trend is the same within all subgroups of the control group (ΔY_s^C is a constant for all subgroups *s*) or (2) the weights in the treatment and control group are identical (so that $(\omega_s^T - \omega_s^C)$ is zero for all *s*).[6]

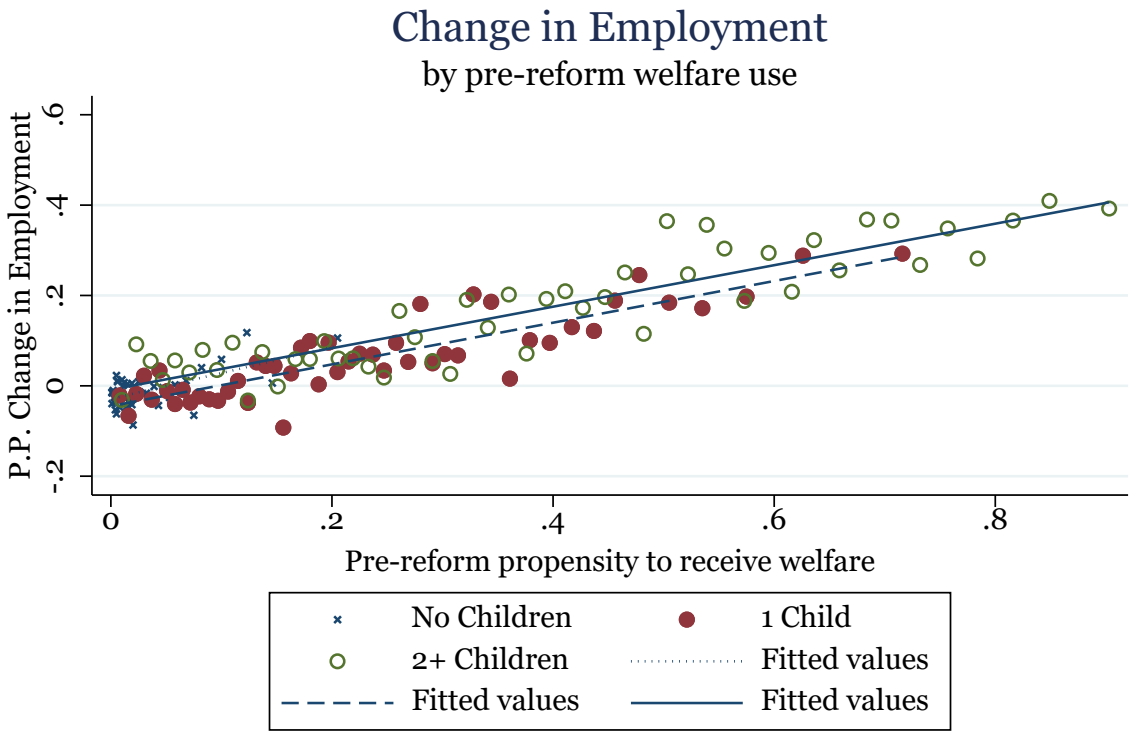
To assess whether the bias term in (3) is likely to be zero, Figure 4 ranks single mothers by percentiles of pre-reform propensity to use welfare (the subgroups *s*) and separates the sample into the treatment and control groups (mothers with two or more children, mothers with one

[5] This is a relatively weak assumption—it might be realistic to make stronger assumptions on subgroups such that subgroup trends should be parallel between treatment and control groups.

[6] A third case, which I do not explore, arises if the imbalance in weights is inversely correlated with the subgroup trends such that the sum of the weighted trends is zero. While this does not imply a violation of the parallel trends assumption, it may imply that additional scrutiny of the research design or into heterogeneous treatment effects may be warranted.

child, childless women). For each treatment group X subgroup sample, the figure presents the percentage point change in employment between 1993 and 2000.

Figure 4



Source: March CPS 1992-2001. Average change in employment before and after reform by percentile of predicted pre-reform propensity to use welfare.

First, Figure 4 shows that the change in employment is strongly related to the predicted pre-reform rate of welfare use for each subgroup. Groups with low rates of pre-reform welfare use saw essentially no increase in employment, whereas groups with high rates of welfare use experienced increases in employment close to 40 percentage points. In other words, subgroups of the treatment and control groups are indeed "trending" at different rates.

