

The Econometric Evaluation of the New Deal for Lone Parents

Peter Dolton
Department of Economics
Royal Holloway, University of London & London School of Economics
peter.dolton@rhul.ac.uk

Jeffrey Smith
Department of Economics
University of Michigan
econjeff@umich.edu

João Pedro Azevedo
Institute of Applied Economic Research (IPEA), Brazil
jazevedo@ipea.gov.br

Version of May18, 2006

We thank participants at numerous presentations of this paper and our related research 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, 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" (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. As such, our findings have relevance both 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 , including a variety of attitudinal measures, we find that these variables add little once we have conditioned flexibly on the lagged outcomes available in our data. Third, relative to a regression-based analysis that also assumed selection on observables, matching does not require an assumption of a linear conditional mean function and allows a careful examination of the support issue. Fourth, we lack access to plausible exclusion restrictions due to the nature of the NDLP program and of our data. While the presence or absence of an instrument does not affect the plausibility of our selection-on-observables assumption, it does reduce the choice set of available evaluation strategies.

We examine the impact of NDLP participation on individuals eligible for NDLP 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 IS, a group we call "the stock", we estimate a reduction in probability of being on IS of 20.45 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 of probability of being on IS of 14.24

percentage points. The difference between the stock and flow estimates suggests a huge one-off gain from exposing 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.

Though surprisingly large, our estimates turn out 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 in detail in the text.

Methodologically, our analyses support the general conclusion in the literature regarding the importance of pre-program outcome measures in reducing (and hopefully eliminating) selection bias in non-experimental studies. Moreover, we show, building on Card and Sullivan (1988), Heckman, Ichimura, Smith and Todd (1998) 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. In our view, 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 small 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 no effect on the impact estimates. This indicates that lagged outcomes embody these otherwise unobserved factors and provides further support for our selection-on-observables evaluation strategy. In a parallel analysis, we also 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.

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, Smith and Azevedo (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 6 compares our estimates to those in the literature. Section 7 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 does not provide participants with additional benefits beyond those for which they already qualified. 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.

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 on their participants 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 survey, in round figures, there were approximately 100,000 participants and the total program costs were around £40.9 million giving a per unit cost of around £400 per participant. This last characteristic 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 literature on evaluating active labor market policies.

2.2 Policy environment

¹ As documented in Dolton et al. (2005) 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, and a 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.

Lone parents in the UK receive means-tested income support (IS) payments that depend on the number of their of their school age children and on the amount of other income they receive. They may 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. The nature of the financial support available to lone parents is made up of a package of IS payments, Child Benefit and WFTC. Their precise financial circumstances will depend most crucially on the income from paid work, and their housing costs. Details of these arrangements and how they have changed over the last ten years are provided in Gregg and Harkness (2003). The access to childcare and its availability and cost vary enormously across the country specifically for children under 4. Access to Day Care, Nursery and Kindergarten also vary according to where the an individual lives.

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: [This ‘employment gap’ is 24 percentage points in the UK whereas in most other OECD countries single mothers are more likely to work than married mothers and indeed in Italy and Spain, for example single mothers have 27 and 23 percentage points *higher* employment rates 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 three other important 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, the UK reorganized its system of employment and training programs in the form of the Job Centre Plus system. This new system includes case management, “one stop” centers, performance standards and all the rest of the currently popular design features for these schemes. Third, 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 push to move them into work. Second, the programme changes helped to make work more attractive relative to IS receipt; when the PA calculated the costs and benefits of work, work may have looked 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 October 1998. Our study focuses on the Phase Three period.

NDLP has expanded its target population and rules of eligibility over time. Initially, NDLP was rolled out to lone parents making new claims for IS whose youngest child was aged over five years and three months. By October of 1998 the roll out was to include those lone parents whose youngest child is aged over five years and three months who had made a 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 diminished over time. In November 2001 (not long after our participants participated), all lone parents not in work or working fewer than 16 hours a week, including those who not receiving benefits, became eligible for NDLP.

The NDLP 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 cleverly 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 excludes 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. 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. (2005) for more details about the definitions of the Primary Sampling Units (PSUs), the exclusion of certain PSUs, and other sampling issues.

Table 1 shows the composition of the sample relative to the population in the selected 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 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 short 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 return 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. (2005) 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) – this one 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. (2005), working in consultation with staff of the Department for Work and Pensions, 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, IB and JSA from June 28, 1999 to the week August 26, 2004. From September 1, 1990 we have data only on JSA spells and that only for spells in progress on June 18, 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 NDLP caseload, and from some of the other evaluation studies, such as Evans et al. (2002; pg. 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 to the side the issues raised in the recent literature on dynamic treatment effects – see e.g. Ichimura and Todd (1999), Frederiksson and Johansson (2003), Abbring and van den Berg (2004), Sianesi (2004) and Heckman and Navarro (2005). 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, LaLonde and Smith (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 address the implications of failing to address the dynamic issue for our estimation method and for the interpretation of our results later in the paper (and we plan to estimate dynamic participation models using these data in future work).

Dolton et al. (2005) examine the fraction of non-participants as defined 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. (2005) 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 (though new data on employment will allow us to use that as a dependent variable in future versions of this paper). As we define it here, benefit receipt means receiving any one of income support, unemployment insurance (called “Job Seekers Allowance” (JSA) in the UK) or incapacity benefit (roughly the UK analog of SSI and SSDI in the US). By using a broad benefit receipt measure, we bring our benefit receipt measure closer to one minus an employment indicator; but we do not get all the way there because some individuals leave the program without obtaining work.

Looking at benefit participation 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, Ham and LaLonde (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 collecting 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. (2005), 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 late in the window, this time interval includes a few pre-program weeks as well.

