DISCUSSION PAPER SERIES

IZA DP No. 5491

The Impact of the UK New Deal for Lone Parents on Benefit Receipt

Peter Dolton
Jeffrey Smith

February 2011

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

The Impact of the UK New Deal for Lone Parents on Benefit Receipt

Peter Dolton
Royal Holloway College, University of London,
London School of Economics and IZA

Jeffrey Smith
University of Michigan,
NBER and IZA

Discussion Paper No. 5491
February 2011

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 5491
February 2011

ABSTRACT

The Impact of the UK New Deal for Lone Parents on Benefit Receipt*

This paper evaluates the UK New Deal for Lone Parents (NDLP) program, which aims to return lone parents to work. Using rich administrative data on benefit receipt histories and a "selection on observed variables" identification strategy, we find that the program modestly reduces benefit receipt among participants. Methodologically, we highlight the importance of flexibly conditioning on benefit histories, as well as taking account of complex sample designs when applying matching methods. We find that survey measures of attitudes add information beyond that contained in the benefit histories and that incorporating the insights of the recent literature on dynamic treatment effects matters even when not formally applying the related methods. Finally, we explain why our results differ substantially from those of the official evaluation of NDLP, which found very large impacts on benefit exits.

JEL Classification: I3

Keywords: program evaluation, active labor market policy, matching, lone parents, New Deal

Corresponding author:

Jeffrey Smith
Department of Economics
University of Michigan
238 Lorch Hall
611 Tappan Street
Ann Arbor, MI, 48109-1220
USA
E-mail: econjeff@umich.edu

* We thank João Pedro Azevedo for research assistance. We are grateful to seminar participants at the University of Arizona, Duke University, Institute for Fiscal Studies, Universität Freiburg, University of Maastricht, University of California-Riverside, Federal Reserve Bank of New York, IUPUI, University of Melbourne, Australian National University and Université Laval as well as audiences at several conferences for helpful comments. We are particularly grateful to Mike Daly, Robert Moffitt, Jeff Grogger, Joe Hotz, Genevieve Knight and Stefan Speckesser. The views expressed herein are our own and do not necessarily reflect the views of the UK Department for Work and Pensions. The usual disclaimer regarding responsibility for errors applies.

1. Introduction

In this paper, we evaluate the New Deal for Lone Parents (NDLP), a large voluntary program for single parents in the United Kingdom (UK). This program, part of a family of welfare-to-work programs introduced by Britain's "New Labour" government in the late 1990s, provides information, referrals and limited financial support to "encourage lone parents to improve their prospects and living standards by taking up and increasing paid work, and to improve their job readiness to increase their employment opportunities" (UK Department for Work and Pensions 2002). Its features resemble those of earlier voluntary programs targeted at a similar population in the United States (US), as well as the less intensive aspects of general employment and training programs such as the US Workforce Investment Act or the Canadian federal-provincial Labor Market Development Agreements. As such, both our methodological and our substantive findings have relevance inside and outside the UK.

Our evaluation applies semi-parametric matching methods to a large administrative dataset rich in lagged outcome measures. Our decision to rely on matching methods has a fourfold motivation: first, the literature clearly indicates the importance of conditioning on lagged outcome variables for reducing (and, hopefully, eliminating) selection bias; we have exceptionally detailed data on these variables. Second, using a subset of our data for which we have detailed survey information, we can examine the value of conditioning on additional variables not present in the administrative data, including a variety of attitudinal measures. Third, relative to conventional analysis that also assumes selection on observed variables but estimates a parametric linear model with main effects in the conditioning variables, matching does not impose linearity on the conditional mean function and allows examination of the extent of common support (i.e. overlap). Fourth, we lack access to plausible exclusion restrictions due to the design and implementation of the NDLP program. While the presence or absence of an instrument does not affect the plausibility of our "selection on observed variables" assumption, it does reduce the choice set of available evaluation strategies.

We examine the impact of NDLP participation on individuals eligible for NDLP in August 2000 who began a spell of NDLP participation between August 1, 2000 and April 28, 2001 using weekly benefit receipt as an outcome measure. Our empirical analysis yields a number of important substantive and methodological findings. On the substantive side, we estimate large (by the standards of experimental evaluations of similar programs in the US) and fairly persistent effects of NDLP participation on the probability of benefit receipt. For NDLP participants in the midst of long spells (at least 66 weeks) of receipt of Income Support (IS), a group we call "the stock", we estimate a reduction in the probability of being on

IS of 17.48 percentage points. In contrast, we estimate that NDLP participants in the midst of relatively short spells of IS receipt, whom we call "the flow", experience a reduction in the probability of being on IS of 5.21 percentage points. The difference between the stock and flow estimates suggests a huge one time benefit from encouraging long-term IS recipients to look for work at a time when other program changes made it more financially attractive for them to do so. The difference also likely reflects the fact that our data allow us to do a better job of controlling for selection in the flow than in the stock.

Though surprisingly large, our estimates are much smaller than those of the official impact evaluation commissioned by the UK Department for Work and Pensions (DWP), conducted by the National Centre for Social Research (NCSR) and reported in Lessof et al. (2003). We explore the sources of these differences.

Methodologically, our analyses support the general conclusion in the literature regarding the importance of pre-program outcome measures in reducing selection bias in non-experimental studies. Moreover, we show, building on Card and Sullivan (1988), Heckman et al. (1998a) and Heckman and Smith (1999), the importance of not just conditioning on lagged outcomes but of doing so *flexibly*. Conditioning on simple summary measures of time on benefit prior to August 2000 yields different, and larger, impact estimates than our preferred measures that embody the rich heterogeneity in IS participation histories present in the data. Our results suggest that the literature has devoted insufficient attention to the importance of flexibility when conditioning on past outcomes.

Using survey data from the official evaluation for a subset of our sample, we show that, once we condition flexibly on lagged outcomes, further conditioning on a variety of measures of attitudes towards work has a large effect on the impact estimates. This indicates that the lagged outcomes we employ do not fully embody these otherwise unobserved factors. These findings suggest the value of further exploring the importance of such variables in other contexts and cast some doubt on the now popular strategy of relying on administrative data alone to evaluate active labor market programs. In a parallel analysis, we find that matched exogenous local area economic variables from the Labour Force Survey do not change the estimates once we flexibly account for the history of IS receipt. This raises questions about the generalizability of Heckman et al.'s (1998a) finding on the importance of conditioning finely on local labor markets.

Our final methodological finding concerns the use of propensity score matching in stratified samples. We find that taking account of the stratification by applying propensity score matching within strata, as suggested by Dolton et al. (2006), rather than ignoring the problem (as in the rest of the literature) makes a difference to our estimates.

The remainder of our paper is organized as follows. Section 2 describes the NDLP program and policy context and Section 3 describes our data. Section 4 outlines our econometric framework. Section 5 presents our main results using the full sample while Section 6 presents analyses for subgroups as well as some secondary analyses. Section 7 compares our estimates to those in the literature. Section 8 concludes.

2. The NDLP Program and Policy Context

2.1 Program basics

The New Deal for Lone Parents is a voluntary program that aims to help lone parents get jobs or increase their hours of work, either directly or by increasing their employability. In its early stages (including the period covered by our data) the NDLP offered participants advice and assistance (in applying for jobs and training courses) and support (in claiming benefits) from a Personal Advisor (PA). The PA also conducted an in-work benefit calculation with the participant, to highlight the potential financial benefits of returning to work or working more. NDLP personal advisors can also approve financial assistance to help with travel costs to attend job interviews, childcare costs or fees for training courses recommended by the PA. Other than these small amounts, NDLP does not provide participants with additional benefits beyond those for which they already qualified.

In the context of this evaluation, the NDLP "treatment" has three important characteristics. The first is heterogeneity resulting from variation among caseworkers in terms of service recommendations and generosity with subsidies, as well as geographic and temporal variation in the extent of available childcare providers and training opportunities. This heterogeneity suggests the potential importance of subgroup differences in mean impacts.[1]

The voluntary nature of NDLP represents its second important characteristic. Simple economic reasoning suggests that voluntary programs will have larger mean impacts than mandatory ones , due to non-random selection into voluntary programs based on expected impacts. This matters in comparing mean impact estimates from NDLP to those from mandatory welfare-to-work programs.

The relatively low intensity and expense of the services offered constitutes the third important characteristic. Over the period of our data, in round figures, there were approximately 100,000 participants

[1] As documented in Dolton et al. (2006) this heterogeneity in the treatment, combined with variability in labor market outcomes in response to treatment, yields widely varying durations of participation in NDLP. In particular, our participants exhibit a highly skewed distribution of durations with a mean of about 39 weeks and long right tail stretching out over 100 weeks. As they discuss in detail, important issues of measurement error and interpretation arise when considering these durations; for this reason, we do not attempt any sort of dose-response analysis in this study.

and the total program costs were around £40.9 million (at 2000 prices) giving a per unit cost of around £400 per participant. This level of expenditures suggests relatively modest mean impacts; while the literature contains a number of examples of expensive programs with small mean impacts, it contains few examples of inexpensive programs with large mean impacts. See Heckman et al. (1999) for a review of the literature on evaluating active labor market policies and Card et al. (2010) for a meta-analysis of recent evaluations.

2.2 Policy environment

In the period we study, lone parents in the UK received means-tested income support (IS) payments that depended on how many school-age children they had and on the amount of other income they received as well as the standard child benefit received by all UK parents. Lone parents could also receive means-tested housing benefits, either in the form of subsidized council housing operated by local governments or assistance with rent in the private housing market, as well as assistance with their local council taxes. If they worked, lone parents received an earnings subsidy via the Working Families Tax Credit (WFTC), a program similar in nature to the Earned Income Tax Credit (EITC) in the US. Their precise financial circumstances depended most crucially on their income from paid work and their housing costs. Access to childcare and its price varied (and still do vary) enormously by geographical location across the UK, especially for children under age four, as did access to state funded nursery school and kindergarten. Specifically the scene changed in 1999 with the first round of the Sure Start national policy which provided integrated learning and childcare for most disadvantaged areas as well as parenting guidance and antenatal and postnatal care. Gregg and Harkness (2003), Gregg, et al. (2009) and Suhrcke et al (2009) provide further information regarding these policies and programs and the wider impact of the NDLP program.

Prior to the advent of NDLP only limited pressure was put on lone parents to work in the UK. IS recipients had to participate in semi-annual "Restart" interviews – see e.g. Dolton and O'Neill (2002) for details and evaluation results – but, particularly in comparison with the long history of welfare-to-work programs in the United States, social and programmatic expectations, as well as financial incentives, helped keep lone parents in the UK at home.

Perhaps not surprisingly, this policy environment led lone mothers to have much lower employment rates than married mothers. Figure 2 of Gregg and Harkness (2003), drawn from OECD (2001), shows an "employment gap" of 24 percentage points in the UK in 1999. In contrast, in most other OECD countries single mothers were more likely to work than married mothers. For example, at the other extreme, in Italy and Spain in 1999 single mothers had employment rates 27 and 23 percentage points *higher* than married

mothers, respectively. This large difference provided part of the motivation for the introduction of the NDLP.

As described in, e.g., Gregg and Harkness (2003), around the same time as the nationwide introduction of NDLP in 1998 four other important labour market changes occurred. First, the Working Family Tax Credit (WFTC) replaced the pre-existing Family Credit (FC). This resulted, in general, in more generous support for working lone parents both directly in terms of larger credits and indirectly via the handling of childcare expenses. Second, a National Minimum Wage was introduced in 1999. Dolton et al. (2010) find that the minimum wage had little effect on employment but a significant positive impact in reducing wage inequality. Third, the UK reorganized its system of employment and training programs in the form of the Job Centre Plus system. This system includes case management, "one stop" centers, performance standards and all the rest of the currently popular design features for these schemes. Finally , in the period after our data, lone parents became subject to mandatory Work Focused Interviews (WFIs) both at the start of their IS spells and at regular intervals thereafter. For more on WFIs and their interaction with NDLP see Coleman, et al. (2003) and Knight et al. (2006).

The policy environment as described here has three main implications for our study. First, the relative lack of programs to push lone parents on IS into work prior to NDLP suggests that many among the stock of NDLP participants in place at the time of NDLP introduction may have needed only a gentle nudge to move them into work. Second, the program changes helped to make work more attractive relative to IS receipt; when the PA calculated the costs and benefits of work, work may have appeared a more attractive option. Third, the new Job Centre Plus system has a stronger focus on employment than earlier UK schemes; part of the estimated mean impact of NDLP likely results from referrals to this improved system.

2.3 Evolution of NDLP over time

An understanding of the development of NDLP over time aids in generalizing the results from this study to more recent cohorts of NDLP participants. In Phase One, a prototype was launched in July and August 1997 in eight locations; see Hales et al. (2000) for an evaluation. In April 1998, Phase Two introduced the program nationally for new and repeat claimants. In Phase Three, NDLP became available to the entire stock of lone parents in September 1999. Our study focuses on the Phase Three period.

NDLP has greatly expanded its target population over time. Initially, NDLP was rolled out to lone parents making new claims for IS whose youngest child was over five years and three months of age. By September 1999 the roll out included those lone parents with a youngest child over five years and three

months of age who had initiated an IS claim prior to April 1998 (i.e. the stock of existing claimants). In April 2000, the target group was extended to include lone parents with youngest children between the ages three and five years three months. Subsequently, the distinction between the target and non-target group has largely disappeared. In November 2001 (not long after our participants joined the NDLP program), all lone parents not in work, or working fewer than 16 hours a week, including those not receiving benefits, became eligible for NDLP.

The NLDP administrative database shows that 577,720 spells of NDLP participation started between October 1998 and December 2003 (which includes a small number of repeat spells). The number of current participants has increased over the life of the program, with noticeable increases in September 1999 when the stock became eligible and again in response to the widening of eligibility in November 2001. By the end of 2003, participation had reached about 100,000 lone parents. These figures demonstrate the importance of NDLP for lone parents on benefit and suggest that it may have equilibrium implications, an issue we return to later.