Second, note that the treatment and control groups are not equally weighted in their composition. Mothers with two or more children are overrepresented in high-welfare use groups, and mothers with one child, and childless women are overrepresented among groups with low pre-reform rates of welfare use. As a result, the bias term in (3) is not likely to be zero—distinct, identifiable subgroups are trending at different rates within group, and these subgroups are weighted

differently across group, which means that even in the absence of treatment, there should be no expectation that the groups should trend in parallel.

How large is the resulting bias? Figure 4 suggests the change in employment is similar across treatment groups within percentiles of welfare use (subgroups). Among women with similar rates of pre-reform welfare use, the change in employment is similar whether they have two or more children, one child, or no children. Specifically, the fitted regression lines for each group are similar to each other, and their y-intercepts are close to zero, indicating that there is little difference in employment across groups after controlling for pre-reform rates of welfare use. In all, this analysis suggests the DD estimate is primarily the result of bias rather than the treatment effect of the EITC. In other words, using the decomposition in (3), not only is the compositional bias term not zero, it appears to "explain" all of the DD estimate.

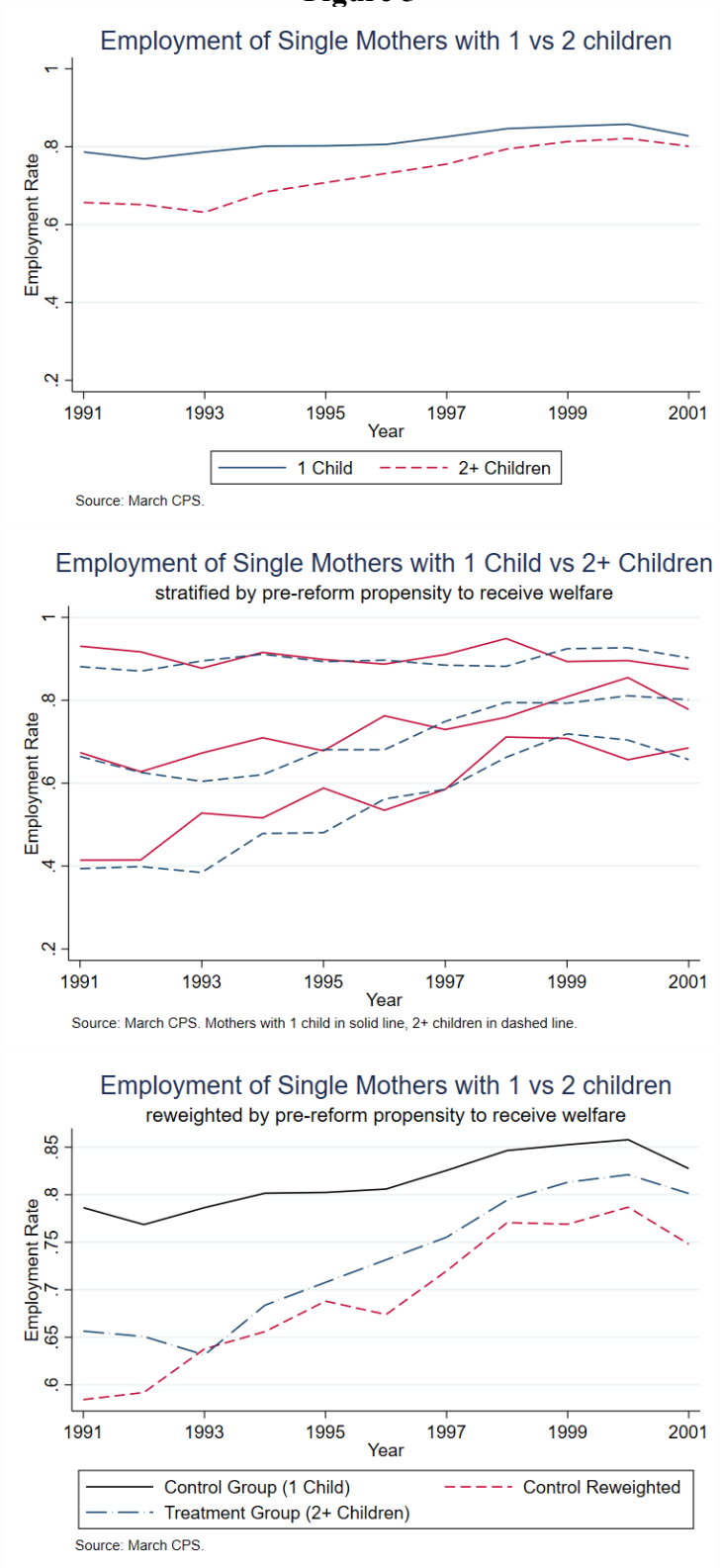
To illustrate this, the top panel of Figure 5 starts with the standard graphical illustration of the DD estimator comparing the employment rate of mothers of 2+ children and mothers with one child. The convergence in the employment rates of these groups after the 1993 expansion is the source of the estimated effect of the EITC in the DD estimator.