4. Methods

4.1 Framework

We adopt the standard evaluation framework in the literature. This framework has many names in the literature; we refer to it here as the Platonic potential outcomes framework, as Plato (approx. 390 BC) clearly outlines the notion of potential outcomes in his allegory of the cave, where the treated see clearly while the untreated see only shadows. Later scholars associated with this framework, such as Neyman (1923), Fisher (1935), Roy (1951), Quandt (1972) or Rubin (1974) merely recast the original Platonic construct.

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 mean impact of treatment on the treated, given by $\Delta_{TT} = E(Y_1 - Y_0 | D = 1) = E(Y_1 | D = 1) - E(Y_0 | D = 1)$ 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 Heckman, Smith and Clements (1997) and Heckman, LaLonde and Smith (1999) 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, one can also tell stories of 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, Lochner and Taber (1998) and Lise, Seitz and Smith (2005).

4.2 Identification using the CIA

We adopt what Heckman and Robb (1985) call a “selection on observables” identification strategy to identify Δ_{TT} . This requires that we adopt what the economics literature calls the Conditional Independence Assumption (CIA) and the statistics literature 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 subset of observed covariates. Following Heckman, Ichimura, Smith and Todd (1998), 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 (where an instrument is a variable that is unrelated to the untreated outcome but affects participation) that we do not observe.

The literature suggests the potential for conditioning flexibly on detailed benefit receipt histories to remove selection bias. Heckman and Smith (1999) and Heckman, Ichimura, Smith and Todd (1998) find this. The Monte Carlo analysis in Section 8.3 of Heckman, LaLonde and Smith (1999) shows that conditioning on lagged outcomes substantially reduces bias for a wide variety of individual outcome processes. We also provide evidence in Section 6.5 that our lagged outcome variables capture the information in variety of survey measures including attitudinal variables. In terms of what determines participation conditional on observables 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 variation in child health status or 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. below.

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 $P(X) = \Pr(D = 1 \mid X)$, 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 .

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 without 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) for additional discussion of matching and more technical detail about alternative matching estimators.

Single nearest neighbor 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. Frölich (2004) demonstrates, in his fine Monte Carlo analysis of alternative matching algorithms, a non-trivial cost in terms of mean squared error 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 strata, as we do in many of our analyses, would become infeasible (within a reasonable time frame) unless we relied on single nearest neighbor matching.

4.4 Matching with stratified samples

Dolton, Smith and Azevedo (2006) provide a simple analysis of the application of matching estimators to stratified samples. They show the desirability of exact or “hard” 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 follow their advice in this paper and construct our estimates separately for each subgroup defined by the stratification variables – namely 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 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 (2005a) for more discussion of methods for doing so. 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 covariate, for the estimates reported here in which we do the matching separately for subgroups defined by the stratification variables. 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. (2005) for more details on this and Smith and Todd (2005b) for further discussion of balancing tests. We consistently find our preferred specification does a good job of balancing the benefit history variables.

We estimate our standard errors using bootstrapping methods with 300 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 (2005), who show the inconsistency of the bootstrap for nearest neighbor matching. Their Monte Carlo analysis suggests that while not zero, the inconsistency in the bootstrap will generally not lead to severely misleading inferences. We plan to pursue the alternative variance estimators in Abadie and Imbens (2004) and/or Politis, Romano and Wolf (1999) in future work, and in the meantime caution the reader to add one or two grains of salt (but not a whole shaker) to our standard errors.

5. Impact Estimates

Figure 1 presents the unadjusted fraction on benefit for NDLP participants and non-participants in our data. It illustrates that, without any adjustments, 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 stronger suggests that participants differ from non-participants in ways related to benefit receipt other than just NDLP participation. Our matching 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 matches 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. (2005) show that the propensity score specification employed in the Lessof et al. (2003) fails to balance the fractions receiving benefits among participants and matched non-participants in their 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, as in Lessof et al. (2003).² 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 in this analysis. That data overflow with information about past histories of benefit receipt, but lacks depth in terms of other variables, with the exception of basic variables such as the number and age of children, the age of the lone parent, and the geographic location of the family required for program administration.

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 a dummy variable for each quarter that indicates whether or not the individual spent at least half the period on benefit. We then concatenate the six dummies into 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: variables indicating the fraction of time on benefit in the pre-program period and a variable measuring the duration of the spell 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 period gets coded as 111000, while the same spell at the end of the pre-program period gets coded as 000111; a variable measuring time on benefit would give the same value to both. Second, relative to using the

² See Appendix C of Phillips et al. (2003) for the details of the National Center propensity score model.

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.

Two important decisions arise in implementing the benefit history strings in our context. 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 for that period's dummy variable in order to code them as a one. 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 10 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 starting a few 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, to defer an analysis of the dynamics of participation to future work.

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 table. 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 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.³ 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.

³ This analysis does not take account of the stratified sampling.

5.2 Exact matching on sampling stratum

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

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 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 of the lone parent, age and disability status of the lone parent (dummies for 10 five-year categories), 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 dummy variables, one for each of the non-empty benefit history strings defined in Section 5.1.⁴ 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 JSA receipt in progress as of June 1999. Recall that our data limit 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 dummy variables for benefit receipt in each of the six weeks prior to the start of the participation window.

⁴ All strings with fewer than 20 observations were pooled into a single category denoted “222222”. This combination includes 19 strings but only 68 observations.

