

Department for Work and Pensions

Research Report No 356

The econometric evaluation of the New Deal for Lone Parents

Professor Peter Dolton, João Pedro Azevedo and Professor Jeffrey Smith

A report of research carried out by the Department of Economics, Royal Holloway, University of London, the Institute of Applied Economic Research, Brazil and the Department of Economics, University of Michigan on behalf of the Department for Work and Pensions

Corporate Document Services

© Crown Copyright 2006. Published for the Department for Work and Pensions under licence from the Controller of Her Majesty's Stationery Office by Corporate Document Services, Leeds.

Application for reproduction should be made in writing to The Copyright Unit, Her Majesty's Stationery Office, St Clements House, 2-16 Colegate, Norwich NR3 1BQ.

First Published 2006.

ISBN 1 84712 035 0

ISBN13 978 1 84712 035 9

Views expressed in this report are not necessarily those of the Department for Work and Pensions or any other Government Department.

Printed by Corporate Document Services.

Contents

Acknowledgements	xi
The Authors	xii
Abbreviations and acronyms	xiii
Samples	xv
Notation	xvii
Summary	1
1 Introduction: The evaluation of the New Deal for Lone Parents	7
Summary	7
1.1 Introduction	8
1.2 The New Deal for Lone Parents	9
1.3 The NatCen report	11
1.4 The policy context	13
1.5 Aims and scope of this evaluation	14
2 Sample design and sampling issues	17
Summary	17
2.1 Introduction	17
2.2 Understanding the sample design	18
2.3 Sampling	19
2.3.1 <i>The overall results</i>	24
2.3 The sampling method	24
2.4.1 <i>Select a sample of eligible non-participants</i>	24
2.4.2 <i>Carry out a postal survey of the sample of eligible non-participants</i>	25
2.4.3 <i>Using administrative data to identify individuals entering NDLP</i>	25

2.5	The administrative datasets	30
2.5.1	<i>NDLP dataset</i>	31
2.5.2	<i>IS dataset</i>	31
2.5.3	<i>IB dataset</i>	32
2.5.4	<i>JSA dataset</i>	32
2.6	Overlapping claims	32
3	Basic notation, the matching methodology and the NatCen model	37
	Summary	37
4	Evaluating the NatCen report	39
	Summary	39
5	The problems of non-response, sample weighting and the window of participation	41
	Summary	41
5.1	Introduction	42
5.2	Non-response and weighting issues	42
5.2.1	<i>The NatCen weighting procedure and calculating new weights for the strata</i>	42
5.2.2	<i>Non-response by NDLP participant status</i>	45
5.2.3	<i>Multivariate analysis of non-response</i>	45
5.3	Impact estimates and non-response	46
5.3.1	<i>Description of the weighting procedure</i>	46
5.3.2	<i>Weighted outcomes</i>	46
5.3.3	<i>The window problem</i>	47
5.4	Using supplementary exogenous conditioning data relating to local labour markets	50
6	Alternative impact estimations	51
	Summary	51
6.1	Introduction	52
6.2	Impacts measured over time	52
6.2.1	<i>Symmetric differences</i>	60
6.2.2	<i>Exact matching on IS receipt histories</i>	63
6.2.3	<i>New propensity score specifications</i>	65
6.2.4	<i>Flow and stock: heterogeneous treatment effects</i>	68
6.2.5	<i>Impact estimates for different subgroups</i>	70
6.3	Graphs of expected outcomes conditional on the propensity score	74
6.3.1	<i>Motivation</i>	74
6.3.2	<i>Conceptual framework</i>	75
6.3.3	<i>Evidence</i>	76

7	Further econometric problems and issues	79
	Summary	79
7.1	Introduction	79
7.2	Matching and stratified sampling	80
7.3	The duration of NDLP participation	84
7.3.1	<i>Conceptual issues related to treatment duration</i>	84
7.3.2	<i>Evidence on the duration of NDLP participation</i>	86
7.3.3	<i>Understanding the heterogeneity in the duration of participation</i>	86
7.3.4	<i>Implications for the analyses in this report</i>	87
7.4	NDLP participation starts at different times	91
7.4.1	<i>Dynamic participation: conceptual issues</i>	91
7.4.2	<i>Evidence of later participation by non-participants</i>	93
7.4.3	<i>A simple sensitivity analysis</i>	94
7.4.4	<i>Implications of dynamic participation for the analysis</i>	94
7.4.5	<i>Changes in NDLP eligibility over time</i>	95
7.5	Multiple spells of NDLP participation	96
7.5.1	<i>Conceptual issues</i>	96
7.5.2	<i>Evidence on multiple spells of NDLP participation</i>	97
7.5.3	<i>Implications of multiple spells of NDLP participation</i>	97
7.6	General equilibrium effects	98
7.6.1	<i>Conceptual issues</i>	98
7.6.2	<i>Implications for the analyses in this report</i>	99
7.7	Effects of other contemporaneous ALMP	99
8	Interpretations and conclusions	101
	Summary	101
8.1	Lessons for economic policy evaluation methodology	105
8.1.1	<i>Use of the administrative data</i>	105
8.2	Future research evaluation of the impact of NDLP	106
8.2.1	<i>Prescriptive use of matching to target potential interviews</i>	106
8.2.2	<i>More careful assessment of the duration nature of the outcome variables in question</i>	106
8.2.3	<i>Long run analysis of the labour market impact of NDLP</i>	107
8.2.4	<i>Cost-benefit analysis of the NDLP programme</i>	107

Appendix A	Logit: Dependent variable NDLP participation using SPSS and STATA	109
Appendix B	Variable descriptions	113
Appendix C	Balancing tests for samples with estimated propensity scores of less than 0.3	115
Appendix D	Balancing tests for NatCen full sample and face-to-face interview respondent sample	123
Appendix E	Summary statistics (mean and standard deviation)	127
Appendix F	Logit coefficients using the administrative data (odds ratio)	135
Appendix G	Model summary statistics by strata	141
Appendix H	Technical annex	143
Appendix I	Basic notation, the matching methodology and the NatCen model	147
Appendix J	Evaluating the NatCen report	159
References		211

List of tables

Table 2.1	Composition of the sample	27
Table 2.2	Selection and response rates for the two key sample sub-groups	28
Table 6.1	Stock/flow heterogeneous average treatment effect on the treated	69
Table 6.2	Subgroups total average treatment effect on the treated	72
Table A.1	Logit: Dependent variable NDLP participation using SPSS and STATA	109
Table B.1	Variable descriptions	113
Table C.1	Balancing tests for samples with estimated propensity scores of less than 0.3	116
Table D.1	Balancing tests for NatCen full sample and face-to-face interview respondent sample	123
Table E.1	Summary statistics (mean and standard deviation)	127
Table F.1	Logit coefficients using the administrative data (odds-ratio)	135
Table G.1	Model summary statistics by strata	141
Table H.1	Trends in average numbers of new entrants to NDLP per month averaged by year: 1998 to 2003	143
Table H.2	Different sample definitions	144
Table H.3	Summary of data collection process	145
Table J.1	Random sample of the propensity scores estimated by NatCen and the authors	160
Table J.2	Six-tiles of the estimated propensity scores	165

Table J.3	Complete pairs of matched face-to-face interview respondents	166
Table J.4	Postal survey respondent sample	167
Table J.5	Matching without replacement in the full sample	167
Table J.6	Matching without replacement in the face-to-face interview sample	167
Table J.7	Number of times non-participant observations used in matching with replacement: full postal survey respondent sample	169
Table J.8	Number of times non-participant observations used in matching with replacement: postal survey respondent sample with estimated propensity scores < 0.3	169
Table J.9	Extract of Balancing Tests for Samples with estimated propensity scores < 0.3	171
Table J.10	Extract of balancing tests for NatCen full sample and face-to-face interview respondent sample	172
Table J.11	Outcome definitions	173
Table J.12	Distribution of NDLP participants and total eligible population by claim duration and child age by Wave 1/ Wave 2 or booster sample	177
Table J.13	Distribution of NDLP participants and total sample by duration of claim and age of youngest child by Wave 1/ Wave 2 or booster sample	179
Table J.14	Non-response to postal questionnaire by participation in NDLP	180
Table J.15	Non-response to face-to-face interview survey by participation in NDLP	181
Table J.16	Sample category – age of youngest child – benefit duration	181
Table J.17	Variables used and respective values	182
Table J.18	Multivariate probit models of non-response to both surveys	184
Table J.19	Relative frequencies of the population, sample and NatCen weights.....	187
Table J.20	Impact estimates from matching adjusted for non-response	188
Table J.21	Impact estimates for different response windows: full postal survey respondent sample	189
Table J.22	Impact estimates for different response windows: full face-to-face interview respondent sample	190
Table J.23	Impact estimates for different response windows: postal survey respondent sample with estimated propensity scores < 0.3	191