The middle panel of Figure 5 decomposes the treatment and control group into subgroups formed by quintiles of predicted pre-reform welfare use, similar to figure 4. For clarity, the figure presents the annual employment rates of women in three of the five quintiles: the lowest predicted quintile of welfare use in 1993, the fourth highest, and the highest quintile.[7]

Figure 5 illustrates the considerable heterogeneity in subgroup trends within the groups. Whereas there is little change over time in employment of groups of mothers with low pre-reform rates of welfare use, employment surges among those with higher rates. Moreover, the time pattern of employment is extremely similar for individuals in the treatment and control group conditional on propensity to use welfare.

[7] The lowest three quintiles have similar rates of predicted welfare use and thus the second and third are omitted for graphical clarity.

Figure 5



Note: Each panel compares the treatment group (blue) to the control group (red). The top panel provides the unweighted data. The second panel stratifies each sample by quintile of predicted pre-reform welfare use (as a proxy for exposure to welfare reform) and compares treatment- and

control-group mothers with similar pre-reform rates of welfare use from the first, fourth, and fifth quintiles. The bottom panel re-weights the control group to have the same average predicted pre-reform exposure to welfare use as the treatment group.

The bottom panel reweights the control group observations (represented unweighted as the black line) to match the pre-reform rates of welfare use observed in the treatment group (presented as the red dashed line), and compares the change in employment to that of the treatment group (the blue dashed line). The reweighted control group closely follows the employment rates of the treatment group. In fact, the reweighting of the control group more than explains the increase in employment over this period, suggesting that all of the difference-in-difference (the source of the variation in the DD estimator) is associated with changes in employment of groups with different exposure to welfare.

4) Identifying and Controlling for Confounding Effects Using Machine Learning

Researchers have long known that welfare reform was a potential confounder in measuring the estimated effect of the EITC in the 1990s, and the analysis above suggests that the confounding variables were available in the data widely used in the literature, which raises the question of whether conventional approaches to identifying covariates to control for or balance are effective. Candidly, the choice of which covariates to include and their functional form is often ad hoc. Even when researchers take care to re-balance treatment and control groups by matching or reweighting, it is not obvious which variables are important to balance.

Moreover, most approaches to matching exclusively use pre-reform information, which may exclude confounding effects. For instance, in synthetic control methods (e.g. Abadie and Gardeazabal 2003) or synthetic difference-in-difference methods (Arkhangelsky et al. 2021), treatment and control groups are reweighted to produce pre-reform trends that are parallel based on the relationship between time trends and characteristics in the pre-reform period. These methods deliberately exclude post-reform information, and as a result, they may exclude covariates that are unimportant in explaining time trends in the pre-reform period but are important in the post-reform period (confounders).

Rather than guess at which covariates or confounding factors are important, I implement off-the-shelf machine learning models to select covariates. While this does not eliminate subjectivity,

because the researcher must offer the model a set of covariates to select among, it systematizes the process to a degree.