3. Sample Design, Sampling Issues and Data

3.1 The sample design

Our analysis employs a stratified, geographically clustered random sample of 64,973 lone parents on IS and eligible for NDLP as of August 2000 sampled in two waves denoted "Wave 1" and "Wave 2" combined with a "booster" sample of eligible new lone parent IS cases drawn from the same areas in October 2000. The sampling scheme excluded a number of geographic areas involved in pilots of NDLP or other programs at the same time. The sampling process also excluded a small number of individuals who had participated in NDLP prior to the sampling. The stratification depends on the age of the youngest child and the length of the parent's spell of IS receipt as of the sampling date.[2] Administrative data on IS recipients define the population.

Table 1 shows the composition of the sample relative to the population in the selected Primary Sampling Units (PSUs), following exclusion of lone parents who had already participated in NDLP. Each row corresponds to one of the 24 strata defined by the age of the youngest child and the duration of the IS spell in progress at the time of sampling. Columns 4 and 5 give the size of the population for the strata in

[2] This sample also forms the starting point for the much smaller sample employed in the Lessof et al. (2003) impact report; see Section 7.1. See Dolton et al. (2006) for more details about the definitions of the Primary Sampling Units (PSUs), the exclusion of certain PSUs, and other sampling issues.

August 2000 (labeled "Wave 1/2") and in October 2000, (the "booster sample") where the October population of interest consists only of lone parents with IS spells of less than three months duration. Column 6 gives the sum of columns 4 and 5. The next three columns indicate the number of NDLP participants in our sample from Waves 1 and 2 and from the booster sample, and the total of these. The next four columns indicate the overall number of sample members in each stratum from the August 2000 sample and the booster sample, the sum of these, and the ratio of the sample to the population. The final column makes it clear that stratification represents an important issue in our data, as the sampling rates range from a low of 0.19 to a high of 0.99 among the strata, where the highest sampling rate relates to those eligible with spells of between 3 and 6 months duration.

Spells in progress at a point in time over-represent long spells relative to their representation in the population of all spells. The literature calls this "length bias". We have a length biased population and, as a result, a length-biased sample. Adding IS spells of less than three months in progress in October 2000 to our population does not convert our population into the population of all spells, rather it undoes the length bias in a crude way and to an unknown extent. Rather than attempting elaborate weighting schemes to obtain estimates for a random sample of all spells, schemes which would have to rely on assumptions about inflow onto IS in periods not in our data, we simply define our population of interest as lone parents eligible for NDLP in August 2000 or, for spells of less than three months in duration, in August or October 2000, in the PSUs employed in Lessof et al. (2003). The somewhat unusual population of interest is unfortunate, but the data essentially force it upon us. We attempt to cope with the length-bias issue by presenting separate estimates by length of IS spell in Section 5.2 below. In addition, unless explicitly noted, all of the full sample analyses presented use weights to undo the stratified sampling, so that they correspond to estimates for the population just defined.

3.2 The data

Our dataset combines extracts from a number of administrative datasets maintained by the UK government for the purpose of administering its benefit programs and active labor market policies. Dolton et al. (2006) describes these data sets in some detail. Like most administrative datasets – see, e.g. the discussions in Hotz and Scholz (2002) or Røed and Raaum (2003) – ours had its share of anomalies and problems, including, but not limited to, overlapping spells on mutually exclusive benefit programs for a number of individuals. As described in Dolton et al. (2006), working in consultation with DWP staff, we spent a substantial amount of time and effort on data cleaning in order to produce the data set ultimately used for this paper. Our

analysis file includes complete data on receipt of IS, Incapacity Benefit (IB – disability insurance) and Job Seekers Analysis (JSA – the analogue of unemployment insurance in the US) from June 28, 1999 to the week of August 26, 2004. For spells in progress on June 28, 1999, we know the starting date of the spell except for spells starting prior to September 1, 1990. We have no information on spells that both start and end prior to June 28, 1999.

3.3 Defining the NDLP treatment

We define participation (or treatment – we use the two terms synonymously) as having an initial NDLP interview during the participation window from August 1, 2000 to April 28, 2001. This is the same definition employed in Lessof et al. (2003a). Our definition of participation differs from the official definition of the NDLP caseload, and from some of the other evaluation studies, such as Evans et al. (2002; p. 29), which employ a more stringent definition that requires involvement in NDLP beyond an initial interview. Similarly, we define as non-participants all lone parents in the sample who do not participate in an initial interview during the participation window described above. Thus, we define participation fairly broadly, so as not to miss any possible impacts of NDLP and, as a consequence, define non-participation relatively narrowly.

Defining participation as we do implicitly puts aside the issues raised in the recent literature on dynamic treatment effects – see e.g. Ichimura and Todd (1999), Abbring and van den Berg (2004), Sianesi (2004), and Heckman and Navarro (2007) and Fredriksson and Johansson (2008). That literature addresses the fact that, contrary to the simple model of a program available in just one period that underlies, e.g., Heckman and Robb (1985) and Heckman et al. (1999), individuals in contexts such as that of the NDLP in fact have a dynamic choice to make. In the period covered by our data, they can participate at any time during their spell of benefit receipt, or not at all. By defining participation in terms of a wide but finite window of time, we ignore both variation in the timing of participation within the participation window as well as future participation by our non-participants after the window and repeat participation by both groups. We discuss the implications of failing to address the dynamic issue for our estimation method and for the interpretation of our results later in the paper.

Dolton et al. (2006) examine the fraction of non-participants (as defined above) during the participation window participating in NDLP following the close of the window. They find a participation rate that starts at zero, climbs to about three percent, and then appears to stabilize. Of our non-participants, about 12 percent participate in NDLP at some point over the period from the close of the participation

window to the end of our data. Turning to repeat participation, about 25 percent of the lone parents we define as NDLP participants have multiple spells of NDLP participation during the period covered by our data. Differences in the incidence of these later spells between participants and non-participants as we define them constitute part of the causal effect of the initial participation. See Dolton et al. (2006) for more about these issues.

3.4 Defining the outcome measure

Our outcome measure of interest consists of benefit receipt. This outcome measure has two important features. First, we care about it for policy reasons; NDLP aims to move lone parents from benefit receipt to work. Second, we can construct it from our data, which do not include information on employment or earnings. As we define it here, benefit receipt means receiving any one of IS, JSA, or IB. By using a broad benefit receipt measure, we come closer to one minus an employment indicator; but we do not get all the way there because some individuals go off benefit without obtaining work (for example, as when they marry someone whose income makes them ineligible).

Looking at benefit receipt rates over time rather than at variables related to exit from the current spell of IS receipt has several advantages. First, our approach takes into account the fact that some NDLP participants may leave IS for a time and then return to IS if they lose their job or find that they cannot effectively combine it with their family responsibilities. In contrast, outcome measures that look at lengths of spells of IS receipt in progress at the time of NDLP participation or of sampling explicitly ignore possible future spells, as do the life tables in the Lessof et al. (2003) report. Outcome measures such as whether an individual ever left IS within a particular time frame also ignore the potential for return to IS. In addition, both types of measures miss any treatment effect that NDLP might have on the duration of future spells of employment or non-employment as in Ham and LaLonde (1996) and Eberwein, et al.) (1997).

Outcome measures that focus only on behavior in the first six months after participation allow too little time for some of those who stop receiving benefits to resume doing so and for individuals who do not participate in NDLP, to find work on their own. As a result, such measures may substantially overstate the impact of NDLP on benefit receipt in the medium and long run.

Our outcome measure consists of benefit receipt measured on a weekly basis; this measure reflects an aggregation of the underlying daily data. As described in Dolton et al. (2006), the variation at the daily level appears less reliable than at the weekly level; moreover, program administration proceeds in terms of weeks rather than days. In all of our analyses, we separately estimate weekly impacts in all weeks for all 24

strata. In reporting overall impact estimates, we take the average of the weekly estimates in what we call the "post-program period", which runs from August 1, 2000 to the week starting August 26, 2004; for individuals participating later in the window, this time interval includes some pre-program weeks as well.

4. Methods

4.1 Framework

We adopt the standard "potential outcomes" evaluation framework. In the usual notation, let Y_1 denote the treated outcome (that realized given participation in NDLP during the participation window) and Y_0 denote the untreated outcome (that realized in the absence of participation in NDLP during the participation window). Let *D* indicate participation, with $D = 1$ for NDLP participants and $D = 0$ for non-participants. We focus on the Average impact of Treatment on the Treated (ATT), given by

$$\Delta_{TT} = E(Y_1 - Y_0 \mid D = 1) = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1),$$

as our parameter of interest. When combined with data on average costs and an estimate of the marginal deadweight cost of taxation, Δ_{TT} allows us to determine whether, from the standpoint of economic efficiency, the NDLP program should be cut or retained. See e.g. Heckman, et al. (1997b) and Djebbari and Smith (2008) for discussions of other parameters of interest in an evaluation context.

Because we include individuals who participate after the participation window within our "untreated" comparison group, the counterfactual we estimate implicitly includes possible future participation in NDLP. This affects the interpretation of our impact estimates and complicates their use in a cost-benefit analysis. In particular, it means that our parameter combines, in a loose sense, impacts from participating versus not with, for some individuals, impacts from participating now rather than later.

Finally, we conduct a partial equilibrium evaluation in this paper. Put differently, we assume the absence of any effects of NDLP participation on non-participants. The statistics literature calls this the "Stable Unit Treatment Value Assumption" or SUTVA for short. As noted in Section 2.3, the NDLP program has a large enough footprint on the labor market that we might expect equilibrium effects. In particular, we might expect displacement of non-participants by participants; this would cause the non-participants in our evaluation to experience worse labor market outcomes (in particular, less work and more time on benefit) than in the absence of NDLP. This, in turn, means that our analysis would overstate the impact of the program. Of course, there could be positive spillovers that lead to a bias in the other direction, as when participants pass along information they learn in the course of participating to non-

participants, or when participants set an example of employment and activity that inspires non-participants. Though potentially important, these effects lie beyond the scope of this paper; we refer the interested reader to discussions in, e.g., Davidson and Woodbury (1993), Heckman et al. (1998b) and Lise et al. (2004).

4.2 Identification using the CIA

We adopt what Heckman and Robb (1985) call a "selection on observables" identification strategy to identify Δ_{TT} (which we will call "selection on observed variables", to emphasize the role of choosing what to observe in designing evaluations). This requires that we adopt what the economics literature calls the Conditional Independence Assumption (CIA) and the statistics literature (rather awkwardly) calls "unconfoundedness". In terms of our notation, we assume that

$$Y_0 \perp D \mid X ,$$

where "$\perp$" denotes independence and *X* denotes a set of observed covariates. In words, we assume independence between the untreated outcome and participation in NDLP, conditional on a set of observed covariates. Following Heckman et al. (1998a), we do not assume the conditional independence of the treated outcome and participation as we do not need it for the treatment on the treated parameter. As discussed in Heckman and Navarro (2004), we therefore allow for certain forms of selection into the program based on impacts.

Substantively, this means that we assume that we observe all the variables, or proxies for all of the variables, that affect both (not either, but both) participation and outcomes in the absence of participation. Conditioning on these variables then removes all systematic differences between the outcomes of participants and non-participants other than the effects of participation. From a different angle, we assume that whatever factors determine participation conditional on *X* are independent of Y_0. Thus, conditional on *X*, participation depends on instruments (variables that affect outcomes only via their effect on participation) that we do not observe. These unobserved instruments generate variation in treatment status conditional on the variables we observe.

A long literature suggests the potential for conditioning flexibly on detailed histories of labor market outcomes to remove selection bias in the context of evaluating active labor market programs; see e.g. Card and Sullivan (1988), Heckman and Smith (1999), Heckman, et al. (1998a), Hotz et al. (2005), Mueser et al. (2007) and Heinrich et al. (2009). In addition, the Monte Carlo analysis in Section 8.3 of Heckman et al. (1999) shows that conditioning on lagged outcomes substantially reduces bias for a wide variety of individual outcome and participation processes. In terms of what determines participation conditional on

observed variables in our context, we expect that it has to do with random differences in information costs and other costs of participation that we do not observe, such as distance to the program office. Finally, because we align our lagged outcome measures relative to the start of the participation window (rather than the actual start of participation), they should do a better job of eliminating selection bias for lone parents starting their spells of NDLP participation early in the window, a prediction we test in Section 6.3.

4.3 Matching algorithm

We apply both cell matching (sometimes called exact matching) and propensity score matching, as developed in Rosenbaum and Rubin (1983). They show that if the conditional independence assumption holds for X, it also holds for $P(X) = \Pr(D = 1 | X)$, the probability of participation given X, also called the propensity score. Matching on the propensity score, a scalar bounded between zero and one, avoids the "curse of dimensionality" inherent in exact matching on multidimensional X.[3]

Propensity score matching constructs an estimated, expected counterfactual for each treated observation by taking predicted values from a non-parametric regression of the outcome variable on $P(X)$ estimated using the untreated observations. Thus, any non-parametric regression method defines a propensity score matching method. In our analysis, we use single nearest neighbor matching with replacement as implemented in the "psmatch2.ado" program for Stata by Leuven and Sianesi (2003). In this method, the estimated expected counterfactual for each treated unit consists of the untreated unit with the nearest propensity score in absolute value. See, e.g. Smith and Todd (2005a), Caliendo and Kopeinig (2008), Busso et al. (2009a, 2009b) and Huber et al. (2010) for additional discussion of matching and more technical detail about alternative matching estimators.

Single nearest neighbor matching throws out a lot of potentially useful information by not making use of multiple untreated observations near a given treated observation when the data provide them. The Monte Carlo analyses by Frölich (2004) and Busso et al. (2009a, 2009b) demonstrate a non-trivial mean squared error cost from choosing single nearest neighbor matching rather than alternative methods, such as kernel matching, that do use multiple untreated observations. We take a pass on those other methods here due to their substantially longer processing time. Constructing weekly impact estimates by stratum, as we do in many of our analyses, became infeasible (with the technology available when we performed our empirical analyses) unless we relied on single nearest neighbor matching.

[3] More accurately, the use of the propensity score pushes the curse of dimensionality back to the estimation of the score, where it is overcome by using a (flexible) parametric model.

4.4 Matching with stratified samples

Dolton et al. (2006, pg. 80-83) provide a simple analysis of the application of matching estimators to stratified samples. They show the desirability of exact matching on the variables defining the strata, particularly (but not exclusively) in contexts where the mean effect of treatment varies in the subgroups defined by the stratification variables. We adopt this approach in this paper and construct our estimates separately for each subgroup defined by the stratification variables (the length of the spell of IS receipt in progress and the age of the youngest child at the start of the participation window) unless otherwise noted.