Table J.24	Impact estimates for different response windows: face-to-face interview respondent sample with estimated propensity scores < 0.3	192
Table J.25	Outcome definitions	193
Table J.26	Impacts from alternative single nearest neighbour matching with replacement using LAD variables	194
Table J.27	Total average treatment effect on the treated according to different samples	195
Table J.28	Card and Sullivan cell match (A, B, C, D and E)	196
Table J.29	Card and Sullivan cell match (B and C)	196
Table J.30	Card and Sullivan cell match except 111111 (A, B, C, D and E)	197
Table J.31	Card and Sullivan cell match except 111111 (B and C)	197
Table J.32	P-value of the joint significance of selected categories on the propensity score estimation	198
Table J.33	Total average treatment effect	198
Table J.34	Proportion of lone parents by age of the youngest child and length of time on IS (%)	198
Table J.35	Subgroups total average treatment effect	199
Table J.36	NDLP claims according study window and participant status	200
Table J.37	NDLP recidivism after 1 April 2001, by participation status	200
Table J.38	NDLP recidivism after 23 October 1998, by participation status	201

List of figures

Figure 2.1	Data collection time line	22
Figure 2.2	Overlapping spells	34
Figure 2.3	Second example of overlapping spells	34
Figure 2.4	Third example of overlapping spell	35
Figure 5.1	Distribution of participants, by participation date	48
Figure 5.2	Distribution of participants according to the number of days between NDLP participation and postal questionnaire response	49
Figure 6.1	Weekly proportion of participants and non-participants on benefit – 1990-2003 (Sample I – Full)	55
Figure 6.2	Weekly proportion of participants and non-participants on benefit – 1997-2003 (Sample I – Full)	55
Figure 6.3	Weekly proportion of participants and non-participants on benefit – 1997-2003 (Sample III – Matched)	57
Figure 6.4	Weekly proportion of participants on benefit 1999-2002 (Sample I – Full)	61

Figure 6.5	Symmetric estimator of the weekly proportion of participants on benefit 1999-2002 (Sample I –Full)	62
Figure 6.6	Weekly proportion of participants and non-participants on benefit – 1997-2003 (Sample I – Specification 3)	68
Figure 6.7	Weekly proportion of participants and non-participants on benefit – 1997-2003 (Sample I – Specification 3). Child between 0 and 3 years old	73
Figure 6.8	Weekly proportion of participants and non-participants on benefit – 1997-2003 (Sample I – Specification 3). Child between 11 and 16 years old	74
Figure 6.9	Participants and non-participants LOWESS non-parametric regression (Sample I –Full)	76
Figure 7.1	Duration of NDLP spell in weeks (Groups A, B and D)	88
Figure 7.2	Duration of NDLP spell in weeks (Group B)	88
Figure 7.3	Proportion of non-participants at April 2001 who subsequently became NDLP participants (Groups C and E)	89
Figure 7.4	Cumulative Proportion of Non-Participants at April 2001 who subsequently became NDLP participants (Group C and E)	89
Figure 7.5	Proportion of Participants and Non-Participants on Benefit 1997-2003 (Sample I) Only if Youngest child age ≥ 5 and ≤ 15)	90
Figure 7.6	Proportion of Participants and Non-Participants on Benefit 1997-2003 (Sample I)C&S, Recent and Long Hist.)	90
Figure H.1	The number of NDLP participants – October 1998 to December 2003	145
Figure H.2	The number of NDLP new entrants – October 1998 to December 2003	146
Figure J.1	Histogram: Propensity score distribution for the face-to-face interviewed sample (Groups B1 and C1)	162
Figure J.2	Histogram: Propensity score distribution for the matched sample (Groups B1, B2, C1 and C2).....	162
Figure J.3	Histogram: Propensity score distribution for the full sample (Groups B and C)	163
Figure J.4	Box-Cox Plot of propensity scores for full sample, by participation status	164
Figure J.5	Box-Cox plots of propensity scores for hex-tiles, by participation status	164
Figure J.6	Symmetric estimator of the weekly proportion of participants on benefit 1999-2002 (Sample II – postal)	201
Figure J.7	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample II – postal)	202

Figure J.8	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – full – weighted by Strata).....	202
Figure J.9	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 1)	203
Figure J.10	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 2)	203
Figure J.11	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3 – Group B only)	204
Figure J.12	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample II – NatCen specification + ps w/ replacement)	204
Figure J.13	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample II – NatCen specification + pre-programme + ps w/ replacement)	205
Figure J.14	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample V – face-to-face)	205
Figure J.15	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) child at between 3 and 5 years old	206
Figure J.16	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) child at between 5 and 11 years old	206
Figure J.17	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) less than 3 months on IS	207
Figure J.18	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) from 3 to 6 months on IS	207
Figure J.19	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) from 6 to 12 months on IS	208
Figure J.20	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) from 12 to 24 months on IS	208
Figure J.21	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) from 24 to 36 months on IS	209
Figure J.22	Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) for over 36 months on IS	209

Acknowledgements

We would like to thank colleagues at the Department for Work and Pensions for their valuable guidance and comments, in particular:

Mike Daly for his critical wisdom throughout; Ellenor Brooks for her sterling editorial work and immense patience.

Karen Grierson, Jessica Vince and Graeme Connor for their provision of the data and their help and advice in understanding its complexities, and Darek Olesinski, Tim Foster and Liz McKeown for the project management.

We wish to thank participants at the one day NDLP conference on the 25 August 2005 especially Joe Hotz, Robert Moffitt, Jonathan Portes, Jeff Grogger, Susan Harkness, Steve Machin and Paul Gregg and our colleagues at PSI especially Genevieve Knight, Stephan Speckasser, Alex Bryson and Richard Dorsett. In addition, we would like to thank the NatCen team, Carli Lessof, Kevin Pickering, Melissa Miller, Miranda Phillips, Susan Purdon and Jon Hales, for the collection of the survey data and their help in understanding their initial analysis.

The Authors

Professor Peter Dolton, Department of Economics, Royal Holloway, University of London.

João Pedro Azevedo, Institute of Applied Economic Research (IPEA), Brazil.

Professor Jeffrey Smith, Department of Economics, University of Michigan.

Abbreviations and acronyms

ALMP	Active Labour Market Programme
ATT	Average Treatment on the Treated
BA	Benefit Agency
CBA	Cost Benefit Analysis
CIA	Conditional Independence Assumption
DWP	Department for Work and Pensions
ES	Employment Service
GMS	Generalised Matching Service
IB	Incapacity Benefit
IS	Income Support
JSA	Jobseeker's Allowance
JTPA	Job Training Partnership Act
LAD	Local Authority District
LFS	Labour Force Survey
LPPA	Lone Parent Personal Adviser
LMS	Labour Market System
NatCen	National Centre for Social Research
NDED	New Deal Evaluation Databases
NDLP	New Deal for Lone Parents
NOMIS	National Online Manpower Information System

PA	Personal Advisor
PSU	Primary Sampling Units
QLFS	Quarterly Labour Force Survey
SDA	Severe Disablement Allowance
TTWA	Travel to Work Area
WFI	Work Focused Interview
WFTC	Working Families' Tax Credit

Samples

Group A	Individuals who participated in the New Deal for Lone Parents (NDLP) and then returned their postal questionnaire.
Group B	Individuals who returned their postal questionnaire and then participated in NDLP.
Group C	Non-participants who returned their postal questionnaire.
Group D	Participants who did not return their postal questionnaire.
Group E	Non-participants who did not return their postal questionnaire.
Group B1	Individuals who returned their postal questionnaire and then participated in NDLP, and who responded to the face-to-face survey.
Group B2	Individuals who returned their postal questionnaire and then participated in NDLP, and who did not respond to the face-to-face survey.
Group C1	Matched non-participants who responded to the face-to-face survey (and responded the postal questionnaire).
Group C2	Matched non-participants who did not respond to the face-to-face survey (and responded to the postal questionnaire).
Group C3	Non-participants who were not matched to participants (and did not respond to the postal questionnaire).

Notation

M	Matching.
PS	Propensity Score.
X_i^A	Administrative data containing information of-economic characteristics of each lone parent eligible to enter the NDLP.
Y_{it}	Vector of outcome variables for individual i at time t .
N^A	Set of lone parents with complete records in the administrative data.
t	Time.
i	Lone parent eligible to enter NDLP.
$t = q$	Date all of the lone parents in the administrative sample received a postal questionnaire.
$Q \in \{0,1\}$	Binary variable indicating response to the postal survey.
X^Q	Vector of characteristics collected from the postal survey.
N^Q	Set of respondents to the postal survey.
N^P	Set of individuals who choose to participate in NDLP.
$D \in \{0,1\}$	Indicates participation in the NDLP.
N^M	The set of individuals who get matched.
$t = f$	Time NatCen attempts to complete a face-to-face interview with each member of N^M .
$t = p$	Denotes the time of NDLP participation.
N^F	Denotes the subset of completing a face-to-face interview.

xviii	Notation
X^F	Denotes the vector of variables (other than outcomes) collected in the face-to-face interview.
$F \in \{0,1\}$	Binary variable indicating completion of the face-to-face interview.
S^Q	Self-reported outcomes collected at one point in time from the postal questionnaire.
S^F	Self-reported outcomes collected at one point in time from the face-to-face interview.
Z	A vector with of various economic and social characteristics of the local area.
L	Lag operator.
W	Vector of observable pre-treatment variables which may include (X, Z, Y).
ω	Weights.
π	Probability of response.
θ	$\frac{1}{\pi}$

Summary

Introduction

The objective of this report is to explore how earlier impact estimates, or additionality, of the New Deal for Lone Parents (NDLP) conducted by the National Centre of Social Research (NatCen) were obtained, to determine the sensitivity of their estimates to changes in the estimation procedure and outcome variables, to look at longer-term outcomes and to provide an insight into the particular econometric difficulties imposed by the evaluation of the NDLP programme.