Table 3 presents the estimates from the propensity score logit 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; results for the other strata are available from the authors upon request. These results show the high levels of joint significance relating to discrete variables associated with: the age the individual, the age of their youngest child, the number of children whereas variables relating to the benefit history are not significant in the determination of being a participant on NDLP. This is understandable for this particular stratum as the individuals in questions have only a very short claimant spell prior up to the NDLP eligibility window,

Figure 3 presents the fraction on benefit in each month from 1997 through 2004 following propensity score matching. Two patterns stand out in Figure 3. First, the propensity score matching does an impressive job of balancing pre-program benefit receipt between the participants and the non-participants. The impact estimates corresponding to the figure appear in the fifth row of Table 6. [JEFF: GIVE THE ESTIMATES AND DISCUSS THEM ONCE THE TABLE IS AVAILABLE].

6.0 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 do single nearest neighbor matching with replacement.

Table 4 presents the estimated mean impacts from this analysis. The first row presents the estimated mean of the weekly impacts on the fraction of time on benefits 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. The ATT for the whole post programme period is 19.42 whereas the impact for the stock is 20.45 and the flow is 14.24.. Secondly, the impacts fall over time. For the stock they fall from 22.49 at three months to 17.07 at 36 months, or by about 20 percent. For the flow they fall from 18.99 at three months to 13.70 at 36 months, or by about 25 percent. 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. The U.S. General Accounting Office (1996) shows that impacts for the U.S. Job Training Partnership Act (JTPA) in the U.S. have impacts that remain quite stable over time; Couch (1992) shows the same for the U.S. National Supported Work Demonstration. Dolton and O’Neill also find a sizeable impact of the Restart program in the UK over 4 years later. In contrast, Hotz, Imbens and Klerman (2000) show that the benefits from the work-first part of the California Greater Avenues to Independence (GAIN), fade out over time. Of the three programs, the services offered by the GAIN most closely resemble 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 “Hetero” gives the treatment effect for the pre-program period; with complete balance of the lagged outcomes this will equal zero. The third column, labeled “Diff” 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, the point estimates increase monotonically in child age, though not strongly so. Lone parents of older children may find it easier to leave home for work when encouraged to do so by NDLP. We find larger but less interpretable differences by the duration of the

benefit 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 4 and 5 show how those impacts play out over time and also graphically illustrate the exact matching on the benefit spell length. In term of time on IS, there is no clear pattern, other than the stock/flow differences already discussed. 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.

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, proceeds 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.

Our analysis in Dolton et al. (2005) examines the NCSR evaluation in great detail, and looks 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 attrition, at least as a function of observable 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) or Mueser, Troske and Gorislawsky (2005).

In contrast, three factors do matter in explaining the difference in estimates. First, flexible conditioning on detailed benefit receipt histories leads to substantially 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. Given the problems with the Lessof et al. (2003) analysis identified in Dolton et al. (2005), we prefer the estimates presented in this paper.

7.2 Comparison to experimental evaluations of US programs

As discussed in Section 5.3, our preferred impact estimate for the flow sample consists of an increase in the probability of not receiving benefits of 14.24 . In this section, we compare these surprisingly large 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), we have greater confidence in them because of 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. We look at impact estimates from the US not because we seek to further advance its cultural and intellectual hegemony but rather 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 if, for example, they get married and become a stay-at-home mom.

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 Aid to Families with Dependent Children (AFDC) or its successor Temporary Aid to Needy Families (TANF).

We begin by considering two expensive, intensive voluntary programs for populations including but 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 must less intensive treatment provided by NDLP. The NSWD provided its participants 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 percent. 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 experiments 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 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.⁵ 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, of course, 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 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 the four programs here, 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 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.

⁵ 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.

As we expect mandatory programs to have lower mean impacts than voluntary ones due to self-selection on impacts, these estimates 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 NSW or the Job Corps). As such, we cannot infer too much from the fact that they fall well below our estimates of the impact of NDLP.

Taken together, the evidence from the experimental literature in the US reinforces our view that our NDLP impact estimates seem too large. This could have several explanations: first, despite our flexible conditioning on the benefit histories some selection on observables may remain; 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 these evaluations and given the greater US experience with programs of this type.

8. Conclusions and Interpretations

Our evaluation of the NDLP using administrative data has yielded a rich harvest of substantive and methodological findings. Substantively, we find that NDLP had a large and 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 spells 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, both the historically low rates of employment among lone mothers in the UK and the lack of much effort to push lone parents on IS into employment in the past. Our more modest, but still substantial, estimated impacts for the remainder of the participant population 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. 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).

Moreover, we find that, used appropriately, the pre-program outcome histories embody the information in survey-based measures of attitudes and demographic characteristics. We suggest that our findings will generalize to other contexts which means that evaluations conducting using only administrative data will have no more bias than survey-based evaluations that rely on information like that collected for the Lessof et al. (2003) evaluation. At the same time, administrative data, though not a panacea, avoid issues of non-response and typically cost a lot less to use. Our analysis thus indirectly highlights 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 analysis shows the value of using a dependent variable that takes into account the potential for return to benefit receipt and the importance of taking account of stratified sampling in applying matching estimators.

Bibliography

Abadie, Alberto and Guido Imbens. 2004. "Large Sample Properties of Matching Estimators for Average Treatment Effects." Unpublished Manuscript, Harvard University.

Abadie, Alberto and Guido Imbens. 2005. "On the Failure of the Bootstrap for Matching Estimators." Unpublished Manuscript, Harvard University.

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.

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.

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. 2005. The Econometric Evaluation of the NDLP. Report prepared for the UK Department of Work and Pensions.

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, Jeffrey Smith and João Pedro Azevedo. 2006. "The Application of Matching Estimators with Stratified Sampling." Unpublished Manuscript, University of Michigan.

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

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.