4.5 Implementation details

We have examined the common support or "overlap" condition at a number of points in the development of our analysis and consistently found that, given our large sample size, it represents only a minor issue. As such, we do not formally impose the common support condition here; see Smith and Todd (2005b) and Lee (2009) for discussions of tests of the common support condition and Crump et al. (2009) and Khan and Tamer (2010) for conceptual and technical background.

We have performed standard balancing tests on all of our conditioning variables in the context of generating estimates using the full sample of administrative data and ignoring the stratification, and we have examined the balance of the lagged outcome variables, which we view as the key covariates, for the estimates reported here in which we do the matching separately by stratum. Indeed, finding imbalance in benefit receipt prior to the start of the participation window when using the specification in the Lessof et al. (2003) report started us down the road toward the more flexible conditioning used here; see the discussion in Section 4 of Dolton et al. (2006) for more details. As described below we consistently find our preferred specification does a good job of balancing the benefit history variables; the sole exception concerns outcomes prior to the start of our complete data for the stock.

We estimate our standard errors using bootstrapping methods with 200 replications. Our bootstrapping operates conditional on the primary sampling units included in the data. As such, we omit any variance component operating at the PSU level. If we interpret our estimates as Sample Average Treatment Effects (SATE) in the spirit of Imbens (2004), then this problem goes away. A more vexing problem arises from the analysis in Abadie and Imbens (2008), who show the inconsistency of the bootstrap for nearest neighbor matching. The Monte Carlo analysis in Abadie and Imbens (2006), unfortunately omitted in the published version, suggests that while not zero, the inconsistency in the bootstrap may not lead to severely misleading

inferences. We leave the pursuit of the alternative variance estimators in Abadie and Imbens (2011) and elsewhere to future work, and in the meantime interpret our standard errors with caution.

5. Impact Estimates

Table A-1 provides descriptive statistics broken down by participation status, and reveals a surprising degree of similarity in the mean observed characteristics of the two groups. Figure 1 presents the unadjusted fraction on benefit for NDLP participants and non-participants in our data. It illustrates that, without any adjustments and despite their relatively similar characteristics, participants have much lower rates of benefit receipt both before and after the start of the participation window. The difference in the period prior to the start of the participation window strongly suggests that participants differ from non-participants in ways related to benefit receipt other than just NDLP participation. Our analysis seeks to eliminate these differences.

5.1 Exact matching on benefit histories

We begin in the spirit of Card and Sullivan (1988) and Heckman and Smith (1999) by performing exact matching based solely on strings that capture much of the detail in individual histories of benefit receipt. This analysis has three primary motivations. First, Dolton et al. (2006) show that the propensity score specification employed in Lessof et al. (2003) fails to balance the fractions receiving benefits among participants and matched non-participants in the Lessof et al. (2003) sample. This indicates that balancing the two groups requires conditioning more flexibly on the benefit history rather than just including the total number of days on benefit.[4] Second, as suggested above, lagged outcomes correlate strongly both with other observed determinants of participation and outcomes and with otherwise unobserved determinants such as tastes for leisure, particular family obligations such as seriously ill or disabled parents or children and so on. Thus, in our view, conditioning on these histories goes a long way toward solving the selection problem. Third, this strategy plays to the strength of the administrative data that we employ. Our data are comprehensive with respect to information about past histories of benefit receipt, but lack depth in terms of other variables, with the exception of basic variables required for program administration such as the number and age of children, the age of the lone parent, and the geographic location of the family. In particular, our administrative data contain no information on schooling or other qualifications.

[4] See Appendix C of Phillips et al. (2003) for the details of the National Centre propensity score model.

To code up our benefit history strings, we first break the period from June 1999 to September 2000 (the period over which we have complete data on benefit receipt) into six 11 week "quarters", where we omit the final week just prior to the start of the participation window. We code an indicator variable for each quarter for whether or not the individual spent at least half the period on benefit. We then concatenate the six indicators to form a string. There are $2^6 = 64$ possible strings, ranging (in binary) from 000000 to 111111. A string of 111111 indicates someone who spent at least half of all six quarters on benefit; similarly, a string of 000000 indicates someone who spent less than half of all six quarters on benefit.

The literature suggests two standard alternatives to the strings we employ here: a variable measuring the fraction of time on benefit in the pre-program period and a variable measuring the duration of the ISspell in progress at the start of the NDLP participation window. Our method has important advantages relative to both. First, relative to a measure of the fraction of time on benefit, the benefit history strings capture the timing of benefit receipt. Using the benefit strings, someone with a 33 week spell at the start of the pre-program period gets coded as 111000, while the same spell at the end of the period gets coded as 000111; a variable measuring time on benefit would give the same value to both. Second, relative to using the duration of the spell in progress at the start of the participation window, the benefit history strings have the advantage of capturing additional spells, if any, during the pre-program period. In addition, our string approach also measures the duration of the spell in quarters and indirectly measures the fraction of time out of work in the sense of the proportion of 1's and 0's.

Two important decisions arise in implementing the benefit history strings. The first concerns how finely to partition the pre-program period. Each additional sub-period doubles the number of possible strings; this in turn consumes degrees of freedom and raises the possibility of common support problems due to strings with participants but no non-participants. On the other hand adding additional sub-periods increases the plausibility of the CIA.

The second, not unrelated, decision concerns the choice of the fraction of time within a period that an individual must be on benefit in order to code them as a one for that period. Setting this value high means that short spells do not count; for example, if we set the cutoff value at 10 of the 11 weeks, then someone with six nine week spells on benefit, one in each 11 week quarter, would be coded as 000000, the same as someone who was never on benefit at all. Setting this value low means that short spells count the same as continuous participation; for example, if we set the cutoff value at being on benefit just one out of the 11 weeks, then someone with six one week spells, one in each 11 week quarter, would be coded 111111, the same as someone continuously on benefit for all 11 months. We chose the 5.5 week cutoff as a compromise,

keeping in mind that few individuals have more than a couple of spells over the entire pre-program period and that the vast majority of spells last at least a couple of months.

Our implementation of the strings has one defect, namely the use of a fixed calendar interval relative to the participation window rather than using time measured relative to the participation decision. As a result of this choice, for some participants the benefit history strings capture their behavior immediately prior to participation, for others they capture behavior a few weeks or months prior to participation. The gain from using fixed calendar dates comes from not having to create phony dates for the non-participants to make their participation decision, as in Lechner (1999) and Lessof et al. (2003). More generally, this strategy flows out of our decision, discussed in Section 4.1 above, not to adopt a dynamic treatment effect framework.

Table 2 presents the results from exact matching on the benefit history strings. The first five columns of the table present the benefit history string for that row, the number of non-participant observations with that string, the average of the weekly probability of benefit receipt over the post-program period among non-participants with that string, the number of participant (treated) observations with that string and the average proportion on benefit in the post-program period among participant observations with that string.

By far the most common string among both participants and non-participants is 111111; the modal benefit history string in both groups represents more or less continuous benefit receipt. A second set of quite common strings, each with several thousand observations in the full sample, consists of strings composed of one or more zeros followed by ones. These almost always represent individuals with a single spell of benefit receipt up to the start of the participation window. A third group of strings with several hundred observations each in the full sample consists of strings with ones followed by zeros followed by ones (in the case of strings ending in zero the new spell of benefit receipt starts in the omitted week before the start of the participation window). These strings represent interrupted spells.

For each string, we construct the string-specific mean impact as the difference in the proportion on benefit in the post-program period between the participants and non-participants in the cell. These differences appear in the column labeled "TT" in each row. We then calculate the weight for each cell; these weights appear in the column labeled "WEIGHT". As we seek to estimate Δ_{TT}, the weight for each string consists of the fraction of the participant observations with that string. We then multiply each string-specific treatment effect by its weight and put the results in the column labeled "CONTR" (for contribution). Summing these yields the overall mean impact estimate for NDLP participation presented in the lower right corner of Table 2.

For the full sample, exact matching on benefit history strings implies that NDLP participation reduces the mean proportion of time spent on benefit in the post-program period by 17.61 percentage points. Though quite large relative to estimates from similar programs in other countries, it nonetheless lies well below the impact estimates reported in Lessof et al. (2003). We put our estimates in the context of the broader literature in Section 7.

A comparison of the impact estimates on the full sample with the corresponding estimates for the sample with the 111111 individuals removed, which we present in the final two columns of Table 2, shows that participants on benefit more or less continuously have a much larger estimated mean impact than other participants.[5] Less formally, the stock has a larger impact than the flow. This difference has two possible sources. It could be that we have simply failed to distinguish strongly enough among the individuals with the 111111 history, leading to more selection bias for this group. Under this interpretation, more weight should be placed on the impact estimate for the other groups, whom we are able to match more finely on their benefit histories. Second, it could be that the NDLP just works better for individuals with very long spells on, or mostly on, benefit.

5.2 Exact matching on sampling stratum

Motivated by the methodological concerns outlined in Section 4.4, we also present estimates based on exact matching only on the sampling strata. As noted in Section 3.1, these strata are defined by the age of the youngest child and the length of the IS spell in progress as of the start of the participation window.

Figure 2 displays the fraction of time on benefit for participants and for non-participants following exact matching on the sampling strata. The underlying matching algorithm corresponds to that in Section 5.1, but with the strata replacing the benefit history strings. Relative to the raw data shown in Figure 1, exact matching by stratum reduces by over half the differences between participants and non-participants in benefit receipt rates prior to the participation window. This figure highlights the potential for ignoring the stratified sampling issue when constructing matching estimates to lead to substantial bias.

5.3 Propensity score matching

In this section we present estimates obtained by propensity score matching using the administrative data. In light of the importance of exact matching on the sampling strata demonstrated in the preceding section, we perform propensity score matching separately within each stratum. That is, within each stratum we estimate

[5] This analysis does not take account of the stratified sampling.

a separate propensity score model (though each one contains the same set of covariates) and we match participants in a given stratum only to non-participants in the same stratum.

The propensity score specification for each stratum includes the sex, age (indicators for 10 five-year categories) and disability status of the lone parent, the number of children in the household, the age of the youngest child, and 12 region dummies (10 for England and one each for Scotland and Wales). In addition, we include three sets of variables related to pre-program benefit histories. First, we include 45 indicator variables, one for each of the non-empty benefit history strings defined in Section 5.1.[6] Second, because over half of the sample has the same string (111111), and because of concerns that we may not have exploited all of the information in the benefit history data for this group, we also add a continuous variable that gives the length of any spell of IS receipt in progress as of June 1999. Recall that our data limits what we can do in this earlier period. Third, in the spirit of Heckman and Smith (1999), we attempt to capture the effects of benefit receipt shortly before the participation decision by including indicator variables for benefit receipt in each of the six weeks prior to the start of the participation window.

Table 3 presents the estimates from the propensity score probit model for the stratum of lone parents with IS spells of less than three months duration and youngest children of age less than three years.[7] The variables related to the age of the lone parent, the age of their youngest child and the number of children all show high levels of statistical and substantive significance. The benefit history variables do not matter much, a finding that makes sense for this stratum given their short histories (and given the limited variation in spell length within the stratum). Figure 3 presents the fraction on benefit in each month from 1997 through 2004 for the treated units and matched, via the propensity score matching, untreated units. Two patterns stand out: First, the propensity score matching does an impressive job of balancing pre-program benefit receipt between the participants and the non-participants.[8] Second, as expected, the propensity score matching yields smaller impact estimates than simply matching on the strata, though the impact estimates remain substantively large.

[6] All strings with fewer than 20 observations were pooled into a single category denoted "222222". This combination includes 19 strings but only 68 observations.
[7] Results for other strata are available from the authors on request.
[8] The impact estimates corresponding to the figure appear in the first row of Table 6.

6. Further Analyses

6.1 Heterogeneous treatment effects: stock and flow

Motivated by our findings in Section 5.1, in this section we present separate propensity score matching estimates for the stock (those with benefit history strings of "111111") and the flow (those with all other benefit history strings). We match exactly on the sample stratum and on whether an individual belongs to the stock or the flow. Within subgroups defined by these exact matches, we estimate the propensity score model defined in Section 5.3 and use the resulting propensity scores to perform single nearest neighbor matching with replacement.

Table 4 presents the estimated mean impacts from this analysis at specific points in time in the post-program period. The first row presents the estimated mean of the weekly impacts on the fraction of time on benefit for the entire post-program period. The following four rows present estimates of the difference in the fraction on benefit between the participants and the matched non-participants at 3, 9, 24 and 36 months after the start of the participation window in August 2000. Three important results emerge from this analysis. First, as in the case of exact matching on the benefit strings in Section 5.1, the mean impact differs quite substantially between the stock and the flow. For example, the ATT for the full sample over the whole post-program period equals 19.09 whereas the impacts for the same period for the stock and flow equal 21.04 and 9.87, respectively. Second, the impacts fall over time for the stock but not the flow. This reduction over time results from catch up by the non-participants rather than from increases in benefit receipt among NDLP participants; thus, NDLP in part speeds up benefit exits that would otherwise occur several months later on their own.

With respect to the fade out of program impacts over time, the wider literature does not provide a clear guide as to what to expect. US General Accounting Office (1996) shows that impacts for the US Job Training Partnership Act (JTPA) remain quite stable over time; Couch (1992) shows the same for the US National Supported Work Demonstration. Dolton and O'Neill (2002) also find a sizeable impact of the Restart program in the UK over four years after random assignment. In contrast, Hotz, Imbens and Klerman (2006) show that the impacts from the work-first part of the California Greater Avenues to Independence (GAIN) program fade out over time. The impacts of the Canadian Self-Sufficiency Project earnings supplement program for single parents also fade out due to control group catch up, as shown in Michalopoulos et al. (2002). Among these programs, the services offered by GAIN most closely resemblance those offered by the NDLP, though it is a mandatory rather than a voluntary program.

6.2 Heterogeneous treatment effects: demographic and benefit history subgroups

In this section we consider heterogeneity in the mean impact of NDLP among subgroups. First, we estimate mean impacts for lone parents with a youngest child in the age intervals [0, 3), [3, 5), [5, 11) and [11, 16] years. We then estimate mean impacts for lone parents in IS benefit spell duration intervals of [0, 3), [3, 6), [6, 12), [12, 24), [24, 36), and 36 or more months at the start of the participation window. The variables that define our univariate subgroups in this section also define, when combined, the sampling strata. The estimates come from exact matching on the sample strata, followed by propensity score matching within sample stratum using the propensity score model in Section 5.3. We then take weighted averages of the estimates from the appropriate strata to obtain the subgroup estimates.