This report also attempts to assess, in a non-experimental way, what the impact of NDLP is. This is done by using statistical matching techniques to try to compare those who participated in NDLP with those who 'looked most like them' but did not participate in NDLP. The problem in trying to do this is that not just those who received the NDLP treatment at a single point in time are observed, but, using administrative data, it is possible to include those who received NDLP every week before and after NDLP eligibility.

Background

In the last eight to ten years the position of lone parents in the UK has changed markedly. This is due to the commitment of the Government to tackle child poverty. As a result the benefits available to lone parents have considerably increased over the 1997 to 2004 period. This has, in large measure, been due to the tax and benefit changes which included the introduction of the Working Families' Tax Credit in 1998 (now Working Tax Credit), the introduction of New Deal for Lone Parents, the National Childcare Strategy and the National Minimum Wage.

Compared to other countries, the lone parent employment rate in the UK is low and the gap between employment rates of lone parents and those of married mothers is higher in the UK than nearly all comparable countries. In the UK this gap stands at 24.5 per cent, whereas in countries like Italy this gap is 27 per cent, and Spain around 23 per cent. Most other countries are in between these two extremes. At the same

time, lone parent employment is rising rapidly in the UK. It has risen by 11.3 per cent over the period 1997 to the present day.

In 1998, the Government introduced NDLP nationally as one of a range of policy initiatives aimed at lone parents. The programme aims 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'* (DWP 2002).

In 2000, the Department for Work and Pensions (DWP) commissioned NatCen to evaluate NDLP and to answer three main questions, among them *'What Impact does NDLP have?'* (Lessoof, Miller et al. 2003). The NatCen report summarised the recent history of policy measures targeted at lone parents and describes the design of the NDLP. Specifically, the NatCen report described who takes part in NDLP, what taking part involves and what those surveyed think of the programme. The NatCen report also contains details of the data available to this research assessment from the initial postal questionnaire and from the subsequent face-to-face interviews. This report examines only the NatCen evaluation methodology and the impact assessment.

The NatCen report suggested that 46 per cent of participants in NDLP had left Income Support (IS) nine months after NDLP participation, compared with only 20 per cent of matched non-participants. This suggests that 26 per cent of lone parents had left IS because of NDLP who would not otherwise have done so. This is a dramatic effect for a voluntary programme.

From the outset it should be clear that what treatment an NDLP participant receives is not entirely straightforward. The key features are that it is a voluntary programme that offers a lot of help and guidance to the participant but also some significant financial inducements, although these do not come in the form of increased benefits. This should be borne in mind in the interpretation of the findings.

Main findings

- The NDLP impact evaluation is not straightforward as it is taking place in a complex environment. The programme itself is a voluntary one in which the recipients can still be in receipt of IS and other benefits. It will, therefore, be unclear how to treat NDLP participants who are not actually on IS – are these individuals to be judged to be in receipt of benefit or not, because they are clearly receiving a state subsidy? The existence of a variety of other important New Deal programmes, most particularly the New Deal for Young People (NDYP), complicates the evaluation of the NDLP. This evaluation conditions on the existing programme environment, within which these other programmes affect the choices and outcomes of participants, eligible non-participants and employers. The impacts of NDLP might differ substantially in an environment without these other programmes. This report briefly discusses these important complexities.

- This research investigated the problems which arise with the consistency of the data which results from merging information containing NDLP, IS, Jobseeker's Allowance (JSA) and Incapacity Benefit (IB) records of lone parents. A number of lone parents appear to have overlapping spells. In most cases these are feasible admissible contemporaneous spells, since people can be on NDLP at the same time as claiming any benefits such as IS, IB or JSA. A set of rules was developed to solve this problem.
- The report investigates the full characterisation of the underlying econometric model which would be necessary to analyse the NatCen data. The model is quite complicated but the version used by NatCen is relatively simple and relies on a number of strong and potentially controversial simplifying assumptions. These assumptions are made explicit, as it is only by recognising them that the treatment effect found by NatCen for the NDLP programme can be put into context.
- The research indicates that although NatCen adopted a number of approaches that are unconventional relative to the econometric matching literature, the estimates obtained are relatively insensitive to these dimensions.
- Adjusting for selective non-response based on observable characteristics makes very little difference to the estimates. Despite these findings, there remains the possibility that an analysis that incorporated selective non-response based on unobserved factors correlated with outcomes might tell a different story.
- Another important issue explored in this research was that the window of time available for an individual to participate in NDLP after returning their postal questionnaire varied among the individuals in the sample. Taken together, the evidence presented suggests that the window definition makes a substantive difference to the resulting impact estimates.
- The research also explored the importance of conditioning on local labour markets, or variables at the local labour market level. This involved merging the evaluation data with the administrative data that contain these variables. It was found that adding these local labour market data did not change the results significantly.
- This assessment of the impact effect of NDLP was conditioned on detailed benefit histories using the administrative data. The focus here was on the characterisation of the past benefit history over the previous six quarters and the most recent weeks. The techniques suggested by Card and Sullivan were followed to summarise the 64 possible sequences of six previous quarters of labour market state information. These dummy variables proved to be very powerful in satisfactorily matching participants and non-participants prior to the sample date.

- The impact of NDLP differed substantially between those clients who were in the stock of benefit claimants when NDLP was introduced compared to those who flowed onto benefits since NDLP became live. The preferred impact estimate was computed using the full administrative dataset for lone parent benefit claimants who were new to the flow of benefit claimants and were not on NDLP most of the time in the preceding six quarters.
- The sampling problem of drawing a representative sample of individuals who are eligible to receive the NDLP treatment and correcting this sample for any observable differences between the administrative sampling units, the postal questionnaire respondents and those who received a face-to-face interview, is non-trivial. The report outlines the appropriate method for weighting the data based on these sampling problems.

Key conclusions

- In 2002 this report was commissioned by the DWP to examine the accuracy and robustness of the results from the NatCen NDLP evaluation in a number of dimensions; in particular, to determine the sensitivity of their estimates to changes in the estimation procedure and outcome variables, and to look at longer-term outcomes.
- A thorough review of all the data available was conducted to assess the impact of NDLP for the cohort who were eligible to participate in this programme in August 2000. Focusing on the simple outcome measure of the probability of being on benefit, it is suggested that the NatCen impact of 26 per cent was too high. The preferred impact, or additionality, estimate of NDLP is 14.24 per cent, using the full administrative dataset. This estimate is based on the mean impact (for up to 30 months) for only those clients who were on the 'flow' (as defined as those individuals not on benefit for over 50 per cent of the time in each of the six quarters prior to the participation window).
- The econometric evaluation in this report makes use of several somewhat innovative features of the NatCen survey sample design. Most importantly, the NatCen data collection strategy involves matching NDLP participants with non-participants using administrative data and information from a postal survey in order to determine which non-participants should receive face-to-face interviews. The econometric literature does not document this prescriptive use of statistical matching.
- The analysis in this report makes explicit the statistical and economic assumptions underlying the NatCen analysis, as well as checking the robustness of their computational and estimation methods. By recreating the logic of the NatCen analysis from scratch, this report clarifies the conceptual issues involved in their analysis and highlights areas that would benefit from further research.

- The econometric work investigated in this report is very much on the boundary of what is known in applied econometric methodology regarding the evaluation of active labour market policy. The reasons are that NDLP policy may involve: treatment for different lengths of time for different individuals; non-participants becoming participants at a later stage; and some NDLP recipients being participants on more than one occasion. These issues are explored and the econometric implications of evaluation in this context are spelt out. However, each of these econometric problems deserves much closer scrutiny in further research work.
- Further research on the evaluation of the NDLP programme is also possible in a more ambitious context. It is possible to link the data made available to us on the benefit and claimant status of each lone parent with what they subsequently earn when in work via Inland Revenue data. If these data becomes available then a more sophisticated analysis of the outcomes of the NDLP programme may be possible. Specifically, it could be possible to perform a cost benefit analysis of exactly what was paid out to the lone benefit recipient under the NDLP programme and link this to what contribution they subsequently make to the Exchequer in the form of income taxes.

It is important to note that this NDLP evaluation examines a particular cohort of NDLP participants; as such, great care should be taken in generalising the results of this evaluation to any other programme or even the same programme at any other point in time.

1 Introduction: The evaluation of the New Deal for Lone Parents

Summary

The New Deal for Lone Parents (NDLP) is a national voluntary programme, aimed at helping lone parents into work, improving job readiness and supporting them in employment. The key features of the programme are the meetings with the Lone Parent Personal Advisers. These must be regarded as an integral part of the 'treatment' of NDLP. These meetings were essentially the same as what are now called Work Focus Interviews (WFIs) (which were not mandatory for the cohort of participants analysed in either this or the NatCen evaluation).