Fisher, Ronald. 1935. *The Design of Experiments*. London: Oliver and Boyd.

Fredriksson, P. and P. Johansson. 2003. "Program evaluation and random program starts". Working Paper No. 2003:1. Institute for Labour Market Policy Evaluation, Uppsala, Sweden.

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

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

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

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.

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. 1998 "Characterising Selection Bias Using Experimental Data." *Econometrica* 66(5): 1017-1098.

Heckman, James, Lance Lochner, and Christopher Taber. 1998. "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. 2005. "Dynamic Discrete Choice and Dynamic Treatment Effects." NBER Technical Working Paper No. 316.

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

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.

Hotz, Joseph, Guido Imbens and Jacob Klerman. 2000. "The Long-Term Gains from GAIN: A Re-Analysis of the Impacts of the California GAIN Program." NBER Working Paper No. 8007.

Hotz, 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.

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.

Knight, Genevieve, Peter Dolton, Jeffrey Smith, Stefan Speckesser, Diana Kasparova, and Pedro Azevedo. 2006. *The Evaluation of Combined Lone Parents Work Focused Interviews and New Deal for Lone Parents, the In-Work Benefit Calculation and Further Re-Analyses of the New Deal for Lone Parents*. London: Policy Studies Institute.

LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4): 604-620.

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.

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 at <http://ideas.repec.org/c/boc/bocode/s432001.html>.

Lise, Jeremy, Shannon Seitz and Jeffrey Smith. 2005. "Equilibrium Policy Experiments and the Evaluation of Social Programs." Unpublished Manuscript, University of Michigan.

Mueser, Peter, Kenneth Troske and Alexey Gorislavsky. 2005. "Using State Administrative Data to Measure Program Performance." Unpublished Manuscript, University of Kentucky.

Neyman, Jerzy. 1923. "Statistical Problems in Agricultural Experiments." *Journal of the Royal Statistical Association*. 2(2): 107-180.

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.

Politis, Demitris, Joseph Romano and Michael Wolf. 1999. *Subsampling*. Springer.

Quandt, Richard. 1972. "Methods of Estimating Switching Regressions." *Journal of the American Statistical Association* 67(338): 306-310.

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.

Roy, A. D. 1951. "Some Thoughts on the Distribution of Earnings." *Oxford Economic Papers* 3(2): 135-146.

Rubin, Donald. 1974. "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies." *Journal of Educational Psychology* 66: 688-701.

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.

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

Table .1Sample Composition.

Strata	Age of youngest Child (years)	IS Spell Duration (months)	Total Eligible Population			NDLP Participants in Sample			Total Sample			Sample Rate
			Wave 1/2	Booster	Total	Wave 1/2	Booster	Total	Wave 1/2	Booster	Total	
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	0.310
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

<i>Pre programme History</i>	<i>Number of non-participants</i>	<i>Proportion of non-participants on Benefit (%)</i>	<i>Number of participants</i>	<i>Proportion of participants on Benefit (%)</i>	<i>Differences (D)-(B)</i>	<i>Proportion of participants on each stratum</i>	<i>Cell Specific Treatment (E)x(F)</i>	<i>Proportion of participants on each stratum (exception of Stratum 111111)</i>	<i>Cell Specific Treatment (E)x(G)</i>
	(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 a dummy 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 dummies 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 Model forStratum 1

	<i>Coef/S.E.</i>	<i>Mean</i>	<i>Std Dev</i>
disabled	-0.901 (0.570)	0.030	0.171
disalen	-0.329* (0.165)	0.157	0.907
W39_2000	0.316 (0.278)	0.970	0.171
W38_2000	-0.216 (0.218)	0.936	0.246
W37_2000	0.275 (0.190)	0.883	0.322
W36_2000	-0.034 (0.159)	0.822	0.382
W35_2000	0.076 (0.141)	0.740	0.439
W34_2000	-0.206 (0.142)	0.641	0.480
ben_proprio	-0.217 (0.176)	0.129	0.280
<i>(categorical variables omitted – available upon request)</i>			
Joint significance of categorical variables	Chi2/ P-val		
Age	13.691 0.057		
Region	10.787 0.461		
AgeYc	10.071 0.018		
NumberChildren	27.731 0.000		
Benefit	22.186 0.877		
Obs	3788		
R-squared	0.061		
LogLike	-857.023		
Chi2	112.207		

Note: (1) Probit estimation;

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

<i>Time Period</i>	<i>ATT</i>	<i>Stock</i>	<i>Flow</i>
All post-programme	19.42 (0.17)	20.45 (0.19)	14.24 (0.21)
3 months on benefit	21.91 (0.1)	22.49 (0.11)	18.99 (0.17)
9 months on benefit	22.28 (0.16)	23.72 (0.16)	15.04 (0.22)
24 months on benefit	18.3 (0.05)	19.13 (0.08)	14.09 (0.2)
36 months on benefit	16.51 (0.15)	17.07 (0.10)	13.7 (0.51)

Estimated Standard errors in parathesis

Note:

(1) Full sample; (2) 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; (3) the time periods refer to points in time after the end of the participation window; (4) the analysis is done separately by stratum and then weighted to take account of the stratified sampling.