Table 5 summarizes the subgroup impact estimates. In Table 5, each row corresponds to the indicated subgroup. The column labeled "Treatment" presents the impact estimate for the entire post-program period. The column labeled "Pre-window difference" gives the treatment effect for the pre-program period; with complete balance of the lagged outcomes this will equal zero. The third column, labeled "Difference" subtracts the pre-program difference from the post-program impact estimate. It thus represents an alternative impact estimate in the spirit of the symmetric differences estimator in Heckman and Robb (1985). As we do a good job of balancing the pre-program benefit histories for most subgroups, we focus our attention on the estimates in the "Treatment" column.

In terms of the age of the youngest child, we find the smallest point estimates for the youngest children and the largest for children ages 3-5, who have recently reached school age. Lone parents of older children may find it easier to leave home for work when encouraged to do so by NDLP. Figures 4 and 5 show how these impacts play out over time. We find larger but less interpretable differences by the duration of the IS spell in progress at the start of the participation window. As expected given the differences between the stock and the flow observed in Sections 5.1 and 6.1, lone parents on benefit more than 36 months have the largest estimated impacts. Figures 6 and 7 present the impacts for the groups with spells of less than three months, and between 2 and 3 years, duration, respectively, and graphically illustrate the exact matching on the IS spell length.

6.3 Window width

Due to the nature of the underlying study design, our benefit history variables all refer to time relative to the start of the participation window on August 1, 2000. We can think of these variables as measurement-error versions of "latent" benefit history variables measured at the time of the participation decision, where the

amount of measurement error increases as time passes during the participation window. Thinking about the variables in this way leads to the conclusion that they should do better at dealing with the selection problem for those who participate early in the participation window relative to those who participate later in the window. To test this conjecture, in the second panel of Table 6 we present impact estimates that define as NDLP participants only those who participate in the first half of the window.

Our analysis yields a very important finding: we obtain a lower impact estimate for the first half of the window, equal to just 5.21 for the flow. The large change in the estimate relative to that obtained using participants over the full window confirms our hypothesis that increasing measurement error throughout the window degrades the quality of the conditioning in the second half of the window. This estimate represents our preferred estimate for the flow. We conjecture that reducing the size of the window again would reduce the estimated impact still more. For the stock, the estimate also falls in magnitude when we restrict our attention to participants in the first half of the window, but the large pre-program difference suggests that we should place little weight on the estimates.

6.4 Benefit histories

The third panel of Table 6 considers several different ways of including benefit history information in the propensity scores. Because we already conduct the entire underlying analysis separately for strata defined in part by the duration of the spell in progress at the start of the window, even the estimates labeled "no benefit history" implicitly condition on benefit duration as described in Table 1. In a linear regression context, this roughly corresponds to interacting every regressor with indicators for the six duration categories that define the strata; as such, it represents fairly flexible conditioning, particularly for the shorter spells. Thus, the four rows in this panel of Table 6 provide information about the marginal value of further conditioning relative to that implicit in constructing separate estimates by stratum.

In particular, the first row of the panel reports estimates for our preferred specification but omitting the benefit history variables. The second specification includes only the fraction of time the individual spent on benefit from June 1999 to August 2000. The third specification includes only the duration of the spell in progress as of August 2000, the start of the participation window. The final row includes only the benefit strings discussed in Section 5.1.

The estimates reveal three patterns of interest. First, for the flow, the "no benefit history" and "only the duration of the benefit spell" specifications imply substantial (over 2.0) pre-program differences in the fraction on benefit. Second, for the stock, all four specifications (along with our preferred specification)

yield pre-program differences of 1.95 or more, with the "duration of the benefit spell" specification leading the field. Third, of the two specifications with a small pre-program difference for the flow, the "just the benefit strings" specification yields the smaller impact estimate; neither differs by very much from our preferred specification estimate. Overall, the additional benefit history variables add surprisingly little to the conditioning implicit in constructing separate estimates by strata, though we can reject many of them based on unsatisfactorily large pre-program differences.

6.5 Geographic information

The analyses in the final panel in Table 6 vary the geographic information we condition on while retaining the other variables in our preferred specification. In the US context, Heckman et al. 1998a) and Heckman et al. (1997a) found conditioning on local labor markets to play a key role in reducing bias in their non-experimental evaluations of the JTPA program. The findings in Friedlander and Robins (1995) using data from various experimental evaluations of US welfare-to-work programs support this conclusion. Similarly, Hoynes (2000) finds that local economic conditions matter for the timing of exits of single parents from the US Aid to Families with Dependent Children (AFDC) program.

In contrast to the literature, we do not find much effect of geography in the NDLP data. To see this, compare the penultimate row of the table, which omits all geographic information from the propensity score, to our preferred specification in the first row, which includes indicator variables for 11 (of 12) regions within the UK. The differences in the estimates are relatively minor for both the stock and flow. The final row in Table 7 replaces the 12 region indicators with measures of unemployment and of income and employment "deprivation" at the level of the Local Authority District (LAD), a much finer level of geographic disaggregation.[9] This detailed geographic information hardly moves the estimates at all. We remain puzzled by the relative unimportance of geographic conditioning information in the UK We conjecture that this finding has one of two causes: either variation in local labor market conditions and other factors gets picked up by the benefit histories or the UK is somehow more homogeneous than the US in terms of local labor markets.

6.6 Attitudinal variables

In Table 7 we examine the importance of attitudinal variables. To do so, we make use of the postal survey administered as part of the official Lessof et al. (2003) evaluation of NDLP. This necessarily limits us to the

[9] See Section 5.4 of Dolton et al. (2006) for more on these measures.

subset of 42, 249 lone parents who responded to the postal survey. The postal survey included a battery of nine attitudinal statements designed to capture the respondent's attachment to work. Respondents gave their reaction to each statement on a five-point Likert scale ranging from "strongly agree" to "strongly disagree". Examples of the statements include "Having almost any job is better than being unemployed" and "It is just wrong for a woman with children under five to go out to work". The last four rows of Table 7 present estimates that include one or both of the attitudinal variables and the benefit history variables. As in Table 6, because we match within strata defined by the age of the youngest child and the duration of the IA spell, even the specifications without our benefit history variables implicitly condition on a coarse measure of IA spell duration.

Focusing on the flow, we find a large pre-program difference when we do not condition on either the benefit history variables or the attitudinal variables in the first row. Conditioning on the benefit histories in the second row largely eliminates the pre-program difference but does little to change the impact estimate. A comparison of the estimates in this row with those in the first row of Table 6 shows that, given our preferred propensity score specification, restricting our attention to the postal survey sample reduces the overall impact and the impact on the stock by several percentage points, but has little effect on the impact for the flow. This apparent selection on impacts represents one benefit of using administrative data not subject to survey non-response. The estimates in the third row of Table 7 leave out the benefit history variables but include the attitudinal variables. To our surprise, inclusion of the attitudinal variables produces an estimate for the flow of just 5.44. Similar but less dramatic reductions occur for the stock and overall. Row 4 reveals that once we condition on the attitudinal variables, adding in the benefit history variables yields a marginally smaller estimate for the flow, while leading to a larger estimate for the stock.

In the NDLP data, the attitudinal variables matter a lot, and appear to take care of a substantial amount of residual selection on unobserved variables that remains after conditioning on the other variables in our preferred specification. Our finding has three implications. First, adding these variables to the administrative data, perhaps via a short in-office survey administered at the first meeting with an NDLP caseworker, would improve the ability of the administrative data to generate compelling impact estimates. Second, in survey-based evaluations, adding these variables represents a good use of interview time. Third, the suggestion sometimes seen in the literature that attitudinal variables such as the ones we examine here might constitute good instruments for participation is clearly inconsistent with their strong relationship with untreated outcomes.

7. Putting Our Estimates in Context

7.1 Comparison to the National Centre Evaluation

The NCSR evaluation presented in Lessof et al. (2003), though examining the same program with (in part) the same data and (in part) the same sample, proceeded very differently than we do. Their evaluation strategy began with a postal survey sent to everyone in the population described in Section 3.1. This postal survey had a response rate of 64.4 percent which, though high by postal survey standards, raised serious concerns about non-response bias. Later on, the NCSR administered a face-to-face interview survey to all of the NDLP participants who responded to the postal survey whose responses did not come after the start of a spell on NDLP, as well as to a matched (using information from the postal survey) sample of NDLP non-participants. The response rate for the face-to-face interview survey was 70 percent; see Phillips et al. (2003), Table 5.5.1. The matching consisted of single nearest neighbor matching without replacement based on propensity scores that included demographics, as well as relatively crude measures of lagged outcomes from the administrative data and attitudinal variables from the postal survey. Lessof et al. (2003) present impacts based on the respondents to the face-to-face interview survey using as their primary outcome measure whether or not an individual left IS within nine months of the start of the participation window. They estimate a rather startling NDLP impact of 26 percentage points on this outcome.

In Dolton et al. (2006), we examine the NCSR evaluation in great detail, and look in particular at various features of the design and implementation that might have biased their estimates. Although we do not attempt a precise decomposition of the difference between their estimates and our own, we find that survey non-response, at least as a function of observed characteristics, does not seem to affect the estimates much. The same holds true for the details of the matching method used, which comports with the general finding in the literature; see e.g. Smith and Todd (2005a), Mueser et al. (2007), or Plesca and Smith (2007).

In contrast, three factors do matter in explaining the difference in estimates. First, flexible conditioning on detailed benefit receipt histories leads to lower impact estimates. Second, using all lone parents who participate in NDLP during the participation window, rather than just those who participate after returning their postal survey, also lowers the impact estimates. Third, using benefit receipt at each point in time, rather than just time to the first exit from IS, modestly decreases the estimates by taking account both of recidivism onto IS and differential movements to other types of benefits among NDLP participants. Given the problems with the Lessof et al. (2003) analysis just described (along with its failure to account for the complex sampling scheme as discussed earlier), and detailed at length in Dolton et al. (2006), we strongly prefer the estimates presented in this paper.

7.2 Comparison to experimental evaluations of US programs

In this section, we compare our estimates to experimental estimates of employment impacts for similar programs serving similar populations in the United States. We focus on the flow estimates in this discussion because they have greater policy relevance (you can only treat the stock once), because we have greater confidence in them due to our lack of detailed information on benefit receipt before 1999 for the stock, and because they correspond better to the population served by US programs. In particular, we highlight the two estimates that we find most compelling among all those we present. The first estimate, 5.21, arises from our preferred propensity score specification for participants in the first half of the participation window. The second, 4.90, comes from our preferred specification augmented with the attitudinal variables and applied to the full sample of postal survey respondents. It appears in the last line of Table 7. The first estimate has the advantages, described above, of looking only at participants in the first half of the window, and also avoids any issues of non-random non-response to the postal survey. The second estimate has the virtue of conditioning on the (obviously quite important) attitudinal variables.

We look at impact estimates from the US because only the US has accumulated a non-trivial body of experimental impact estimates for similar programs administered to similar populations. We focus on employment impacts as these correspond most closely to our impacts on benefit receipt, keeping in mind the fact that some individuals in the NDLP context may leave benefit but not enter employment. We consider two sets of programs: voluntary programs aimed at disadvantaged women and mandatory welfare-to-work programs for lone mothers on benefit. In the US context, "on benefit" means in receipt of AFDC or its successor Temporary Aid to Needy Families (TANF).

We begin by considering two expensive, intensive voluntary programs for populations including but not limited to lone mothers: the National Supported Work Demonstration (NSWD) and the Job Corps. If we assume that more inputs, in the form of program expense, should generate larger program impacts, then these programs provide a (perhaps distant) upper bound on what we might expect from the much less intensive treatment provided by NDLP. The NSWD provided its participants with intensive work experience in a supportive environment for several months. Table 4.6 of Hollister and Maynard (1984) shows an impact on employment in months 25-27 after random assignment of 7.1 percentage points. The Job Corps provides intensive training in job and life skills over several months in a residential setting to disadvantaged young adults. Figure VI.8 of Schochet et al. (2001) shows impacts on the fraction of weeks employed in the fourth year after random assignment of 0.041 for female participants.

In our view, the "other services" treatment stream for adult women from the US National Job Training Partnership Act (JTPA) may represent the best overall analog to the NDLP among the programs considered here. The population served by JTPA included all disadvantaged women, not just lone mothers on benefits, but, as shown in Exhibit 4.10 of Bloom et al. (1993), women with some AFDC experience represent 46.1 percent of the experimental sample for this stream. The "other services" treatment stream, defined based on treatments recommended prior to random assignment, includes mainly job search assistance and other low intensity services. Exhibit 4.13 of Bloom et al. (1993) presents experimental impact estimates on employment rates over the first six quarters after random assignment measured using detailed survey data on employment spells.[10] The table shows an impact on the probability of employment of 0.045 in the fifth quarter after random assignment and 0.023 in the sixth quarter, with an average of 0.043 over all quarters.

Gueron and Pauly (1991), LaLonde (1995), Friedlander and Burtless (1995) and Hamilton et al. (2001) (and many others) summarize the results from a number of experimental evaluations of mandatory welfare-to-work programs for AFDC recipients. Table 5-1 in Friedlander and Burtless (1995) presents five year employment impacts from experimental evaluations of four mandatory welfare-to-work programs from the 1980s. These evaluations measure employment as the number of quarters with non-zero earnings in administrative earnings data from state Unemployment Insurance (UI) systems. Under the assumption that treatment and control group members work the same fraction of each quarter, we can translate these estimates into impacts on employment probabilities. For these four programs, this yields impacts of 0.029 (= 0.55 / 20), 0.046 (= 0.97 / 21), 0.029 (= 0.41 / 14) and 0.049 (= 0.97 / 20). Table 4.1 of Hamilton et al. (2001) presents similar evidence for several welfare-to-work programs implemented in the late 1990s. The impacts they report range from -0.1 to 1.6 quarters. Almost all of them lie in the range from 0.3 to 0.8 quarters. Under the same assumption as before, with a denominator of 20 quarters, 0.8 quarters corresponds to an increase in employment probability of about 0.04.