In 2000 NatCen was commissioned by the Department for Work and Pensions (DWP) to conduct a quantitative assessment of NDLP as part of a large scale evaluation of who takes part in the programme, what participation involves, and to assess the impact of NDLP on lone parents' outcomes.

In 2002 the authors of this report were commissioned by DWP to examine the accuracy and robustness of the results from the NatCen report along a number of dimensions; in particular, to determine the sensitivity of their estimates to changes in the estimation procedure and outcome variables, and to look at longer-term outcomes.

It is important to note that this evaluation examines a particular cohort of NDLP participants; as such, great care should be taken in generalising these results to other cohorts of NDLP participants or indeed to other similar Active Labour Market Programmes (ALMPs) in other countries.

1.1 Introduction

There are estimated to be about 1.75 million lone-parent families in the UK, with about 2.9 million dependent children (Evans et al. 2003). In 1998 the Government introduced NDLP nationally, as one of a range of policy initiatives aimed at lone parents. The programme aims 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'* (DWP 2002).

In 2000, the DWP commissioned NatCen to evaluate NDLP and to answer three main questions, among them; *'What Impact does NDLP have?'* (Lessof, Miller et al. 2003).

The task of this report is to conduct a thorough investigation of exactly how NatCen obtained their impact estimates, to determine the sensitivity of their estimates to changes in the estimation procedure and outcome variables, and to look at longer term outcomes. An insight into the particular econometric difficulties imposed by the evaluation of the NDLP programme is also provided.

This report (in the following chapters and their respective appendices) presents the main findings of the evaluation:

- **This chapter** provides background information about NDLP and its' policy context, and summarises the main findings of the NatCen quantitative evaluation. It also describes the aims and scope of this research.
- **Chapter 2** describes the sample design, the sampling method, and the main administrative datasets used in this evaluation.
- **Chapter 3** sets out the basic notation used throughout the report, explains the statistical matching methodology which is used to calibrate the impact assessment and discusses the econometric model implicit in the NatCen evaluation design.
- **Chapter 4** focuses on the main findings of the NatCen report. Special emphasis is given to the propensity score calculations and the investigation of its applicability to this data. The common support problem, the matching algorithm used and balancing test on appropriate covariates, is discussed. The sensitivity of the results to alternative matching methods is also examined.
- **Chapter 5** provides a further sensitivity analysis of the estimates to the non-response biases and how the NDLP participation window differing for different respondents is handled.
- **Chapter 6** estimates alternative NDLP impacts over a longer period of time. The robustness of the impact results to different matching techniques and different characterisations of the pre-treatment labour market history, is examined. The expected NDLP impacts conditioned on the propensity score are investigated and whether the impact effect is different for different subgroups based on the age of the lone parent's child and whether they live in the north or south of the country is explored.

- **Chapter 7** raises further econometric issues related to the econometric evaluation of NDLP. This discussion is essential if the limitations of the present research are to be appreciated and the scope for more advanced technical econometric analysis is to be understood.
- **Chapter 8** provides an interpretation of the main findings on the impact of the NDLP programme and how it relates to the existing international literature.

1.2 The New Deal for Lone Parents

The New Deal for Lone Parents is a voluntary programme whose objectives are to:

- help and encourage lone parents to improve their prospects and living standards by taking up, or increasing hours of, paid work;
- improve the job readiness of lone parents to increase their employment opportunities.

In the early stages of NDLP (and at the time the data were collected) the participant was offered advice and support from a Lone Parent Personal Adviser (LPPA). The terminology has now changed and the programme now provides advice and guidance through Work Focus Interviewers whose job is, through a series of interviews or contacts, to:

- encourage and motivate all lone parents to identify their skills and develop confidence;
- provide support and guidance to clients who are ready to search for work and provide practical support in finding and applying for jobs;
- improve awareness and knowledge of benefits and tax credits and provide practical support organising and applying for them;
- help with the transition from claiming Income Support (IS) into work by providing 'better off' calculations, assisting with in-work benefit claims and liaising with employers;
- identify and provide access to education or training courses with a 'direct' work outcome to increase job readiness (including Jobcentre Plus programmes), and provide practical support applying for education and training;
- provide practical support in finding childcare, applying for child maintenance and liaising with the Child Support Agency;
- offer in-work support.

The NDLP participant is not given any extra IS or other benefit than they would otherwise be eligible for, but they may be eligible for financial help with travel costs to attend job interviews, childcare costs or fees for training courses recommended by the LPPA. Money is available to meet such expenses but only if approved by the LPPA. Payments for childcare and course fees are made direct to the provider at a rate of up to £67.50 for one child and £100 for two or more children.

The consideration of exactly what treatment an NDLP participant receives is, therefore, not entirely straightforward. The key features are that it is a voluntary programme but one that offers a lot of help and guidance to the participant, and also some significant financial inducements, although these do not come in the form of increased benefits. This is a moot point, and care must be taken in the analysis in terms of how individuals who are on NDLP but not in receipt of IS are treated (in terms of whether they have left benefit or not), if this is to be the crucial outcome measure. Are NDLP participants who are not on IS treated as being still on benefit? This should be borne in mind in the interpretation of the findings.

In the assessment of the NDLP impact effect, it should also be noted that NDLP has developed and changed over time since its initial introduction. Hence, in essence, the nature of the NDLP treatment has itself changed. Below is a brief outline of how NDLP has changed, but this issue is returned to in Chapter 7.

- A prototype was launched in July and August 1997 in eight locations with the former Employment Service and Benefits Agency, each responsible for four of these areas. This was known as Phase One and was evaluated in its own right (Hales, Roth et al. 2000). In April 1998, the programme was introduced nationally for new and repeat claimants (Phase Two). NDLP was then rolled out fully to all lone parents on IS in October 1998. This is known as Phase Three and is the focus of the present study.
- NDLP has also expanded its target population and rules of eligibility. Initially NDLP was only available to lone parents on IS; the primary method used to target prospective participants were invitation letters, which were sent to target groups, that is those lone parents whose youngest child was aged over five years and three months. Those with younger children (the non-target group) could join the programme by putting themselves forward. The target group was then extended to include lone parents whose youngest child was aged over three years. Over time, the distinction between the target and non-target group has diminished, as the number of lone parents coming forward have been found to be similar among those with older and younger children. Since November 2001, eligibility has been extended to all lone parents who are not in work or are working fewer than 16 hours a week, including those who do not receive benefits but who want the support and advice offered by the programme.

- There have been some instances where the voluntary nature of help given to lone parents has not been preserved. While NDLP itself has remained voluntary, two related policy developments (the ONE Pilots and WFI meetings) have introduced compulsory interviews at the start of a claim for benefit and at intervals while the claim is ongoing. These other initiatives have been evaluated separately (Coleman, Rousseau et al. 2003).

Overall, 577,720 new lone parents spells can be identified in the NDLP database between October 1998 and December 2003¹. Figure H.1 (in Appendix H) shows monthly totals of programme participants (flow plus stock) from October 1998 to December 2003. A clear trend of increasing numbers of participants over the life of the programme is evident. The pattern was that participation rose rapidly to a peak in September 1999 as the stock of eligible existing lone parents was introduced to the programme. Then participation levelled off at around 60,000 and stayed at approximately this level until October 2001. When the eligibility criteria were extended in November 2001, participation rose more or less continuously until its level of 100,000 by the end of 2003.

Figure H.2 (in Appendix H) shows the monthly total of new entrants onto NDLP between November 1998 and December 2003. As noted by Evans *et al.* (2003) the overall trend in numbers is less straightforward to discern from this graph because of the substantial seasonality. Table H.1 gives the average monthly number of new entrants for each year to enable the reader to identify overall trends in the number of monthly entrants per year. As can be seen, the monthly average number of NDLP new entrants has been increasing throughout the entire period.

Looking more closely at Figure H.2, seasonal variation in the numbers of new entrants can be explained from the logical forces of the annual cycle. Most marked is the fall of the number of new entrants every December to around 5,000 or less. The peak months for new entrants are March, May and September as these logically coincide with the beginning of new school terms as a new group of lone parents becomes more eligible to seek work when their children start school. The months of April, June, July and August have less new entrants due to the school holidays.

1.3 The NatCen report

NatCen was commissioned by the DWP in 2000 to conduct an impact evaluation of the NDLP. The study is described in more detail in the full report (Lessof, Miller et al. 2003) and the accompanying technical report (Phillips, Pickering et al. 2003).

A systematic review of the detailed contents of the NatCen report is not provided, but a familiarity with the results of the NatCen report is presumed, which is briefly summarised where appropriate. The NatCen report summarises the recent history

¹ As Evans et al (2003), noticed, this overall figure is inflated by the fact that some participants have joined the programme more than once.

of policy measures targeted at lone parents and describes the design of NDLP. Specifically, the NatCen report describes who takes part in NDLP, what taking part involves and what those surveyed think of the programme. The NatCen report also contains details of the data available to this research assessment from the initial postal questionnaire and the subsequent face-to-face interviews. This work examines only the NatCen evaluation methodology and their impact assessment (contained in Chapter 7 of their report).