Table 5 Estimating Treatment Effects for Subgroups

<i>Description</i>	<i>Treatment</i>	<i>Hetero</i>	<i>Diff</i>
Children at the age of [0,3)	16.96 (0.25)	-0.97 (0.06)	17.93 (1.16)
Children at the age of [3,5)	18.62 (0.16)	0.11 (0.08)	18.51 (0.54)
Children at the age of [5,11)	20.38 (0.17)	-0.04 (0.06)	20.42 (0.59)
Children at the age of [11,16)	21.3 (0.25)	-1.12 (0.14)	22.42 (1.44)
On IS for less than 3 months	16.14 (0.19)	-0.54 (0.08)	16.68 (0.73)
On IS from [3,6) months	13.34 (0.29)	-3.11 (0.21)	16.45 (2.24)
On IS from [6,12) months	20.97 (0.2)	-1.11 (0.19)	22.08 (1.37)
On IS from [12,24) months	15.8 (0.45)	0.26 (0.16)	15.55 (3.96)
On IS from [24,36) months	18.17 (0.13)	-0.46 (0.08)	18.64 (0.42)
On IS for more than 36 months	25.72 (0.21)	0.03 (0.01)	25.69 (0.80)

Note:

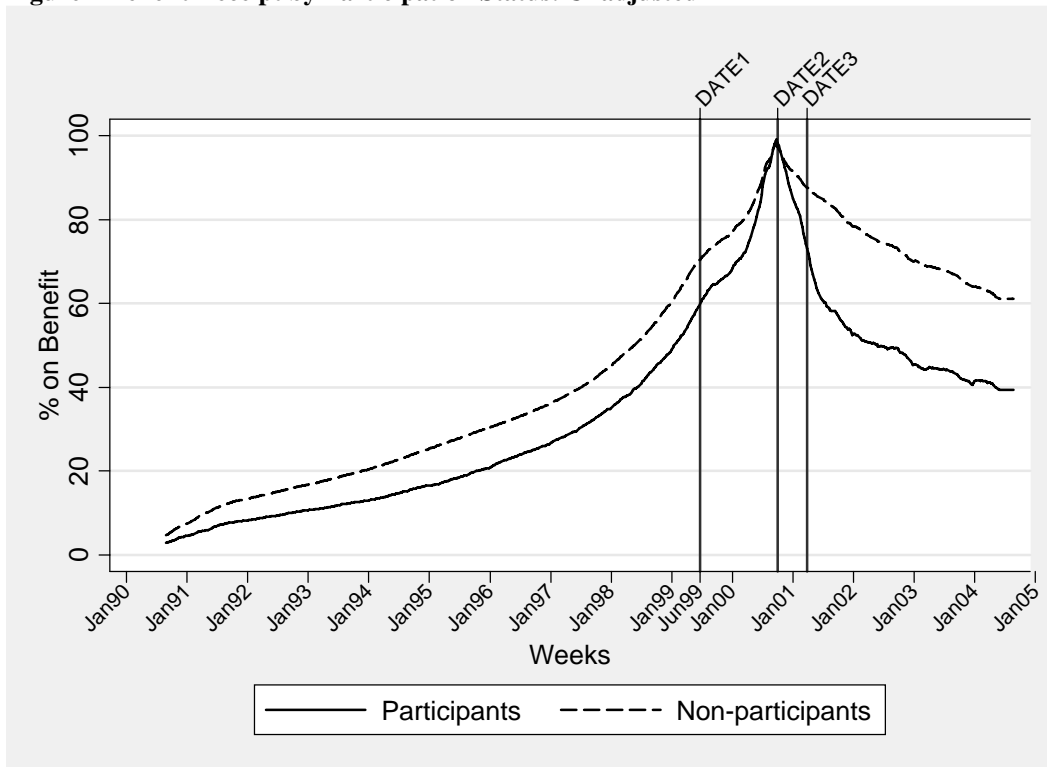
(1) Estimated Standard errors in parenthesis; (2) the impact estimates refer to the entire post-programme period; (3) The subgroups are define according to the administrative database used to draw the initial NatCen sample; (4) the age of the children refers to the youngest child of the household as of August/2000; (5) the analysis is done separately by stratum and then weighted to take account of the stratified sampling.

Table A-1. Variables used in the propensity score

<i>Variable Name</i>	<i>Description</i>
_lsex_2	sex==2
_lageband_2	ageband==2
_lageband_3	ageband==3
_lageband_4	ageband==4
_lageband_5	ageband==5
_lageband_6	ageband==6
_lageband_7	ageband==7
_lageband_8	ageband==8
_lageband_9	ageband==9
_lageband_10	ageband==10
_lgor_2	gor==2
_lgor_3	gor==3
_lgor_4	gor==4
_lgor_5	gor==5
_lgor_6	gor==6
_lgor_7	gor==7
_lgor_8	gor==8
_lgor_9	gor==9
_lgor_10	gor==10
_lgor_11	gor==11
_lgor_12	gor==12
_lpcageyc_2	pcageyc==2
_lpcageyc_3	pcageyc==3
_lpcageyc_4	pcageyc==4
_lpcageyc_5	pcageyc==5
_lpcageyc_6	pcageyc==6
_lpcnumbch_3	pcnumbch==3
_lpcnumbch_4	pcnumbch==4
_lpcnumbch_5	pcnumbch==5
_lben_prehi_1	ben_prehist3==1
_lben_prehi_10	ben_prehist3==10
_lben_prehi_11	ben_prehist3==11
_lben_prehi_100	ben_prehist3==100
_lben_prehi_101	ben_prehist3==101
_lben_prehi_110	ben_prehist3==110
_lben_prehi_111	ben_prehist3==111
_lben_prehi_1000	ben_prehist3==1000
_lben_prehi_1001	ben_prehist3==1001
_lben_prehi_1011	ben_prehist3==1011
_lben_prehi_1100	ben_prehist3==1100
_lben_prehi_1101	ben_prehist3==1101
_lben_prehi_1111	ben_prehist3==1111
_lben_prehi_10000	ben_prehist3==10000
_lben_prehi_10001	ben_prehist3==10001
_lben_prehi_10011	ben_prehist3==10011
_lben_prehi_10110	ben_prehist3==10110
_lben_prehi_10111	ben_prehist3==10111
_lben_prehi_11000	ben_prehist3==11000