Like Hämäläinen et al. (2008), we expect mandatory programs to have lower mean impacts than voluntary ones due to self-selection on impacts. Thus, the estimates just described represent a lower bound on what to expect from the NDLP, with the caveat that some of these programs provide modestly more intensive services than the NDLP (though much less intensive than the NSWD or the Job Corps). As such, we cannot infer too much from the fact that their impacts fall well below our estimates for the NDLP.

[10] Unfortunately, the 30 month impact report in Orr et al. (1996) lacks a similar table and focuses almost entirely on earnings rather than employment impacts.

Taken together, the evidence from the experimental literature in the US suggest that the two estimates we highlighted as most compelling for the flow look reasonable. They modestly exceed the estimates for the mandatory welfare-to-work programs, as we would expect if individuals who choose to participate in NDLP can at least somewhat anticipate their gains from participation. At the same time, they end up below the experimental estimates for the more intensive US programs.

In the end, we remain concerned that the NDLP impact estimates for the stock seem too large. This could have several explanations: first, despite our flexible conditioning on the benefit histories some selection on unobserved variables likely remains; our estimates suggest positive selection on unmeasured ability or motivation. Second, many benefit leavers may not go into employment. The broader results from the US evaluations lead us to doubt the importance of this explanation. Third, the NDLP may have negative spillover effects, perhaps due to displacement, on our comparison group; we view this as plausible but unlikely to bias the estimates by more than a percentage point or two. Fourth, the UK may do a better job of running these programs or they may work better in the UK economic environment. Both of these explanations seem unlikely to account for much either, given the strong economy in the US in the time period corresponding to most of the evaluations discussed above and given the greater US experience with programs of this type. Finally, we note that only the first of these explanations applies more to the stock than to the flow. Given the reasonableness of our flow estimates, we lean toward the first explanation.

8. Conclusions and Interpretations

Our evaluation of the NDLP using administrative data has yielded a number of substantive and methodological findings. Substantively, we find that the NDLP had an economically meaningful impact on the time spent on benefit by participants. The impacts we estimate vary both with participant characteristics and over time. We find much larger impacts for lone parents who participate during a long spell of IS receipt ("the stock"), which we interpret as the effect of the program at pushing individuals near the margin into employment. This represents a one-time windfall for the government and reflects, in our view, the historically low rates of employment among lone mothers in the UK, the lack of much effort to push lone parents on IS into employment in the past and concurrent policy reforms designed to "make work pay". Due to the limitations of our data, the estimates for the stock also likely contain residual positive selection bias. We feel more confident in our modest, but positive and statistically and substantively significant, impacts for the "flow", and argue that these estimates provide a better guide to future policy. We also find that the impacts of NDLP fade out modestly over time, as non-participant benefit receipt levels slowly fall. Thus,

some fraction of the effect of NDLP comes from speeding up exits from benefit that would otherwise occur a few months later.

Methodologically, our analysis has a number of implications for the conduct of future evaluations. First, and most importantly, we show the importance of flexible conditioning on pre-program outcomes for removing selection bias. Simple summary measures of outcomes, such as the number of months on benefits or the length of the spell in progress at the time of participation, though helpful, lose an important part of the information contained in the outcome histories. Thus, our findings reinforce the lessons in Card and Sullivan (1988) and Heckman and Smith (1999).

Second, we find a surprisingly large effect from conditioning on the attitudinal variables measured on the postal survey. The use of such variables, which could even be collected routinely at intake and included in administrative data, may represent an important avenue for the improvement of non-experimental evaluation. Of course, further research on which attitudes to measure and how best to measure them needs to complement our own findings.

Third, we illustrate the value and importance of taking account of complex sample designs in evaluations using matching. We adopt a simple strategy, namely constructing separate estimates by stratum. This strategy requires a relatively large sample size per stratum, as in our data, to work well. The natural alternative strategy uses weights to take account of the complex sampling design; our strategy here has the value of highlighting the importance of the implicit interaction between the IS spell duration and the other conditioning variables.

Fourth, we show the value of focusing on participants in NDLP who participate early in the participation window that defines the treatment, given that we measure our benefit history variables relative to the beginning of the window. This conclusion follows from the rising importance of measurement error in these variables over time within the window. This finding reinforces the recent focus in the literature on the so-called “dynamic treatment effects” framework, that seeks to estimate the impact of participation at a particular point in time relative to non-participation at that time, conditional on variables measured up to that time.

Fifth, contrary to earlier findings using the data from the US JTPA evaluation, we find that conditioning on variables related to local labor markets, either at a very detailed level or at a regional level within the UK, has little effect on the resulting estimates. This raises some questions about the generality of the earlier findings and suggests the value of additional research on when local labor markets matter and why.

Sixth, our analysis highlights the potential for non-experimental evaluations using administrative data to produce credible impact estimates. Administrative data allow the flexible conditioning on benefit receipt histories that we find important. They also avoid issues of non-random non-response associated with survey-based evaluations and generally cost a lot less to produce and use than survey data. We thus indirectly highlight the value of producing relatively clean and well-documented administrative data that both outside researchers and program staff will find easy to use.

Finally, our preferred non-experimental estimates for the "flow" sample coincide well with (the upper end of) our priors based on experimental evaluations of similar programs for similar populations in the US. Of course, we did not fully specify our analysis plan in advance but instead undertook an iterative reproach wherein we refined our analysis over time as we learned more about the data, the evaluation design we inherited in part from the NCSR, and the NDLP program itself. This fact reduces but does not eliminate the value of our findings as evidence for the proposition that well designed non-experimental evaluations that build on the knowledge available in the literature (and have large samples to work with) can produce credible estimates that line up with the experimental literature. At the same time, the fact that the NCSR evaluation found an impact (on a related dependent variable) several times larger than our own preferred impact estimates reinforces the general point that non-experimental impact estimates will likely always have larger non-sampling variation than experimental ones.[11]

[11] Though the non-sampling variation in experimental estimates is not zero; see the evidence and discussion in Heckman and Smith (2000).

Bibliography

Abadie, Alberto and Guido Imbens. 2006. “On the Failure of the Bootstrap for Matching Estimators.” NBER Technical Working Paper No. 325.

Abadie, Alberto and Guido Imbens. 2008. “On the Failure of the Bootstrap for Matching Estimators.” *Econometrica* 76(6): 1537-1557.

Abadie, Alberto and Guido Imbens. 2011. “Bias-Corrected Matching Estimators for Average Treatment Effects.” *Journal of Business and Economic Statistics* 29(1): 1-11.

Abbring, Japp and Gerard van den Berg. 2004. “Analyzing the Effect of Dynamically Assigned Treatments Using Duration Models, Binary Treatment Models and Panel Data Models.” *Empirical Economics* 29(1): 5-20.

Bloom, Howard, Larry Orr, George Cave, Stephen Bell and Fred Doolittle. 1993. *The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months*. U.S. Department of Labor Research and Evaluation Report 93-C.

Busso, Matias, John DiNardo and Justin McCrary. 2009a. “New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators.” Unpublished manuscript, University of California at Berkeley.

Busso, Matias, John DiNardo and Justin McCrary. 2009b.”Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects.” Unpublished manuscript, University of California at Berkeley.

Caliendo, Marco and Sabine Kopeinig. 2008. “Some Practical Guidance for the Implementation of Propensity Score Matching.” *Journal of Economic Surveys* 22(1): 31-72.

Card, David, Jochen Kluve and Andrea Weber. 2010. “Active Labor Market Policy Evaluations: A Meta-Analysis.” NBER Working Paper No. 16173.

Card, David and Daniel Sullivan. 1988. “Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment.” *Econometrica* 56(3): 497-530.

Coleman, Nick, Nicola Rousseau and Matt Laycock. 2003. *National Evaluation of Lone Parents Adviser Meetings: Findings from a Longitudinal Survey of Clients*. BMRB Social Research.

Couch, Kenneth. 1992. “New Evidence on the Long-Term Effects of Employment and Training Programs. *Journal of Labor Economics* 10(4): 380-388.

Crump, Richard, V. Joseph Hotz, Guido Imbens and Oscar Mitnik. 2009. “Dealing with Limited Overlap in Estimation of Average Treatment Effects.” *Biometrika* 96(1):187-199.

Davidson, Carl and Stephen Woodbury. 1993. “The Displacement Effect of Reemployment Bonus Programs.” *Journal of Labor Economics* 11(4): 575-605.

Dolton, Peter, Azevedo, João Pedro and Jeffrey Smith. 2006. The Econometric Evaluation of the New Deal for Lone Parents. UK Department of Work and Pensions Research Report No 356.

Dolton, Peter and Donal O’Neill. 2002. “The Long-Run Effects of Unemployment Monitoring and Work-Search Programs: Experimental Evidence from the United Kingdom.” *Journal of Labor Economics* 20(2): 381-403.

Dolton, Peter, Chiara Rosazza Bondibene and Jonathan Wadsworth. 2010 “The UK National Minimum Wage in Retrospect.” *Fiscal Studies* 31(4): 509-534.

Djebbari, Habiba and Jeffrey Smith. 2008. “Heterogeneous Impacts in PROGRESA.” *Journal of Econometrics* 145: 64-80.

Eberwein, Curtis, John Ham and Robert LaLonde. 1997. “The Impact of Classroom Training on the Employment Histories of Disadvantaged Women: Evidence from Experimental Data.” *Review of Economic Studies* 64(4): 655-682.

Evans, Martin, Jill Eyre, Jane Millar and Sophie Sarre. 2003. *New Deal for Lone Parents: Second Synthesis Report of the National Evaluation*. University of Bath Centre for Analysis of Social Policy.

Fredriksson, P. and P. Johansson. 2008. “Dynamic Treatment Assignment.” *Journal of Business and Economic Statistics* 26(4): 435-445.

Friedlander, Daniel and Gary Burtless. 1995. *Five Years After: The Long-Term Effects of Welfare-to-Work Programs*. New York: Russell Sage Foundation.

Friedlander, Daniel and Philip Robins. 1995. “Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods.” *American Economic Review* 85(4): 923-937.

Frölich, Markus. 2004. “Finite Sample Properties of Propensity Score Matching and Weighting Estimators.” *Review of Economics and Statistics* 86(1): 77-90.

Frost, Robert. 1920. “The Road Not Taken.” In: Robert Frost (ed.), *Mountain Interval*. New York: Henry Holt.

Gregg, Paul and Susan Harkness. 2003. “Welfare Reform and Lone Parents’ Employment in the UK.” CMPO Working Paper No. 03/072.

Gregg, Paul, Susan Harkness and Sarah Smith. 2009. "Welfare Reform and Lone Parents' Employment in the UK." *Economic Journal* 119(535): F38-F65.

Gueron, Judith and Edward Pauly. 1991. *From Welfare to Work*. New York: Russell Sage Foundation.

Hales, Jon, Wendy Roth, Matt Barnes, Jane Millar, Carli Lessof, Mandy Glover and Andrew Shaw. 2000. *Evaluation of the New Deal for Lone Parents: Early Lessons from the Phase One Prototype - Findings of Surveys*. Department for Social Security Research Report 109.

Ham, John and Robert LaLonde. 1996. "The Effect of Sample Selection and Initial Conditions in Duration Models: Evidence from Experimental Data." *Econometrica* 64(1): 175-205.

Hämäläinen, Kari, Roope Uusitalo, and Jukka Vuori, 2008. "Varying Biases in Matching Estimates: Evidence from Two Randomised Job Search Training Experiments." *Labour Economics* 15(4): 604-618.

Hamilton, Gayle, Stephen Freedman, Lisa Gennetian, Charles Michalopoulos, Johanna Walter, Diana Adams-Ciardullo and Anna Gassman-Pines. 2001. *National Evaluation of Welfare to Work Strategies: How Effective Are Different Welfare-to-Work Approaches? Five-Year Adult and Child Impacts for Eleven Programs*. New York: Manpower Demonstration Research Corporation.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd. 1998a "Characterising Selection Bias Using Experimental Data." *Econometrica* 66(5): 1017-1098.

Heckman, James, Hidehiko Ichimura and Petra Todd. 1997a. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64(4): 605-654.

Heckman, James, Lance Lochner, and Christopher Taber. 1998b. "Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labor Earnings with Heterogeneous Agents." *Review of Economic Dynamics* 1(1): 1-58.

Heckman, James, Robert LaLonde, and Jeffrey Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In Orley Ashenfelter and David Card (eds.), *Handbook of Labor Economics, Volume 3A*. 1865-2097.

Heckman, James and Salvador Navarro. 2004. "Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models." *Review of Economics and Statistics* 86(1): 30-57.

Heckman, James and Salvador Navarro. 2007. "Dynamic Discrete Choice and Dynamic Treatment Effects." *Journal of Econometrics* 136(2): 341-396.

Heckman, James and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions." In James Heckman and Burton Singer (eds.), *Longitudinal Analysis of Labor Market Data*. New York: Cambridge University Press. 156-246.

Heckman, James and Jeffrey Smith. 1999. "The Pre-Programme Dip and the Determinants of Participation in a Social Programme: Implications for Simple Programme Evaluation Strategies." *Economic Journal* 109(457): 313-348.

Heckman, James and Jeffrey Smith. 2000. "The Sensitivity of Experimental Impact Estimates: Evidence from the National JTPA Study" in David Blanchflower and Richard Freeman (eds.), *Youth Employment and Joblessness in Advanced Countries*, Chicago: University of Chicago Press for NBER, 331-356.

Heckman, James, Jeffrey Smith and Nancy Clements. 1997b. "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies* 64(4): 487-537.

Heinrich, Carolyn, Peter Mueser, Kenneth Troske, Kyung-Seong Jeon and Daver Kahvecioglu. 2009. "New Estimates of Public Employment and Training Program Net Impacts: A Nonexperimental Evaluation of the Workforce Investment Act Program." IZA Discussion Paper No. 4569.

Hollister, Robinson and Rebecca Maynard. 1984. "The Impacts of Supported Work on AFDC Recipients." In Robinson Hollister, Peter Kemper and Rebecca Maynard (eds.), *The National Supported Work Demonstration*. Madison: University of Wisconsin Press. 90-135.