It is, however, important to note some of the broad characteristics of the NatCen study. A cohort of lone parents was selected for the evaluation between August and October 2000 and data was collected about them through to the end of 2001. At the time of the NatCen evaluation, all lone parents on IS were eligible to join NDLP and the target population had recently been extended from lone parents with children of five years and older to those with children of three years and over. Other changes were also introduced to the programme during the course of NatCen evaluation, for example, the introduction of compulsory WFIs. As a result, it is important to remember that the NatCen evaluation took place during a particular time period and measures the effect of a programme in a particular phase, within a particular social and economic context, on a particular cohort of individuals (Lessoof, Phillips et al. 2003). The circumstances impose qualifications which apply to the general applicability of the NatCen report, which also apply to this report.

In considering the NatCen report, in particular several somewhat innovative features of their sample design and of their evaluation design complicate the evaluation: Firstly, and most importantly, the NatCen data collection strategy involves matching NDLP participants with non-participants using administrative data and information from a postal survey in order to determine which non-participants should receive face-to-face interviews. The econometric literature does not document this prescriptive use of statistical matching. As far as is known, unpublished evaluations of various Canadian ALMPs constitute the only earlier example of this technique. It poses a number of interesting technical problems, especially in relation to non-response. Secondly, the sample design involves two surveys: a postal survey and a face-to-face interview survey. This means that the final sample embodies two different survey non-response processes, which can potentially bias the evaluation results in two different ways. Finally, several of the outcomes of interest to the DWP involve the duration of time an individual spends in a labour market state, eg, the length of time they remain on IS following NDLP participation. As with the literature on how to assess treatment effects with outcomes that are measured at different points in time or are indeed themselves duration variables remains in its infancy, additional difficulties arise on this dimension. In particular, as will be shown, the choice of outcome variable looms large in this context, and can have strong implications for the conclusions drawn in an evaluation.

1.4 The policy context

Over the period 1997 to date, the position of lone parents in the UK has changed remarkably. This is due to the commitment of the Government to tackle child poverty. As a result, the benefits available to lone parents have considerably widened and deepened over the 1997 to 2004 period. This has, in large measure, been due to the tax and benefit changes which included the introduction of the Working Families' Tax Credit (WFTC) in 1998 (now Working Tax Credit), the introduction of NDLP, the National Childcare Strategy and the National Minimum Wage.

It is well reported by Gregg and Harkness (2003) that there are several stylised facts about the nature of lone parents in the UK: Firstly we know that lone parent employment in the UK is very low relative to other countries. Indeed, the gap between employment rates of lone parents and those of married mothers is higher in the UK than nearly all comparable countries. In the UK this gap stands at 24.5 per cent, whereas in countries like Italy the same gap is 27 per cent and Spain around 23 per cent. Most other countries are in between these two extremes. At the same time we know that lone parent employment is rising fast in the UK. Gregg and Harkness (2003) report that it has risen by 11 per cent over the period 1983 to 2002, so that it now stands at around 52 per cent.

Although the number of lone parents in work has been rising, the number in receipt of WFTC have also been rising. Part of this increase is due to the falling numbers of lone parents on IS.

There have been very few attempts at the evaluation of the impact of NDLP. Gregg and Harkness (2003) use the Labour Force Survey (LFS) to examine the difference in employment rates between lone parents and other comparator groups over the period 1992 to 1998 and 1998 to 2002. In essence, they examine the difference between these two groups of people over the pre- and post-reform period. Then they take differences in these estimates. It is suggested that the overall effect of the welfare reforms including NDLP and the changes to benefits and WFTC have had an effect of five per cent on lone parent employment rates.

The Gregg and Harkness (2003) result needs some clarification, as it might be tempting to try to reconcile their impact estimate with the ones which have been described in this report. This temptation should be avoided as there is no clear way of comparing the two impact assessments. This is evident for three reasons: firstly NatCen are looking at aggregate employment rates, secondly they compare lone parents with married mothers and other groups, and thirdly they do not attempt to distinguish between those individuals who have been on the NDLP programme and those who have not. In contrast, the results presented in this report are for the outcome measure 'being in receipt of benefits' for a particular cohort of people who were eligible for NDLP in September 2000. Secondly, this analysis compares directly the outcome for those lone parents who have been NDLP participants with a

matched group of lone parents who have not been on NDLP. Hence, it should be understood that the impact assessments of the treatment effect of NDLP of around 14 per cent on being in receipt of benefits, cannot be compared with aggregate employment effects.

1.5 Aims and scope of this evaluation

The task of this report is to examine the accuracy and robustness of the results from the NatCen report along a number of dimensions. Specifically, we were asked to:

- examine the sensitivity of the NatCen analysis, in order to understand how their propensity score estimates were derived, how the samples for the face-to-face interviews were selected and how the impact estimates were constructed. This necessitated the replication of their analysis;
- explore specific issues surrounding the ‘windows’ (length of time between the return of the postal survey and participation in NDLP) used to define the analysis sample and issues surrounding lone parents who participated in NDLP while employed;
- explore the issues of non-response and weighting;
- produce impact estimates for key variables such as numbers off benefits, and numbers into jobs and report on the possible range within which the true impact of NDLP lies;
- review key outputs from the evaluation, in particular: draft reports, summaries and technical papers;
- lend support to the overall evaluation approach and methodology;
- examine longer-term outcomes of NDLP up to the end of 2003.

The analysis in this report will, therefore, make explicit the statistical and economic assumptions underlying the NatCen analysis, as well as checking the computational details of what NatCen did, and examine the robustness of their computational and estimation methods. By recreating the logic of the NatCen analysis from scratch, the conceptual issues involved in their analysis will be clarified and areas that would benefit from further research highlighted.

To review the computational procedures utilised by NatCen an answer will be sought to the following basic questions:

- Is the logit model used by NatCen to estimate the propensity scores for the NDLP participation the best available? If not, what are the best alternatives?
- How were the propensity scores for the participation decision in the NDLP programme estimated?

- Do the estimated propensity scores generated by the NatCen logit model pass the 'balancing test' suggested by Rosenbaum and Rubin (1983) and routinely employed in matching analyses in the published literature? The 'balancing test' examines the statistical equivalence of the distributions of the covariates included in the propensity score model, conditional on the scores.
- How much difference does it make to match using the 'psmatch.do' algorithm written for Stata rather than the Blaise programme written and used by NatCen?
- Does it make a difference to the results if the matching is done with or without replacement?

To examine the robustness of the NatCen matching, the following was checked:

- How much difference does it make if, in computing the impact estimates, outcomes only for matched pairs of participants and non-participants are compared, matched by the NatCen Blaise algorithm (or the psmatch Stata routine), wherein both members of the pair completed face-to-face interviews?

To examine the estimation procedures utilised by NatCen the following can be checked:

- How much difference does it make to the impact estimates if they are adjusted for non-response based on observable characteristics to the postal survey and/or the face-to-face interview survey?
- How much difference does the definition of the participation window make to the impact estimates?

In addition, this paper will seek to establish:

- How much difference does it make to the estimated treatment effect if a participant was already in a job at the date of (or just prior to) participation in the NDLP?
- How much difference does it make to the impact estimates if alternative outcome measures are used from the administrative data that track participation over time, rather than the cumulative incidence outcome measures employed by NatCen?
- How much difference does it make to the estimated treatment effect if additional variables drawn from the administrative data, and from the face-to-face interviews, are utilised to construct the matching estimates?
- The extent to which a careful analysis of the data in the sample relating to age of the youngest child and duration of IS claim make a difference to the size of the estimated effects of the programme.
- The size of the treatment effect when considering the 'flow' sample of those entering a claim rather than the 'stock' sample of those on a long-term IS claim at the time of the survey.

2 Sample design and sampling issues

Summary

This chapter introduces the main datasets used in this research. The chapter is particularly relevant given that it introduces the sample definitions that will be used throughout this report. It is important to note that the same definitions are being used in this report as the National Centre for Social Research (NatCen) used in their report, and that the cohort sample of the Income Support (IS) eligible population used in this analysis is fully comparable with that which was used by NatCen.

This chapter also explores the logical problems with the data which result from merging the New Deal for Lone Parents (NDLP), IS, Jobseeker's Allowance (JSA) and Incapacity Benefit (IB). A number of clients appear to have overlapping spells. In most cases these are feasible admissible contemporaneous spells since people can be on NDLP at the same time as claiming any benefits such as IS, IB or JSA. A person can also claim both IS and IB at the same time. However, some overlaps are technically impossible, since the benefit eligibility rules formally prevent somebody from simultaneously receiving IS and JSA or IB and JSA. Together these 'illegal' overlapping spells constitute slightly less than 1.0 per cent of the total sample of 69,829 claims. It is worth noting that the development of the Master Index by the Department for Work Pensions (DWP) will make the merging process of the different extracts described in this chapter substantially simpler.

2.1 Introduction

In this chapter the different datasets used in this analysis are examined. Any statistical evaluation of an Active Labour Market Programme (ALMP) must rely on how good the data quality is to facilitate the detailed analysis. The sample design on the data collected by NatCen for the evaluation of NDLP is complex and unusual.