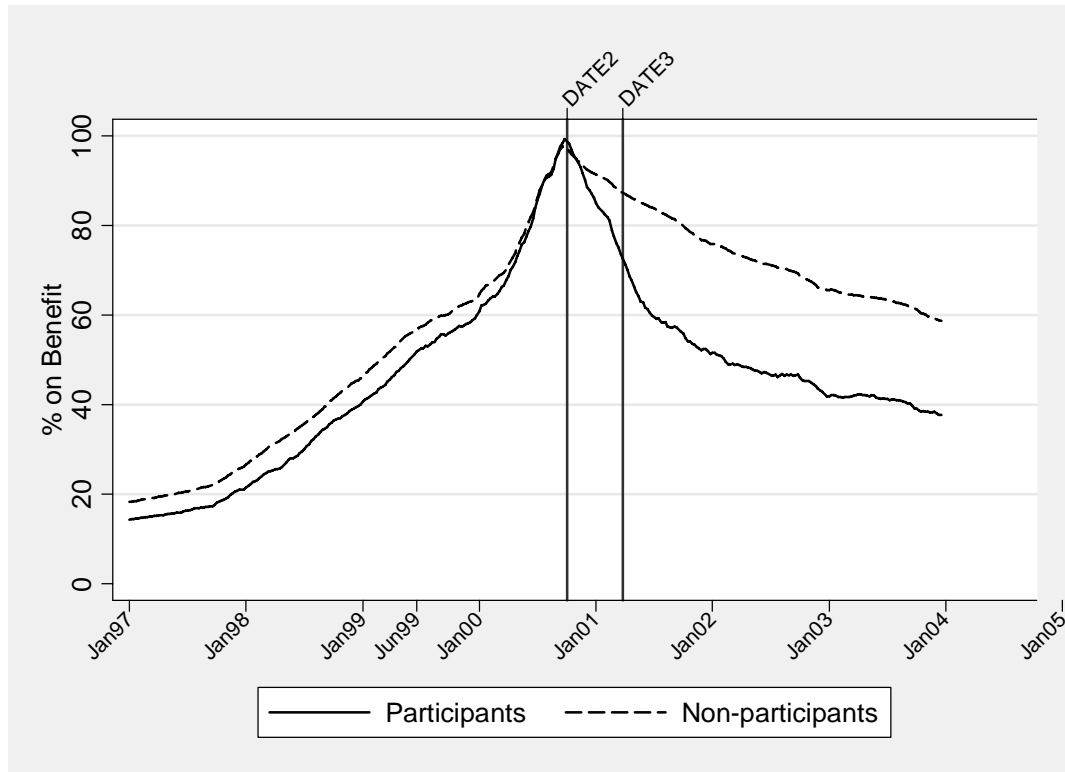
_lben_prehi_11001	ben_prehist3==11001
_lben_prehi_11011	ben_prehist3==11011
_lben_prehi_11100	ben_prehist3==11100
_lben_prehi_11101	ben_prehist3==11101
_lben_prehi_11111	ben_prehist3==11111
_lben_prehi_100000	ben_prehist3==100000
_lben_prehi_100001	ben_prehist3==100001
_lben_prehi_100011	ben_prehist3==100011
_lben_prehi_100101	ben_prehist3==100101
_lben_prehi_100111	ben_prehist3==100111
_lben_prehi_101011	ben_prehist3==101011
_lben_prehi_101101	ben_prehist3==101101
_lben_prehi_101111	ben_prehist3==101111
_lben_prehi_110000	ben_prehist3==110000
_lben_prehi_110001	ben_prehist3==110001
_lben_prehi_110011	ben_prehist3==110011
_lben_prehi_110111	ben_prehist3==110111
_lben_prehi_111000	ben_prehist3==111000
_lben_prehi_111001	ben_prehist3==111001
_lben_prehi_111011	ben_prehist3==111011
_lben_prehi_111100	ben_prehist3==111100
_lben_prehi_111101	ben_prehist3==111101
_lben_prehi_111110	ben_prehist3==111110
_lben_prehi_111111	ben_prehist3==111111
_lben_prehi_222222	ben_prehist3==222222
disab_pr	Disabled lone parent
disalen	Length of the lone parent disability
ben_a2118	On Benefit at W39_2000
ben_a2117	On Benefit at W38_2000
ben_a2116	On Benefit at W37_2000
ben_a2115	On Benefit at W36_2000
ben_a2114	On Benefit at W35_2000
ben_a2113	On Benefit at W34_2000
ben_proprio	Proportion of time on Benefit prior to June/1999