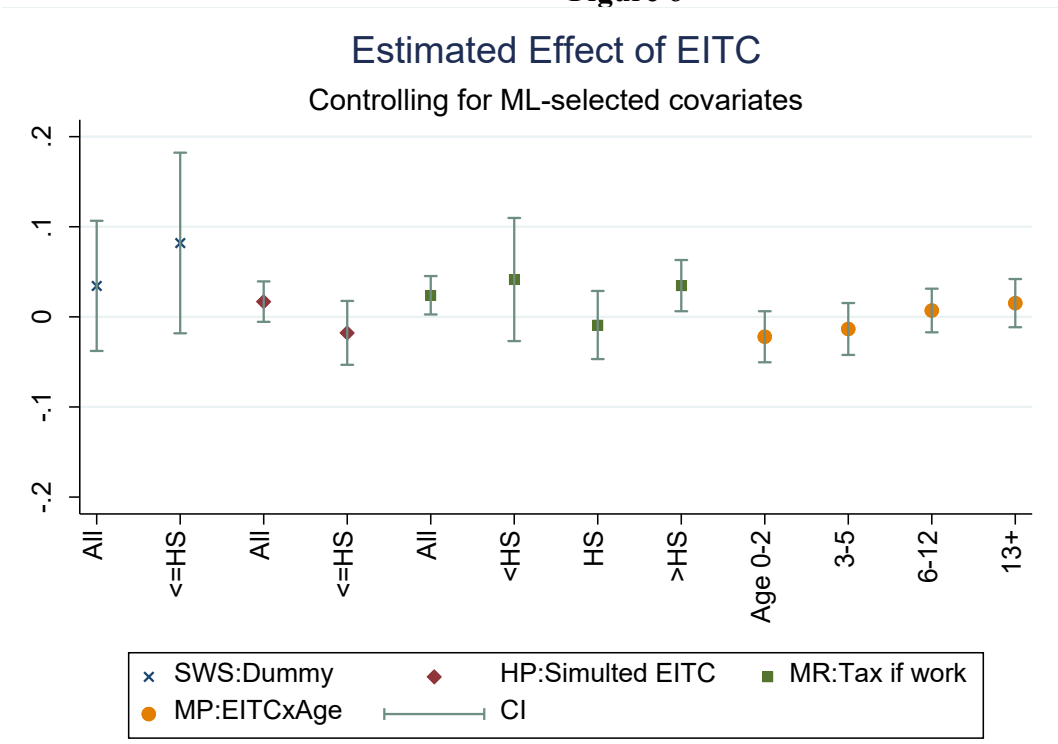
Specifically, I use the double debiased machine learning (DDML) algorithm proposed in Chernozhukov et al 2018. In short, I generate a large set of potential covariates (e.g. marital status, age, educational attainment, age of youngest child, state by year effects, race; predicted pre-reform welfare use and employment) and interaction terms between demographic characteristics, including pre-reform welfare use, and time effects. The model then selects those covariates that have the highest predictive explanatory power for the outcome to include as controls, and for predicting treatment status, to use to balance treatment and control groups. The model estimated using three ML modes: Lasso regression, random forest, and Gradient Boosted Trees.

The machine learning approach identifies pre-reform welfare use and time effects in pre-reform welfare use (as well as other indicators of welfare use, like age of youngest child) as important predictors of both changes in employment and treatment status. This isn't surprising given the graphical evidence presented earlier, but it is reassuring that an off-the-shelf algorithmic approach identifies the confounding variables directly.

Results Controlling for Confounding Effects of Welfare Reform

Figure 6 presents the DDML estimates of the effect of the EITC including controls and reweighting for covariates identified by the model. In contrast to the replication in Figure 2, the estimated effects are attenuated toward zero and typically not statistically significantly different from zero. In short, after controlling for these confounding factors, the EITC appears not to have increased the labor supply of single mothers.

Figure 6



Note: These figures present the estimated DDML effects of the EITC in the same specifications presented in Figure 2.

Have these estimators been "debiased" by the inclusion of these covariates? To test this, I replicate the placebo tests above with the specification that includes the ML-selected covariates. The resulting placebo estimates have a mean of zero and are uncorrelated with pre-reform rates of welfare use, suggesting the bias has been eliminated.

5) Conclusion

The analysis in this paper suggests that the estimated effect of the 1993 Earned Income Tax Credit is biased by confounding effects of welfare reform, and that the identifying assumptions in commonly used DD estimators are falsified in the data. One contribution of the paper is to propose methods to assess identifying assumption in such models; while identifying assumptions obviously cannot be proved to be true, they should at least not be contradicted in data available to researchers.

The analysis of this paper also suggests that welfare reform policies had a larger impact on employment (and likely, other outcomes often attributed to the 1993 EITC expansion, like

changes in welfare use, earnings, poverty, income, and human capital accumulation) than previously believed. Ethnographic narratives suggest that changes in the culture of welfare and the efforts to reduce caseloads were important contributors to changes in employment and welfare use; this paper suggests that exposure to such policies, as measured by predicted pre-reform rates of welfare use, explain a great deal of the change in employment during this time period.