Specifically, it involves the use of administrative register data to identify those who are eligible for NDLP, and then issues a postal questionnaire to a sizeable sample of these people. The design was then to wait to see which of these individuals in this sample actually participated in NDLP. These people were matched statistically to those 'who most looked like them' in the sample of non-participants – then and only then – were all these individuals sought for a face-to-face interview. This prescriptive use of matching was innovative.

Section 2.2 presents the sample design developed by NatCen. Section 2.3 describes the sampling method of selecting a representative sample of people eligible for NDLP in August 2000 and outlines the issue of getting a geographically representative sample from different Travel to Work Areas (TTWAs). This section also discusses the drawing of a Wave 2 booster sample in October 2000 to restore a balance of the flow of new entrants to the scheme.

Section 2.4 describes in careful detail the main datasets used in our analyses, distinguishing between those who were only in the administrative data and those who responded (or not) to the postal questionnaire and face-to-face interview. The actual samples we used in the analyses are made clear.

Section 2.5 describes the main administrative datasets used in the analysis and Section 2.6 explains in some detail how the IS, IB, JSA and NDLP labour market history data was reconciled. Actual cases are used as examples to clarify how data which contained erroneous overlapping spells on different benefits was organised and corrected, and presents the rules used in this research to address this issue.

2.2 Understanding the sample design

The NatCen evaluation of NDLP has, of necessity, a rather complex sample design. Chapter 2 of the NatCen report fully describes that design. This section briefly summarises key aspects of the design that play a role in this study because of their potential effects on the magnitudes and interpretation of the impact estimates presented in the NatCen report.

The NatCen design uses administrative data to draw a stratified sample around 65,000 people from the eligible population to participate in NDLP in August 2000, complemented with a further sample of around 5,000 people from the eligible population using a subsequent extract from October 2000 (see Chapter 2 of NatCen (2003b) for further details). The final stratified sample of this population, which included around 70,000 people, received a postal survey. It is important to note that the postal questionnaire design team was especially concerned with the role of unobserved covariates, in particular, attitudes on the propensity of NDLP participation, and hence, collected a set of such variables prior to NDLP participation to use on their propensity score specification. On the basis of the information collected in the postal survey, NatCen matched individuals who participated in NDLP within a given calendar time period to non-participants with similar characteristics. NatCen then

attempted to conduct a face-to-face interview with all of the participants and their associated matched non-participants. The final NatCen report estimates the impact of NDLP participation by comparing the outcomes of NDLP participants who completed a face-to-face interview with the outcomes of non-participants who completed a face-to-face interview.

The time line in Figure 2.1 summarises the principle features of the sample design. It shows the temporal relationships between the definition of the population of interest, the two postal and face-to-face interview surveys and programme participation as defined in the NatCen study. At time $t = 1$, corresponding to August 2000, the set of lone parents on IS in the administrative data who meet the sample inclusion criteria defines the population of interest. Everyone in this population received a postal survey as described in the NatCen report. Respondents to the postal survey return their survey at time $t = q$, between October 2000 and March 2001.

Participation, as measured in the NatCen study, takes place at time $t = p$, in the interval from August 2000 to 28 April 2001. The NatCen study divides the participants into three groups based on whether, and when, they returned their postal surveys. Group B participants returned their postal surveys prior to participation in NDLP, so that $q < p$. Group A participants return their postal surveys after starting to participate in NDLP, so that $p < q$. Group D participants did not return their postal surveys. The majority of the NatCen analysis focuses on the Group B participants, omitting the Group A participants due to concerns that participation in NDLP before responding to the questionnaire may influence responses on the postal survey. Group D participants who do not respond to the postal survey are omitted from the NatCen analysis. The non-participant sample includes lone parents who participate in NDLP after 28 April 2001. This practice is fairly standard in evaluation research, though it is often not recognised that it has important implications for the nature of the parameter estimated in the evaluation. Lechner (1999), Sianesi (2001a) and Fredriksson and Johansson (2002) provide discussions of this issue, which will be returned to later (in Chapter 7).

Respondents to the postal survey who also complete the face-to-face interview survey do so at time $t = f$, between November 2001 and February 2002. The face-to-face interviews take place after an interval of several months in order to allow time for the effects of participation to become apparent in the labour market.

2.3 Sampling

The sampling exercise built on the approach used in the pilots. The main differences were that the definition of geographical units was conducted on a national scale, rather than in small areas. Specific geographical areas were excluded to avoid overlap with existing pilots (though an error was made and some correction was needed during the study). Individuals were sorted into groups according to characteristics (as in the pilot) but selected proportionately rather than having a fixed number from each group.

The construction of the sample used by NatCen was complex. There were basically five stages involved in the construction of weights to ensure that the sample used for the outcome evaluation was representative of the population:

- firstly, defining the appropriate geographical basis of the investigation by sampling from the Primary Sampling Units (PSUs);
- secondly, selecting the sample from the PSUs;
- thirdly, allowing for the fact that the sample size of the sample drawn at August 2000 (Wave 1) was too small due to an error in the original sample, a Wave 1 replacement sample was drawn from the same August 2000 population. In addition to correct for low sample numbers on short claim durations, a second sample (Wave 2) was drawn in October 2000;
- fourthly, taking account of the different distribution of age of youngest child and duration of present claim in the sample and the population;
- fifthly, in using only the sample respondents the NatCen outcome measures could be biased due to non-response and hence they attempted to weight to control for this based on some observable characteristics.

This section explains each of these steps and provides a rationale for where departures from the NatCen report occur, in order to analyse the full administrative data.

Stage One: Definition of primary sampling units

For the purposes of the postal pilot, DWP provided NatCen with IS records for April 2000 excluding participants known at that time. This data set was interrogated to find out the number of lone parents in each postcode district and sector and so provided information that could be used to divide the country into about 500 geographical units. These formed the basis for identifying the final primary sampling units.

Rather than creating geographical units based solely on the number of lone parents in each post code district or sector based on a map of the country, TTWAs were used as the basic building blocks. Where necessary, TTWAs were split or combined while attempting to preserve their integrity as far as possible. Using TTWAs meant that each geographical unit would have some relevance to local labour markets. In addition, it meant that the information available from Nomis4 about each TTWA could be used (combined with other information collected for postal districts) to stratify the primary sampling units.

Initially, the TTWA that each lone parent lived in was identified and the number of lone parents in each postcode sector within TTWAs was listed. Using a postcode map to identify adjacent postcode districts and postcode sectors, clusters of postcode sectors were selected to arrive at sensible geographical building blocks.

The aim was to ensure that each geographical unit contained between 1,500 and 2,500 eligible lone parents, although these boundaries were treated flexibly when appropriate. Some TTWAs contained between 1,500 and 2,500 lone parents and so were treated as whole primary sampling units; other TTWAs contained many more than 2,500 lone parents and so were sub-divided. In the remaining areas, TTWAs contained only a few lone parents and so were combined with adjacent TTWAs.

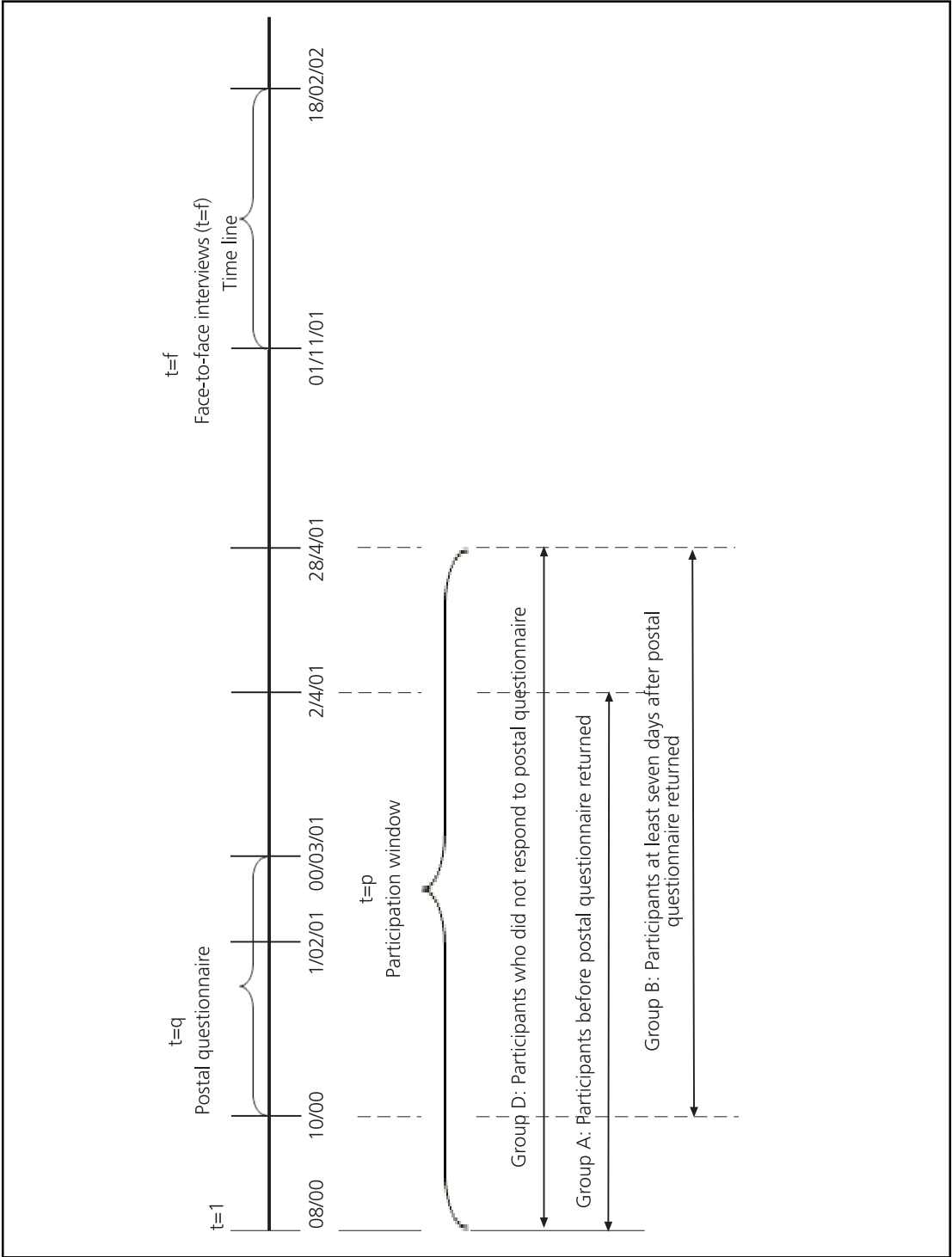
This process generated 510 primary sampling units as well as some 'surplus' geographical areas that were defined as out of scope. These were postcode districts in Northern Ireland, and those in England, Scotland and Wales that were beyond the scope of the NatCen field force; for example, in areas north of the Caledonian Canal which represent a very small percentage of the total UK (and lone parent) population.