Hoynes, Hillary. 2000. "Local Labor Markets and Welfare Spells: Do Demand Conditions Matter?" *Review of Economics and Statistics* 82(3): 351-368.

Hotz, V. Joseph, Guido Imbens and Jacob Klerman. 2006. "Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Reanalysis of the California GAIN Program." *Journal of Labor Economics* 24(3): 521-566.

Hotz, V. Joseph, Guido Imbens and Emily Mortimer. 2005. "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations." *Journal of Econometrics* 125: 241-270.

Hotz, V. Joseph and Karl Scholz. 2002. "Measuring Employment and Income Outcomes for Low-Income Populations with Administrative and Survey Data." In *Studies of Welfare Populations: Data Collection and Research Issues*. National Research Council: National Academy Press. 275-315.

Huber, Martin, Michael Lechner and Conny Wunsch. 2010. "How To Control for Many Covariates? Reliable Estimators Based on the Propensity Score." IZA Discussion Paper No. 5268.

Ichimura, Hidehiko and Petra Todd. 1999. “Alignment Problem in Matching Estimators for Longitudinal Data.” Unpublished manuscript, University of Pennsylvania.

Imbens, Guido. 2004. “Nonparametric Estimation of Average Treatment Effects Under Exogeneity.” *Review of Economics and Statistics* 86(1): 4-29.

Khan, Shakeeb and Elie Tamer. 2010. “Irregular Identification, Support Conditions and Inverse Weight Estimation.” *Econometrica* 78(6): 2021-2042.

Knight, Genevieve, Stefan Speckesser, Jeffrey Smith, Peter Dolton, and João Pedro Azevedo. 2006. Lone Parents Work Focused Interviews/New Deal for Lone Parents: Combined Evaluation and Further Net Impacts. Department of Work and Pensions Research Report no 368.

LaLonde, Robert. 1995. “The Promise of Public-Sponsored Training Programs.” *Journal of Economic Perspectives* 9(2): 149-168.

Lechner, Michael. 1999. “Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification.” *Journal of Business and Economic Statistics* 17(1): 74-90.

Lee, Wang-Sheng. 2009. “Propensity Score Matching and Variations on the Balancing Test.” Unpublished manuscript, University of Melbourne.

Lessof, Carli, Melissa Miller, Miranda Phillips, Kevin Pickering, Susan Purdon and Jon Hales. 2003. New Deal for Lone Parents Evaluation: Findings from the Quantitative Survey. Department for Work and Pensions WAE Report 147.

Leuven, Edwin and Barbara Sianesi. 2003. “PSSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.” [Available from http://fmwww.bc.edu/RePEc/bocode/p.]

Lise, Jeremy, Shannon Seitz and Jeffrey Smith. 2004. “Equilibrium Policy Experiments and the Evaluation of Social Programs.” NBER Working Paper No. 10283.

Michalopoulos, Charles, Doug Tattrie, Cynthia Miller, Philip Robins, Pamela Morris, David Gyarmati, Cindy Redcross, Kelly Foley and Reuben Ford. 2002. *Making Work Pay: Final Report on the Self-Sufficiency Project for Long-Term Welfare Recipients*. Ottawa: SRDC.

Mueser, Peter, Kenneth Troske and Alexey Gorislavsky. 2007. “Using State Administrative Data to Measure Program Performance.” *Review of Economics and Statistics* 89(4): 761-783.

OECD. 2001. *OECD Employment Outlook 2001*. Paris: OECD.

Orr, Larry, Howard Bloom, Stephen Bell, Fred Doolittle, Winston Lin and George Cave. 1996. *Does Training Work for the Disadvantaged?* Washington, DC: Urban Institute Press.

Phillips, Miranda, Kevin Pickering, Carli Lessof, Susan Purdon and Jon Hales. 2003. Evaluation of the New Deal for Lone Parents: Technical report for the Quantitative Survey. Department for Work and Pensions WAE Report 146.

Plesca, Miana and Jeffrey Smith. 2007. "Evaluating Multi-Treatment Programs: Theory and Evidence from the U.S. Job Training Partnership Act" *Empirical Economics* 32(2-3): 491-528.

Røed, Knut and Oddbjørn Raaum. 2003. "Administrative Registers – Unexplored Reservoirs of Scientific Knowledge?" *Economic Journal* 113(488): F258-F281.

Rosenbaum, Paul and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41-55.

Schochet, Peter, John Brughardt and Steven Glazerman. 2001. *National Job Corps Study: The Impacts of Job Corps on Participants' Employment and Related Outcomes*. Princeton: Mathematica Policy Research.

Sianesi, Barbara. 2004. "An Evaluation of the Swedish System of Active Labour Market Programmes in the 1990s." *Review of Economics and Statistics* 86(1): 133-155.

Smith, Jeffrey and Petra Todd. 2005a. "Does Matching Overcome LaLonde's Critique of Nonexperimental Methods?" *Journal of Econometrics* 125(1-2): 305-353.

Smith, Jeffrey and Petra Todd. 2005b. "Rejoinder." *Journal of Econometrics* 125(1-2): 365-375.

Suhrcke, Marc, Carmen de Paz Nieves, Cristina Otano and Adam Coutts. 2009 "Lone Parent Policies in the UK: The Impact of the New Deal for Lone Parents (NDLP) on Socioeconomic and Health Inequalities". Mimeo.

UK Department for Work and Pensions. 2002. "Lone Parent Brief." Unpublished Manuscript, Sheffield.

US General Accounting Office. 1996. Job Training Partnership Act: Long-Term Earnings and Employment Outcomes. GAO/HEHS-96-40.

Table 1 Sample Composition.

Strata	Age of youngest Child (years)	IS Spell Duration (months)	*Total Eligible Population*			*NDLP Participants*			*Total Sample*			
			Wave 1/2	Booster	Total	Wave 1/2	Booster	Total	Wave 1/2	Booster	Total	Sample Rate
1	[0,3)	[0,3)	1,853	10,139	11,992	119	130	249	1,834	2107	3,941	0.329
2	[0,3)	[3,6)	6,198	0	6,198	124	0	124	1,814	0	1,814	0.293
3	[0,3)	[6,12)	11,405	0	11,405	238	0	238	3,503	0	3,503	0.307
4	[0,3)	[12,24)	17,883	0	17,883	320	0	320	5,354	0	5,354	0.299
5	[0,3)	[24,36)	11,347	0	11,347	139	0	139	2,869	0	2,869	0.253
6	[0,3)	[36,∞)	21,122	0	21,122	122	0	122	4,174	0	4,174	0.198
7	[3,5)	[0,3)	782	3,899	4,681	69	53	122	779	688	1,467	0.313
8	[3,5)	[3,6)	2,071	0	2,071	161	0	161	2,055	0	2,055	0.992
9	[3,5)	[6,12)	3,264	0	3,264	93	0	93	1,303	0	1,303	0.399
10	[3,5)	[12,24)	5,838	0	5,838	115	0	115	1,810	0	1,810	0.31
11	[3,5)	[24,36)	4,899	0	4,899	106	0	106	1,428	0	1,428	0.291
12	[3,5)	[36,∞)	22,425	0	22,425	267	0	267	4,568	0	4,568	0.204
13	[5,11)	[0,3)	1,435	7,401	8,836	134	106	240	1,419	1046	2,465	0.279
14	[5,11)	[3,6)	3,932	0	3,932	334	0	334	3,885	0	3,885	0.988
15	[5,11)	[6,12)	5,687	0	5,687	205	0	205	2,825	0	2,825	0.497
16	[5,11)	[12,24)	9,819	0	9,819	193	0	193	2,981	0	2,981	0.304
17	[5,11)	[24,36)	7,337	0	7,337	124	0	124	2,213	0	2,213	0.302
18	[5,11)	[36,∞)	48,290	0	48,290	497	0	497	9,660	0	9,660	0.200
19	[11,16)	[0,3)	815	4,384	5,199	48	69	117	807	827	1,634	0.314
20	[11,16)	[3,6)	2,229	0	2,229	55	0	55	646	0	646	0.290
21	[11,16)	[6,12)	3,189	0	3,189	51	0	51	911	0	911	0.286
22	[11,16)	[12,24)	5,207	0	5,207	47	0	47	1,027	0	1,027	0.197
23	[11,16)	[24,36)	3,849	0	3,849	41	0	41	909	0	909	0.236
24	[11,16)	[36,∞)	29,990	0	29,990	271	0	271	6,027	0	6,027	0.201
Total			230,866	25,823	256,689	3,873	358	4,231	64,801	5,028	69,829	0.272

Table 2. Exact Matching on Benefit History Strings

Benefit history string	*Number of non-participants*	*Proportion of non-participants on benefit (%)*	*Number of participants*	*Proportion of participants on benefit (%)*	*Differences*	*Proportion of participants in each stratum*	*Cell-specific treatment effect contribution*	*Proportion of participants in each stratum (excluding stratum 111111)*	*Cell-specific treatment effect contribution*
					(D)-(B)		*(E)x(F)*		*(E)x(G)*
	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)
111111	42,408	82.3	2,276	60.1	-22.2	0.54	-11.91	-	-
000001	4,154	62.2	404	51.3	-10.9	0.1	-1.03	0.21	-2.24
000000	2,502	66.2	225	47.4	-18.8	0.05	-1	0.11	-2.16
000011	3,658	65.3	349	55.8	-9.6	0.08	-0.78	0.18	-1.7
011111	2,598	72.5	196	55.6	-16.9	0.05	-0.78	0.1	-1.69
000111	2,651	69.1	206	57.1	-12	0.05	-0.58	0.11	-1.26
001111	2,198	70	165	57.2	-12.8	0.04	-0.5	0.08	-1.08
100001	367	61.2	41	49	-12.2	0.01	-0.12	0.02	-0.26
101111	330	76.1	28	58.5	-17.6	0.01	-0.12	0.01	-0.25
110111	329	79.2	26	64.6	-14.6	0.01	-0.09	0.01	-0.19
111011	382	73.8	28	60.5	-13.4	0.01	-0.09	0.01	-0.19
111100	210	71.8	20	54.4	-17.4	0	-0.08	0.01	-0.18
111101	354	72.6	21	56.3	-16.3	0	-0.08	0.01	-0.17
110001	328	60.4	21	45.6	-14.9	0	-0.07	0.01	-0.16
110011	479	68.7	36	61.1	-7.6	0.01	-0.06	0.02	-0.14
Others	2,646		193				-0.32		-0.54
Total	65,594		4,235				-17.61		-12.33

Notes: (1) In the spirit of Card and Sullivan (1988), we adopt the following approach. First, we break the period from June 1999 to September 2000 (the period over which we have complete data on benefit receipt) into six 11 week "quarters", where we omit the final week just prior to the start of the participation window. We code an indicator variable for each quarter that indicates whether or not the individual spent at least half the period on benefit. We then concatenate the six different indicators into a string. There are $2^6 = 64$ possible strings, ranging from 000000 to 111111. A string of 111111 indicates someone who spent at least half of all six quarters on benefit; similarly, a string of 000000 indicates someone who did not spend at least half of any of the six quarters on benefit. (2) This analysis does not take the sample stratification into consideration.

Table 3 . Propensity Score Probit Model for Stratum 1

	Coef/S.E.	*Mean*	*Std Dev*
Disabled	-0.901	0.030	0.171
	-0.570		
Length of time disabled	-0.329*	0.157	0.907
	-0.165		
On benefit at week 39 in 2000	0.316	0.970	0.171
	-0.278		
On benefit at week 38 in 2000	-0.216	0.936	0.246
	-0.218		
On benefit at week 37 in 2000	0.275	0.883	0.322
	-0.190		
On benefit at week 36 in 2000	-0.034	0.822	0.382
	-0.159		
On benefit at week 35 in 2000	0.076	0.740	0.439
	-0.141		
On benefit at week 34 in 2000	-0.206	0.641	0.480
	-0.142		
Proportion of time on benefit prior to June 1999	-0.217	0.129	0.280
	-0.176		
(categorical variables omitted – available upon request)			
Joint significance of categorical variables	Chi-squared statistic / p-value		
Age	13.691		
	0.057		
Region	10.787		
	0.461		
Age of youngest child	10.071		
	0.018		
Number of children	27.731		
	0.000		
Benefit history variables	22.186		
	0.877		
Observations	3788		
R-squared	0.061		
Log-likelihood	-857.023		
Chi squared statistic	112.207		

Table 4. Estimated Treatment Effects for the Stock and the Flow

Time Period	*ATT*	*Stock*	*Flow*
All post-programme	19.09	21.04	9.87
	[0.43]	[0.65]	[0.57]
3 months after start of window	18.2	22.55	8.78
	[5.29]	[7.65]	[6.73]
9 months after start of window	22.28	25.93	10.84
	[1.95]	[2.89]	[2.68]
24 months after start of window	18.98	20.13	9.99
	[0.80]	[1.23]	[1.10]
36 months after start of window	15.83	17.33	8.77
	[2.61]	[4.11]	[3.17]

Notes: (1) Full sample; (2) estimated bootstrap standard errors appear in square brackets (3) we define the stock as those individuals who spent more than 50 percent of the weeks in each of the six "quarters" prior to the start of the NDLP participation window on benefit and we define the flow as the complement of the stock. In terms of the benefit history strings, the stock consists of individuals with a value of 111111 and the flow consists of everyone else; (4) the analysis is done separately by stratum; the overall estimate consists of a weighted average of the estimates for individual strata.

Table 5 Estimating Treatment Effects for Subgroups.

Description	*Treatment effect*	*Pre-window difference*	*Difference*
Children at the age of [0,3)	17.2	-0.88	18.08
	[0.89]	[0.37]	[0.98]
Children at the age of [3,5)	21.43	1.21	20.22
	[0.82]	[0.36]	[0.91]
Children at the age of [5,11)	20.6	0.26	20.33
	[0.84]	[0.35]	[0.93]
Children at the age of [11,16)	18.03	1.27	16.76
	[0.85]	[0.45]	[0.98]
On IS for less than 3 months	16.6	1.35	15.25
	[0.83]	[0.58]	[1.03]
On IS from [3,6) months	11.73	-2.27	14.01
	[0.94]	[0.61]	[1.13]
On IS from [6,12) months	18.12	1.33	16.78
	[0.87]	[0.60]	[1.07]
On IS from [12,24) months	15.84	0.96	14.87
	[0.96]	[0.54]	[1.12]
On IS from [24,36) months	20.3	0.22	20.09
	[0.92]	[0.47]	[1.05]
On IS for more than 36 months	24.93	0.82	24.11
	[1.07]	[0.24]	[1.12]

Notes: (1) Estimated bootstrap standard errors appear in square brackets; (2) the impact estimates refer to the entire post-programme period; (3) the subgroups are define according to the administrative database used to draw the initial National Centre sample; (4) the age of the children refers to the youngest child in the household as of August 2000; (5) the analysis is done separately by stratum; the overall estimate consists of a weighted average of the estimates for individual strata.