Figure 1 Benefit Receipt by Participation Status: Unadjusted



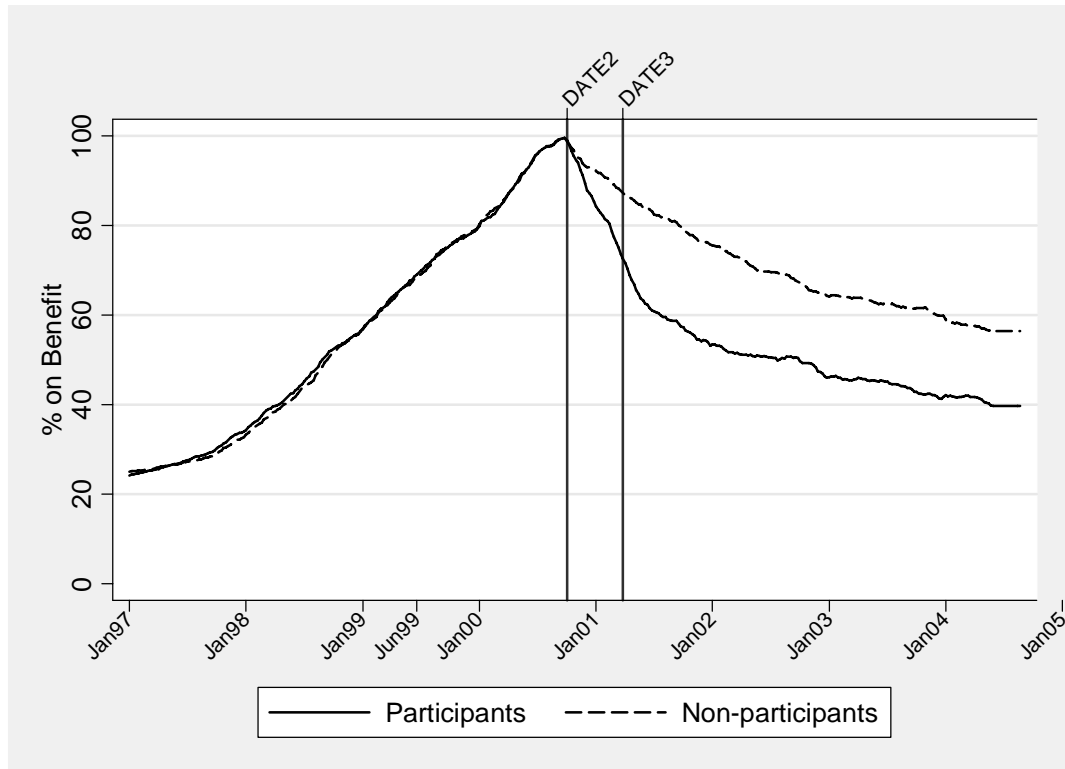
Note: (1) Date2 and Data3 are the start and end date of the participation window, respectively; (2) this estimation does not take into account the sample design.

Figure 2 Benefit Receipt by Participation Status: Matching on Strata



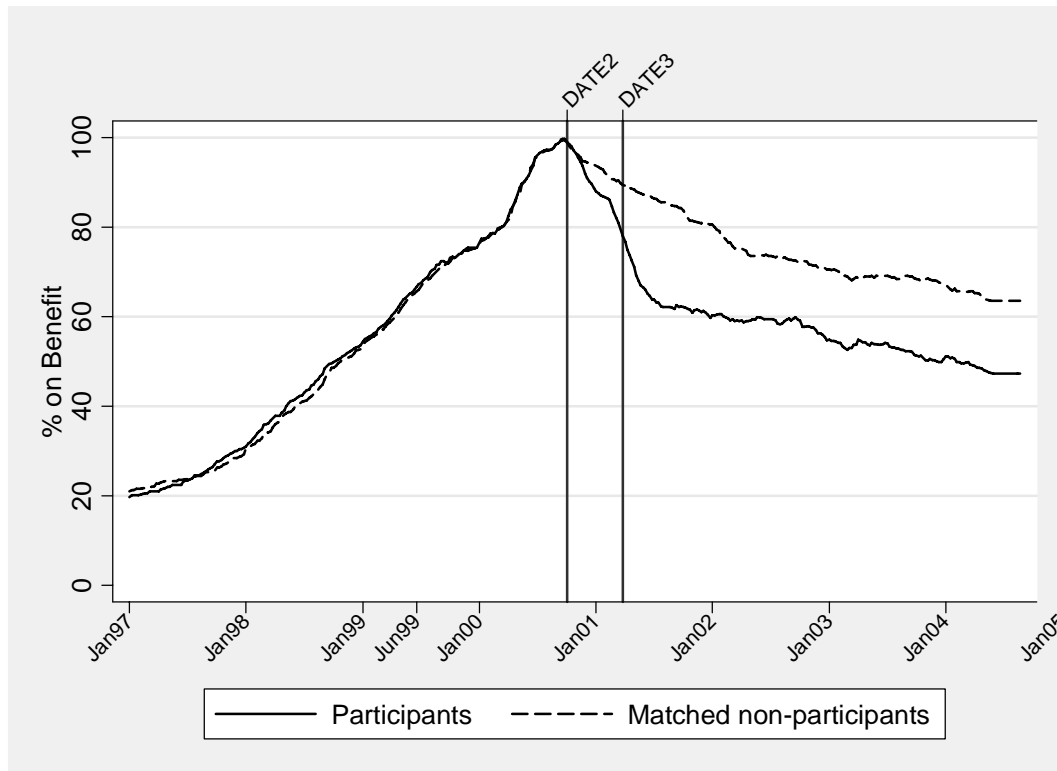
Note: (1) Date2 and Data3 are the start and end date of the participation window, respectively; (2) this estimation does take into account the sample design.