After that, a process was undertaken to further modify the primary sampling units to remove postcode sectors or districts where:

- ONE action and control areas were located;
- Work Focused Interview (WFI) pilots were located;
- NatCen had carried out piloting for the evaluation.

Some further adjustments were made to avoid postcode districts overlapping between TTWAs or PSUs and the size of the eligible lone parent population in each PSU was then recalculated using the most up to date IS records available. It was expected that the number of eligible lone parents in each of these areas would remain broadly similar over time. This provided the opportunity to check that this was the case, and to ensure that the selection of each PSU was based on the actual size of the eligible population. Because of these exclusions and readjustments, some PSUs were combined with others or removed from the sampling exercise (where there were less than 250 eligible lone parents left within a PSU).

Figure 2.1 Data collection time line²



² The data include three observations with recorded face-to-face interview dates of 6 January 2001, 9 January 2001, and 18 June 2001. These dates are necessarily incorrect as they lie outside the time period for the face-to-face interviews. The errors probably result from faulty internal clocks on the computers used for the CAPI interviews. As the data required to correct these errors was not available, the observations in what follows are ignored.

Stage Two: Selecting 130 Primary Sampling Units

The target issued sample size was 70,000. It was recognised that because of the time it would take to select a sample of this size, by the time the questionnaire was issued there would be no one in the sample who had made a fresh claim within the last few months. Accordingly, it was decided to draw an initial sample of 65,000 cases and then a further sample of 5,000 to 6,000 in the same areas at a later date (using subsequent downloads of the IS records) so as to ensure a reasonable proportion of fresh claims.

- (A) In order to do this, 130 PSUs were selected with 500 lone parents in each. For each PSU, the August 2000 IS records were used to calculate the size of the eligible lone parent population excluding:
 - a. lone parents over retirement age;
 - b. lone parents with the youngest children over 16;
 - c. lone parents who have already participated in NDLP;
- (B) the labour market situation in each PSU based on the unemployment rate in TTWAs for men and women combined (with high, medium and low rates of unemployment);
- (C) the NDLP participation rate based on the number of participants in each PSU as a fraction of the total lone parent population on IS; and
- (D) the population density in each PSU.

The PSUs were sorted based on these criteria and the desired number of PSUs were randomly selected. Each PSU was sampled with a probability proportional to the size of the eligible lone parent population in each PSU.

Stage Three: Selection of individuals from the sampling frame

New cases to boost the original sample (Wave 1 Replacement sample and Wave 2 Booster Sample)

The first and main sample comprised 48,000 lone parents selected from the August 2000 IS records. A replacement sample of 16,700 cases was selected from the August 2000 IS records – this was known as a Wave 1 Replacement sample by NatCen. It consisted of new cases drawn from largely the same³ PSUs selected at Wave 1. This sample was selected to compensate for an error made when drawing the initial sample. The Wave 1 Replacement sample has already removed individuals who had already participated in NDLP according to the October New Deal Evaluation Database.

³ Phillips *et al.* (2003) indicates that NatCen used 34 fresh PSUs to replace 24 ineligible PSUs to use a sample of 139 PSUs.

A major problem with the Wave 1 and Wave 1 Replacement sample was that it did not contain individuals with very short claims of under three months' duration. For this reason a Booster Sample was undertaken of 5,027 cases to rectify this problem.

NatCen believe that the selection of the Wave 1 Replacement and Booster Sample for the new cases has not been compromised by the need to re-sample some of the original cases. The number of eligible cases issued for the main sample was 64,973. Therefore to achieve a complete sample of over 70,000 cases (as proposed) would require a further 5,027 cases to be selected for the new cases booster sample. The number of eligible new cases in the previously selected PSUs was 6,824. From these, a fixed number (41) or all cases (if less than 41) were sampled within each PSU. This resulted in 5,090 new cases being selected amounting to a total sample of 70,063 cases (see footnote 3).

Stage Four and Stage Five: involve allowing for the Stata associated with duration of claim and age of the youngest child and allowing for the bias associated with non-response. This report describes what NatCen did and what procedures were used in Section 5.

2.3.1 The overall results

After all the sampling stages reported above were replicated from the NatCen original data, a position was reached that enabled the computing of the outcome estimates which are reported in this document. In this analysis the outcome figures are actually plotted using the weights which incorporate the effect of the 24 stata. For each of the main outcome figures this procedure consisted of computing 24 separate estimates for each of the 700 weeks in the data. The final outcome is the weighted average of these 24 estimates.

2.4 The sampling method

Tables 2.1 and 2.2 are drawn from the draft NatCen report (NatCen, 2003a), highlighting the different samples that result from the sample design illustrated in Figure 2.1 (in the Technical annex). In particular, they define six different types of observations in the data, denoted groups A to E. It is important to clarify what role each of these samples plays in the subsequent analysis.

2.4.1 Select a sample of eligible non-participants

The first stage was to collect data from lone parents on the factors thought likely to influence participation. This was achieved by selecting a very large sample (69,851) of eligible lone parents who had yet to participate in NDLP from all those eligible for NDLP in August and October 2000. The sample was drawn from two downloads of the Income Support Computer System (ISCS) database. Any sample drawn from a cross-section of a population will tend to over-represent the 'stock' of claimants and under-represent the annual 'flow' of those who enter the programme, particularly

since many of these will leave quickly. As a result, a Booster sample of new entrants was selected to compensate for this under-representation. Areas with above average rates of participation were also over-sampled.

2.4.2 Carry out a postal survey of the sample of eligible non-participants

All of these 69,851 lone parents were sent a postal questionnaire (cost prohibiting any other survey mode on this scale) covering issues that were thought to relate to participation in NDLP. A copy of the questionnaire is included in Appendix A of Phillips et al (2003) and the main findings are presented elsewhere (Lessof et al, 2003). In summary, questions were asked about number and age of children, current and past work experience, barriers to work, attitudes to work and childcare and aspirations for work and training in the future. Other questions were included such as access to a phone, ethnicity, and qualifications. Many of the topics included in the questionnaire were identified from existing literature about lone parents and from the literature about propensity score matching.

In total, 42,273 completed questionnaires were returned. Since some cases were identified as ineligible, this represented an overall response rate of 64.4 per cent. Fieldwork took place between October 2000 and April 2001 and consisted of up to five mailings (first contact and then a series of reminder letters and replacement questionnaires) to each sample member. As part of the postal survey, respondents were informed that the data would be analysed by members of the research team at NatCen and the (then) Employment Service. As well as providing assurances about confidentiality, a reference was made at the end of the questionnaire to the possibility that respondents would be invited to take part in further research in the future.

2.4.3 Using administrative data to identify individuals entering NDLP

On a monthly basis, the sample of 69,851 individuals was tracked using the NDLP evaluation database so that all individuals who subsequently participated in NDLP could be identified. 'Participation' itself was defined as having an initial NDLP interview. It should be noted that this definition differs from the official definition of NDLP caseload, and from some of the other evaluation studies, which, for example, might have a more stringent definition relating to involvement beyond an initial interview (see Evans et al., 2002: p.29). This also applies to the definition of non-participation, which in this report, refers to all those not classified as a participant (ie. lone parents who had not had an initial interview with NDLP). This is in contrast to some other studies, in which lone parents who had an initial interview but who were not added to the caseload may be included as non-participants.