Table 6. Sensitivity Analyses: Window Width, Benefit History Variables and Local Labour Market Variables

Specification	Full		Flow		Stock	
	Treatment	*Pre Diff*	*Treatment*	*Pre Diff*	*Treatment*	*Pre Diff*
Our preferred specification						
Our preferred specification	19.09	0.46	9.87	0.49	21.04	1.97
	[0.43]	[0.19]	[0.57]	[0.17]	[0.65]	[0.45]
Alternative Group of Participants						
Our preferred speciation with participants from the first half of the window	14.74	1.24	5.21	0.06	17.48	6.23
	[0.90]	[0.44]	[0.82]	[0.28]	[1.00]	[0.63]
Alternative Benefit History Specifications						
No benefit history variables	18.19	2.52	10.02	2.05	19.78	2.16
	[0.45]	[0.23]	[0.58]	[0.25]	[0.64]	[0.49]
Only the fraction of time on benefit from June 1999 to August 2000	18.23	1.22	9.36	0.98	19.06	1.95
	[0.44]	[0.21]	[0.55]	[0.20]	[0.66]	[0.49]
Only the duration of the benefit spell in progress at the start of the window	18.53	2.17	10.93	2.53	20.37	3.66
	[0.43]	[0.22]	[0.58]	[0.29]	[0.64]	[0.51]
Just the benefit history strings	17.79	0.85	8.11	0.22	19.79	2.37
	[0.43]	[0.21]	[0.55]	[0.19]	[0.64]	[0.47]
Alternative Geographical Specifications						
No geographic information	20.11	-0.37	10.66	-0.68	21.91	2.12
	[0.52]	[0.16]	[0.58]	[0.16]	[0.68]	[0.43]
Detailed local council area variables	20.38	0.75	9.6	-0.42	20.5	1.28
	[0.40]	[0.20]	[0.54]	[0.17]	[0.61]	[0.47]

Notes: (1) Estimated bootstrap standard errors appear in square brackets; (2) the impact estimates refer to the entire post-programme period; (3) the analysis is done separately by stratum; the overall estimate consists of a weighted average of the estimates for individual strata.

Table 7. Attitudinal Variables

Description	*Sample*	Full		Flow		Stock	
Preferred Model		*Treatment*	*Pre-Diff*	*Treatment*	*Pre-Diff*	*Treatment*	*Pre-Diff*
No attitudinal variables and no benefit history information	Postal Survey Respondents	21.37	3.29	9.27	2.12	25.06	4.21
		[0.50]	[0.27]	[0.75]	[0.22]	[0.75]	[0.61]
No attitudinal variables but including benefit history information	Postal Survey Respondents	20.18	0.93	9.58	0.08	24.84	0.34
		[0.49]	[0.22]	[0.74]	[0.23]	[0.79]	[0.57]
Attitudinal variables included but no benefit history information.	Postal Survey Respondents	17.8	1.3	5.44	0.80	13.01	-2.08
		[0.65]	[0.35]	[0.59]	[0.29]	[0.98]	[0.84]
Includes attitudinal variables and benefit history information.	Postal Survey Respondents	17.53	0.31	4.9	0.41	17.6	2.44
		[0.63]	[0.33]	[0.57]	[0.26]	[0.99]	[0.87]

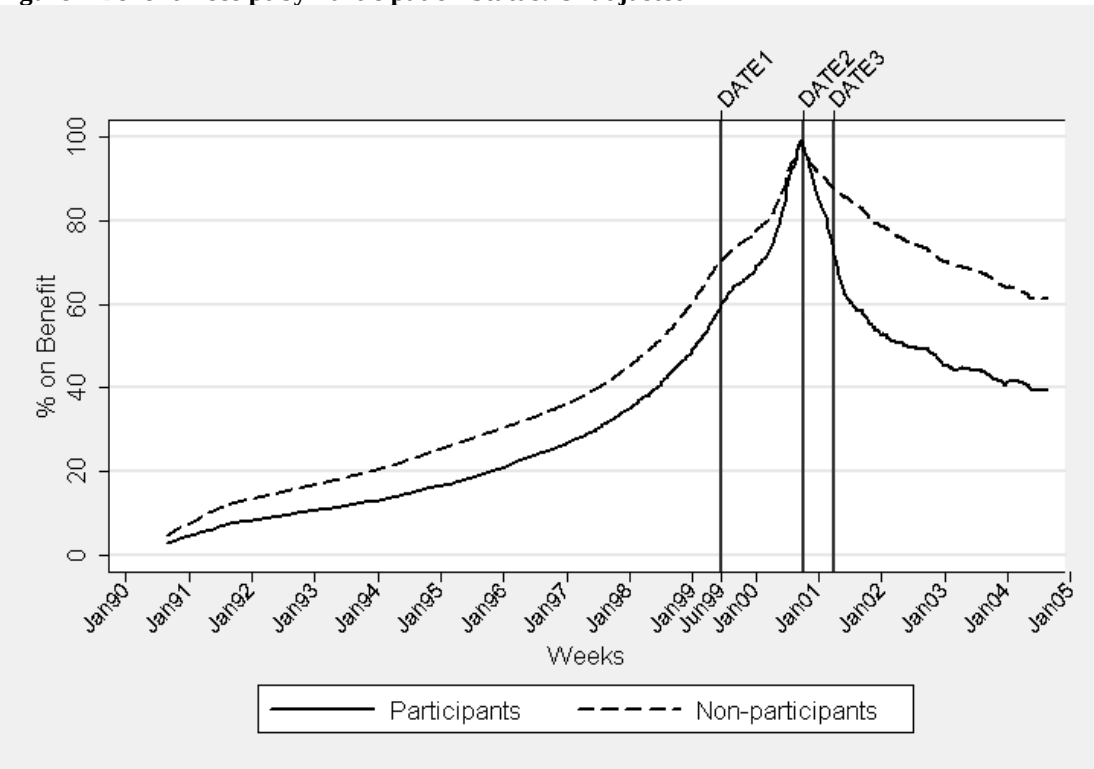
Notes: (1) Estimated bootstrap standard errors appear in square brackets; (2) the impact estimates refer to the entire post-program period; (3) the analysis is done separately by stratum and the overall estimate consists of a weighted average of the estimates for the individual strata; and (4) the postal survey sample includes 42,249 observations.

Table A-1 Descriptive Statistics for Variables
Included in the Propensity Score Model

Description	Participants	Non-participants
Sex = male	0.06	0.06
Age (category) = 20 - 24	0.16	0.15
Age (category) = 25 - 29	0.21	0.20
Age (category) = 30 - 34	0.23	0.23
Age (category) = 35 - 39	0.19	0.19
Age (category) = 40 - 44	0.10	0.11
Age (category) = 45 - 49	0.04	0.05
Age (category) = 50 - 54	0.01	0.02
Region = north west	0.16	0.15
Region = merseyside	0.06	0.05
Region = yorkshire & humber	0.05	0.05
Region = east midlands	0.07	0.07
Region = west midlands	0.06	0.08
Region = south west	0.08	0.06
Region = eastern	0.04	0.04
Region = london	0.14	0.17
Region = south east	0.10	0.10
Region = wales	0.08	0.07
Region = scotland	0.11	0.12
Age of the youngest child (category)== 2 yrs	0.08	0.08
Age of the youngest child (category)= 3 to 5 yrs	0.28	0.23
Age of the youngest child (category)= 6 to 7 yrs	0.12	0.12
Age of the youngest child (category)= 8 to 11 yrs	0.20	0.18
Age of the youngest child (category)= 12 to 19 yrs	0.11	0.13
Number of children (category) = 2	0.34	0.33
Number of children (category) = 3	0.11	0.15
Number of children (category) = 4 to 8	0.04	0.07
Benefit history = 000001	0.03	0.09
Benefit history = 000010	0.13	0.47
Benefit history = 000011	0.10	0.06
Benefit history = 000101	0.08	0.06
Benefit history = 001001	0.05	0.04
Benefit history = 010001	0.04	0.03
Benefit history = 100001	0.05	0.04
Benefit history = 100101	0.01	-
Benefit history = 110001	0.01	-
Benefit history = 110111	0.01	0.01
Benefit history = 111000	0.01	0.01
Benefit history = 111101	0.01	-
Benefit history = 111110	0.01	-

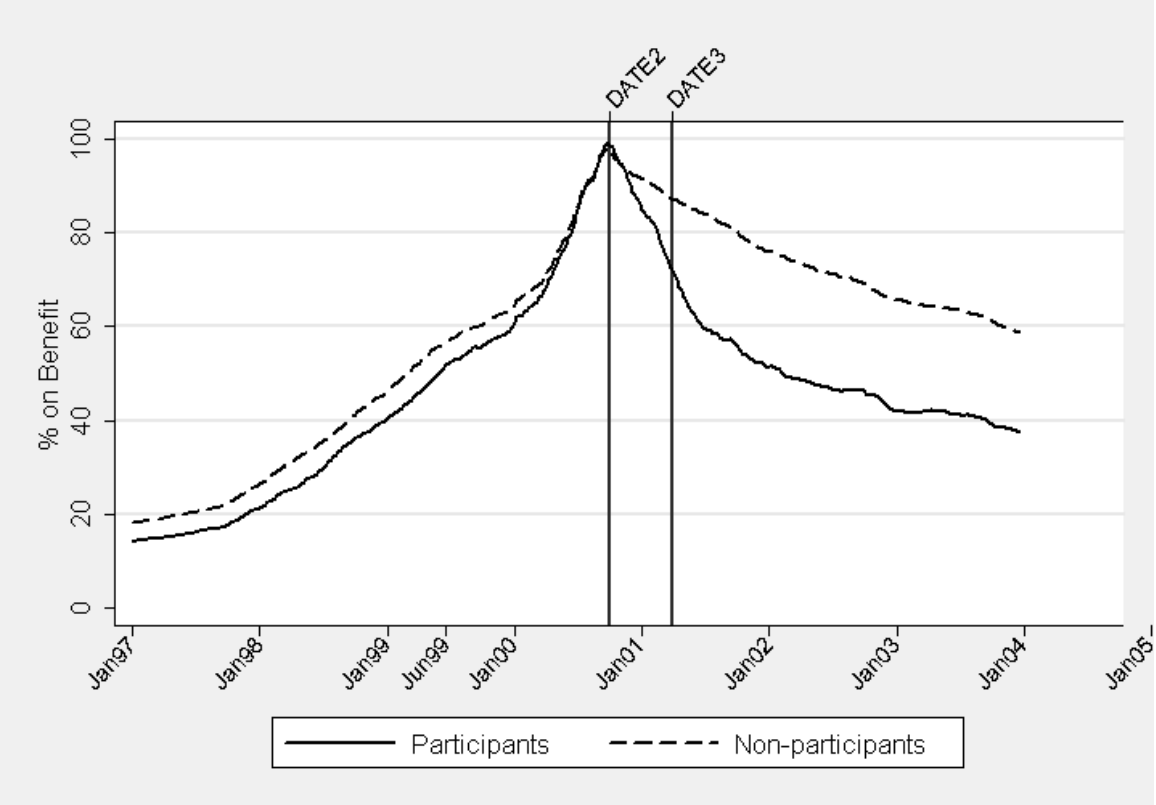
Benefit history = 111111	0.01	0.01
Disabled lone parent	0.55	0.66
On benefit in week prior to benefit window	0.99	0.98
On benefit in week two weeks prior to benefit window	0.99	0.98
On benefit in week three weeks prior to the benefit window	0.98	0.98
On benefit in week four weeks prior to the benefit window	0.97	0.97
On benefit in week five weeks prior to the benefit window	0.96	0.96
On benefit in week six weeks prior to the benefit window	0.94	0.95
Proportion of time on benefit from June 1999 to August 2000	0.40	0.50

Figure 1 Benefit Receipt by Participation Status: Unadjusted



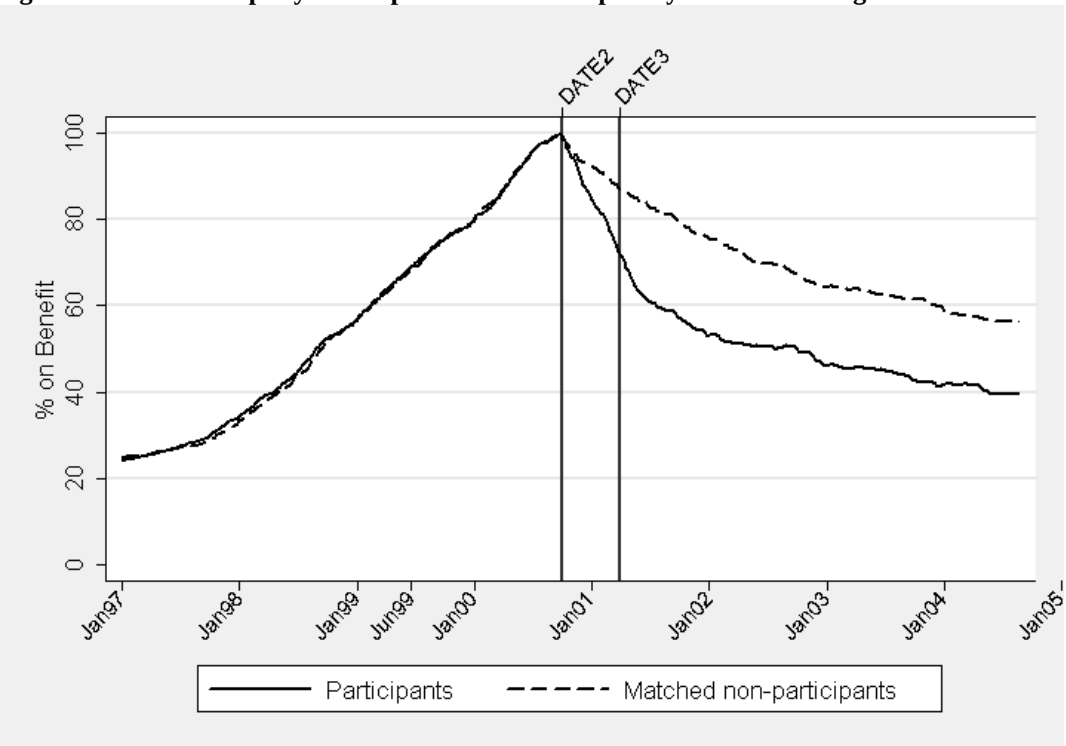
Notes: DATE1 is the data at which complete benefit history data become available. DATE2 and DATE3 define the participation window. These estimates ignore the sample stratification.