If the changes in employment were caused by efforts to reduce caseloads and work requirements—the "sticks" of welfare reform—rather than the "carrot" of the EITC, that implies that the changes in the wellbeing of affected single mothers was more negative than generally believed, particularly among single mothers with infants and young children who were the most affected by these changes. Dissatisfaction with these changes may be a contributor to recent calls by advocates and policymakers to restore cash benefits that are not conditioned on work, ranging from universal basic income to paid parental leave.

Finally, the evidence of the apparent success of the 1993 EITC expansion in increasing employment and earnings, reducing poverty and welfare use, and improving the economic environment of children continues to reverberate in policy debates over the design of related social policies like the child tax credit. Supported by published evidence of the efficacy of the 1993 EITC expansion, "virtually all gains in spending on the social safety net for children since 1990 have gone to families with earnings, and to families with income above the poverty line" according to Hoynes and Whitmore Schanzenbach (2018). And recent efforts to increase transfers to non-working families remain contentious (see, e.g. Corinth et al. 2021). This paper suggests that the changes in employment in the 1990s were attributable more to the withdrawal of welfare than to other factors, indicating both that the policy was effective in changing behavior, but also that it reduced the wellbeing of affected families. This implies that the tradeoffs in designing social policies are more difficult than previously believed.

References [Incomplete]

Corinth, Kevin, Bruce Meyer, Matthew Stadnicki, and Derek Wu. 2021. "The Anti-Poverty, Targeting, and Labor Supply Effects of the Proposed Child Tax Credit Expansion." Becker Friedman Institute for Economics Working Paper No. 2021-115. https://bfi.uchicago.edu/working-paper/2021-115/.

DeParle, Jason. *American Dream: Three Women, Ten Kids, and a Nation's Drive to End Welfare*. Viking, 2004.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*, Volume 225, Issue 2, Pages 254-277, ISSN 0304-4076, https://doi.org/10.1016/j.jeconom.2021.03.014.

Grogger, Jeffrey. 2003. "The Effects of Time Limits, the EITC, and Other Policy Changes on Welfare Use, Work, and Income Among Female-Headed Families." *Review of Economics and Statistics*, 85(2): 394-408.

Grogger, Jeffrey. 2004. "Time limits and welfare use." *Journal of Human Resources* 39.2: 405-424.

Hoynes, H.W. and Patel, A.J., 2018. Effective policy for reducing poverty and inequality? The Earned Income Tax Credit and the distribution of income. *Journal of Human Resources*, *53*(4), pp.859-890.

Hoynes, Hilary W., and Diane Whitmore Schanzenbach. 2018. "Safety Net Investments in Children." *Brookings Papers on Economic Activity*, Spring, 89–150.

Kleven, H., 2024. "The EITC and the Extensive Margin: A Reappraisal." Forthcoming, Journal of Public Economics.

Levine, P.B. and Whitmore, D.M., 1997, January. The impact of welfare reform on the AFDC caseload. In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association* (Vol. 90, pp. 24-33). National Tax Association.

Looney, A. and Manoli, D., 2016. Are there returns to experience at low-skill jobs? Evidence from single mothers in the United States over the 1990s.

Meyer, Bruce and Dan Rosenbaum. 2001. "Welfare, the Earned Income Tax Credit, and the Labor Supply of Single Mothers." *Quarterly Journal of Economics*, 116(3): 1063-1114.

Michelmore, K. & N. Pilkauskas, 2021. "Tots and Teens: How Does Child's Age Influence Maternal Labor Supply and Child Care Response to the Earned Income Tax Credit?," *Journal of Labor Economics*, University of Chicago Press, vol. 39(4), pages 895-929.

Nichols, A. and Rothstein, J., 2015. The Earned Income Tax Credit. In *Means-Tested Transfer Programs in the United States*, Volume 1 (pp. 137-218). University of Chicago Press.

Strain, M. R. and Whitmore Schanzenbach. 2021. "Employment effects of the Earned Income Tax Credit: Taking the long view." *Tax Policy and the Economy*, 35(1), 87-129.