Figure 3 Benefit Receipt by Participation Status: Propensity Score Matching



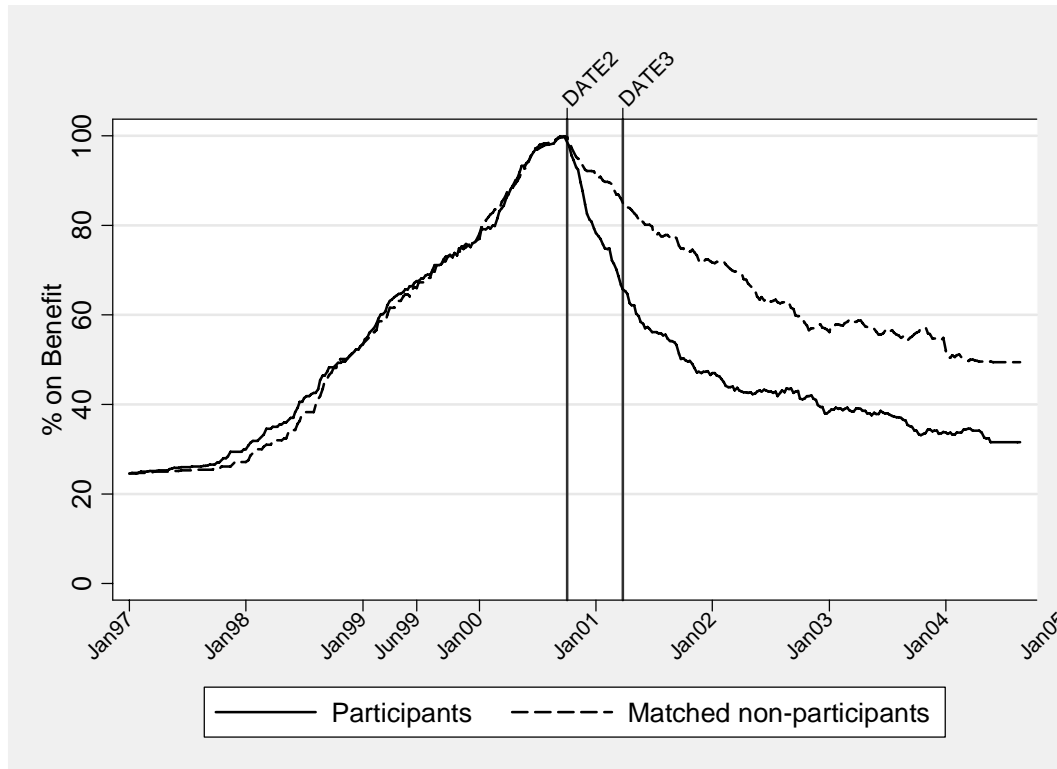
Note: (1) Date2 and Data3 are the start and end date of the participation window, respectively; (2) this estimation does take into account the sample design “hardmatching” by stratum; (3) nearest neighbor propensity score matching with replacement and common support.

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



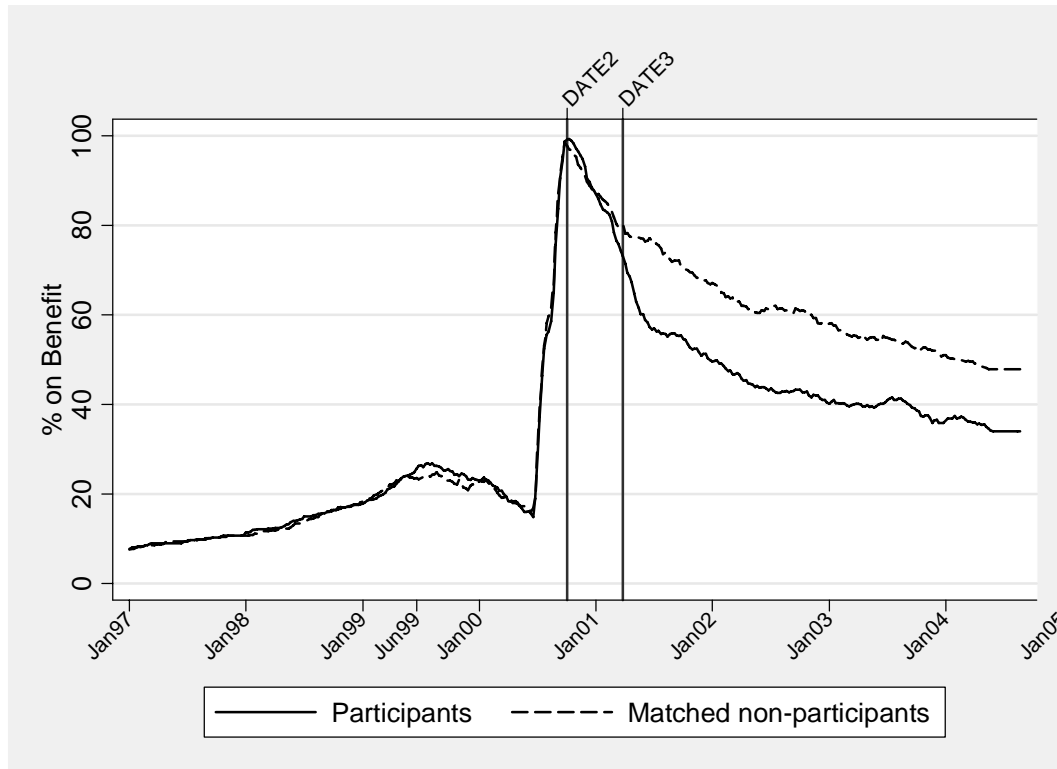
Note: (1) Date2 and Data3 are the start and end date of the participation window, respectively; (2) this estimation does take into account the sample design “hardmatching” by stratum; (3) nearest neighbor propensity score matching with replacement and common support.

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



Note: (1) Date2 and Data3 are the start and end date of the participation window, respectively; (2) this estimation does take into account the sample design “hardmatching” by stratum; (3) nearest neighbor propensity score matching with replacement and common support.

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



Note: (1) Date2 and Data3 are the start and end date of the participation window, respectively; (2) this estimation does take into account the sample design “hardmatching” by stratum; (3) nearest neighbor propensity score matching with replacement and common support.