Figure 2 Benefit Receipt by Participation Status: Matching on Strata



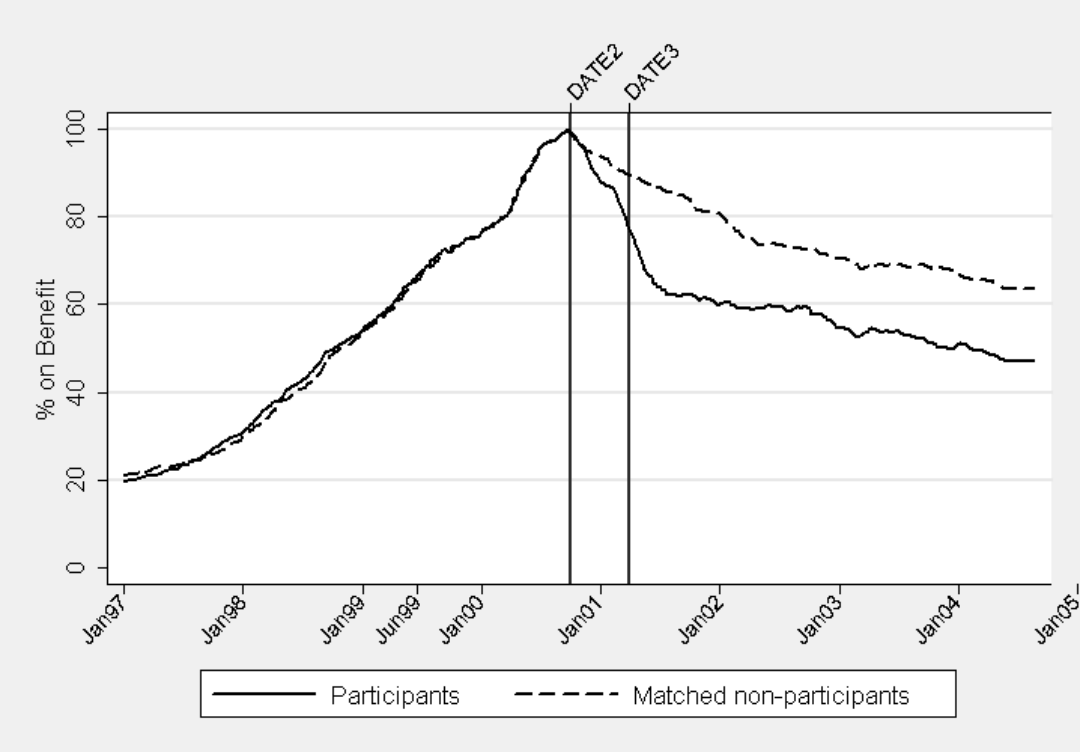
Notes: DATE1 is the data at which complete benefit history data become available. DATE2 and DATE3 define the participation window. These estimates ignore the sample stratification.

Figure 3 Benefit Receipt by Participation Status: Propensity Score Matching



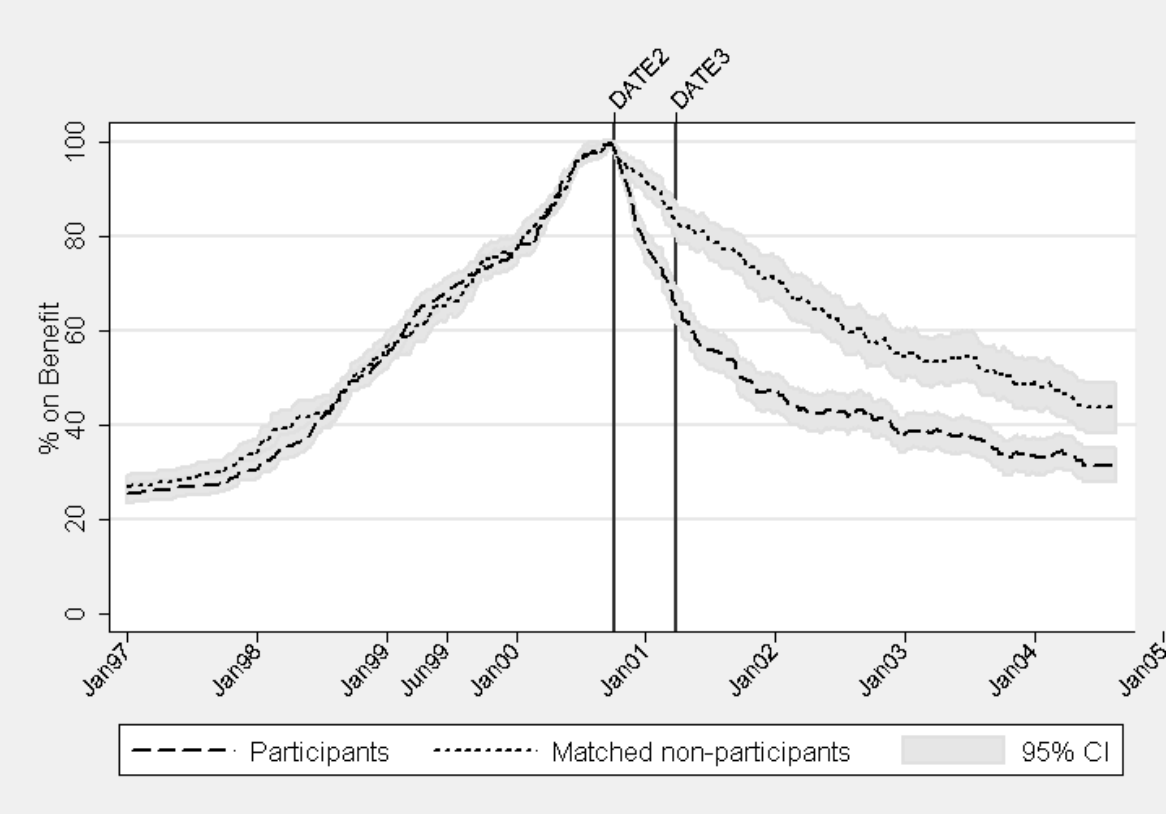
Notes: DATE2 and DATE3 define the participation window. The estimates in the figure represent a weighted (by the fraction of treated units in the stratum) average of separate propensity score matching estimates obtained for each stratum.

Figure 4. Benefit Receipt by Participation Status: Propensity Score Matching Youngest Child Age 0-3.



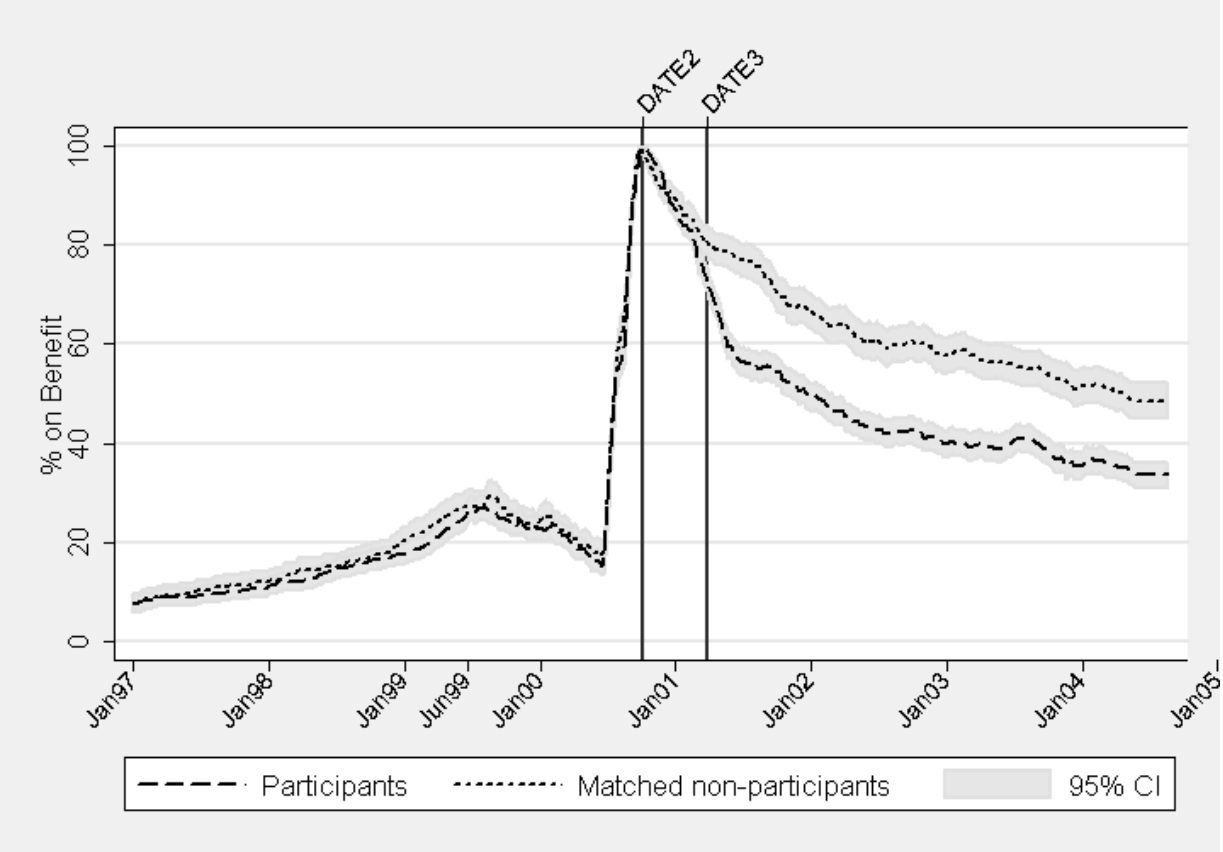
Notes: DATE2 and DATE3 define the participation window. The estimates in the figure represent a weighted (by the fraction of treated units in the stratum) average of separate propensity score matching estimates obtained for each stratum in this subgroup.

Figure 5. Benefit Receipt by Participation Status: Propensity Score Matching Youngest Child Age 11-16.



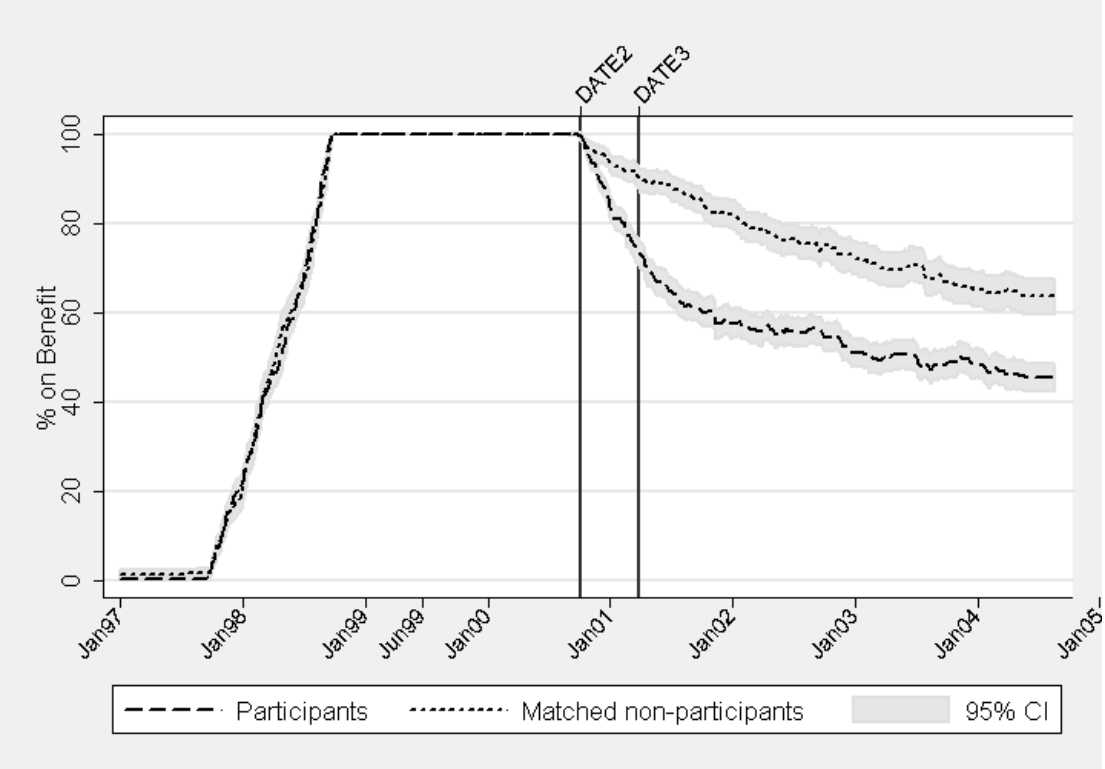
Notes: DATE2 and DATE3 define the participation window. The estimates in the figure represent a weighted (by the fraction of treated units in the stratum) average of separate propensity score matching estimates obtained for each stratum in this subgroup.

Figure 6 Benefit Receipt by Participation Status: Propensity Score Matching On IS for Less Than 3 Months.



Notes: DATE2 and DATE3 define the participation window. The estimates in the figure represent a weighted (by the fraction of treated units in the stratum) average of separate propensity score matching estimates obtained for each stratum in this subgroup.

Figure 7 Benefit Receipt by Participation Status: Propensity Score Matching On IS for 2 to 3 years.



Notes: DATE2 and DATE3 define the participation window. The estimates in the figure represent a weighted (by the fraction of treated units in the stratum) average of separate propensity score matching estimates obtained for each stratum in this subgroup.