The tracking process was continued throughout the 'observation window' from August 2000 (ie. from the point the first sample was drawn) until April 2001 (when it was necessary to select the participants and non-participants for the face-to-face survey).

Although the whole sample was tracked, the main focus was on the 42,273 who had responded to the postal survey, since these were the individuals for whom the additional information to analyse propensity to participate was available.

By the time the observation window was closed, the 42,273 postal survey respondents could be divided into these three groups:

- (A) Those who had participated in NDLP before returning their questionnaire (1,209). These individuals were discarded from subsequent analysis on the grounds that their questionnaire responses would be affected by NDLP and so were not valid predictors of the propensity to participate.
- (B) Those who responded to the postal survey and then participated in NDLP during the observation window. This group of 1,787 individuals (just 4.2 per cent of the total) was selected as the sample of 'participants'.
- (C) Those who did not participate either before returning their questionnaire or within the eight month window (39,277). This group of non-participants formed the pool from which the matched comparison sample was selected.

It should be noted that the 27,578 sample members who did not respond to the postal survey were also tracked in order to observe whether they participated in NDLP. Among this group, 1,249 participated in the programme by the time the observation window was closed and 26,329 did not. For completeness, these are referred to as Group D and Group E. These cases were not selected for the next stage of the study though they were (alongside the full original sample) analysed using administrative data alone.

Table 2.1 Composition of the sample

	Participants	Non-participants
	Individuals identified as participants on the NDLP database within the study window, i.e., between August 2000 and April 2001	Individuals who were not identified as participants on the NDLP database within the study window
Cases for whom we have administrative and postal survey data	Group A (1,209) Individuals who participated in NDLP and then returned their postal questionnaire. Because of this they could not be used for the subsequent sampling	Group C (39,277) Non-participants who returned their postal questionnaire This group formed the pool from which the matched sample of non-participants was selected
	Group B (1,787) Individuals who returned their postal questionnaire and then participated in NDLP This group formed the participant sample	
	Group D (1,249) Participants who did not return their postal questionnaire. Because of this they could not be used for the subsequent sampling	Group E (26,329) Non-participants who did not return their postal questionnaire. Because of this they could not be used for the subsequent sampling

Source: NatCen draft report, Figure 2.2.

Table 2.2 Selection and response rates for the two key sample sub-groups

	Group B (1,787) The participant sample	Group C (39,277) The pool from which the matched sample of non-participants was selected
	Group B1 (1,269) Participants who responded to the survey.¹ Of these: a) 1,071 confirmed in the survey that they were participants b) The survey participation status of a further 10 could not be checked c) 188 said in the survey that they were non-participants Group B2 (517) Participants who did not respond to the face-to-face survey	Group C1 (1,253) Matched non-participants who responded to the survey.² Of these: a) 1067 confirmed they were non-participants during the survey b) For 12 their survey participation status could not be checked c) During the survey 174 said they were participants Group C2 (576) Matched non-participants who did not respond to the survey Group C3 (37,446) Non-participants who were not matched to participants

¹ This figure excludes one participant who responded to the survey but did not give consent for the interview data to be used in analysis. This case is removed from all discussions relating to survey data, but included in administrative or postal data analyses.

² This figure excludes two non-participants who responded to the survey but did not give consent for the interview data to be used in analysis. These cases are removed from all discussions relating to survey data, but included in administrative or postal data analyses.

Source: NatCen draft report, Figure 2.2.

Table H.2 summarises the different sample definitions used in the statistical analysis of later chapters. Most of the samples are simple aggregations of the categories presented in Table J.9 and Table 2.2, or slight variations using the propensity score value as an additional criteria (in order to avoid potential common support problems), such as samples IV and VI.

The samples VII to XI are the ones that address the problem of non-participants joining NDLP after the study window. Sample VII does that by dropping only the non-participants, while the remaining samples drop both the non-participants and their respective counterparts.

Different sample definitions

It is important to note that in this report impact estimates using only the information in the administrative data were also constructed by comparing the participants in Groups A, B and D with the non-participants in Groups C and E. The disadvantage of this approach lies in the inability to condition on the many covariates collected in the postal survey, a disadvantage partially mitigated by the rich information on individual labour market and programme participation histories available in the administrative data. The main advantage of this approach is that any bias which may result from non-response to the survey is avoided. In addition, this approach retains a larger sample size with which to compute an impact estimate.

Several features of the sample design raise statistical issues. These include:

- Some individuals return their postal surveys much earlier than others within the window of time allowed. The timing of the return may affect the measurement of time-varying variables on the survey, variables that, in turn, may play a role in the matching. Such changes affect the nature of the covariate alignment achieved through the matching and can thereby affect the impact estimates.
- The window for NDLP participation runs over eight months but the Group B participants include only those who return their postal questionnaires prior to participation. If the timing of the survey return in relation to participation is correlated with outcomes in the absence of participation or with the person-specific impacts from participation, reliance on the Group B1 participants may compromise the validity of the NatCen impact estimates.
- The face-to-face interviews took place over a four month period. Combined with differences in the timing of participation during the period over which it was measured, this leads to large differences among participants in the time period available for outcome measurement. The maximum is 17 months for a participant who responded to the postal survey immediately in October 2000 and had a face-to-face interview in February 2002. The minimum is just six months, for a person who started NDLP in April 2001 and had a face-to-face interview in November 2001. If the timing of participation is correlated with the impact of the programme, or with the speed with which the effect of the programme manifests itself in the labour market, then these differences among participants could affect the validity of the impact estimates.

- Although high by the relatively low standards of postal surveys, the postal survey for the NDLP evaluation experienced a substantial amount of non-response. As previously noted, NatCen utilised the postal data in selecting both the participants and the matched non-participants for the face-to-face interviews. A correlation between whether or not a lone parent responds to the postal survey and their labour market outcomes in the absence of participation, or between non-response and their impact from NDLP participation, could affect the validity of the NatCen impact estimates.
- There was also sizeable non-response to the face-to-face interviews. As with the postal survey, correlations between non-response and either outcomes in the absence of participation or the person-specific impact from the programme would compromise the validity of the NatCen impact estimates.
- The matching procedure used to select the non-participants to receive face-to-face interviews made only modest use of the detailed information on previous labour market histories available in the administrative data. If the omitted aspects of the labour market histories affect both participation in NDLP and (potential) outcomes in the absence of participation conditional on the variables included in the analysis, failure to include them will bias the resulting impact estimates.

A further issue is that the gap between the questionnaire and participation varies – the longer it is, the greater the chance that things had changed (particularly attitudes) so that participation is based on information which is not captured.

2.5 The administrative datasets

The sample used in this analysis is the same as the one used in the NatCen report, and comprises a stratified sample of the NDLP eligible population. Therefore, all the lone parents in this sample were IS claimants in August and October 2000, but not necessarily on NDLP at that time.

However, in order to estimate the effect of NDLP over a longer period of time, the NatCen work was complemented with more recent extracts from the Labour Market System (LMS) and the Generalised Matching Service (GMS).

The first extract contained the information from NDLP participants and includes all NDLP participants since the introduction of the programme. The data also include information on all Lone Parent Personal Adviser (LPPA) meetings that have taken place since the meetings were introduced in April 2001. The information for this dataset comes from the LMS database.

The remaining three extracts provide us with information on the lone parent eligible population. This gives us details of all lone parents on IS, JSA, IB or Severe Disablement Allowance (SDA). The information here comes from the GMS database. This is an historical database that provides information on claims going back to mid-1999.

All the databases used were spell-based files which contained one or more records per person. An individual will have multiple records if they: (1) receive benefits from more than one programme (IB, IS, SDA, JSA or NDLP); (2) stop receiving benefits and then receives benefits again at a later date; or (3) change their benefit amount from a specific programme.

The only other reason duplicate records could occur is if some personal detail differs, e.g., if the date of birth (DOB) is different in the two records then two different clients would be identified due to a data input error. Although this problem is minimal there are a few such cases where the client's serial number and benefit information are the same but two separate records exist.

Below the variables used in each of the data extracts are described, as well as the assumptions made when merging them into one consistent benefit and state history for each individual claimant.

2.5.1 NDLP dataset

The NDLP data extract has 36,153 spells corresponding to all NDLP enrolments and WFI meetings for the 27,431 benefit claimants who were sent the postal questionnaire and were on NDLP or had a WFI meeting up until the 22 December 2003. This shows that many of the individuals eligible for NDLP included in the NatCen study (69,829 individuals) have not yet become NDLP participants⁴.

During the data cleaning process 14,189 spells had to be dropped because 'stcasedt' was missing. The reason for this missing data is the fact WFI and NDLP spells are recoded in the same database, and there were a number of lone parents who attended a WFI but did not enter the NDLP programme.

2.5.2 IS dataset

The IS data extract had 92,540 spells from all IS claims with a Lone Parent Premium, for a total of 69,829 benefit claimants who were sent the postal questionnaire and had an IS claim before the 22 December 2003.

The start and end dates⁵ for a client being on IS did not match exactly with the overall IS claim start and end dates⁶, because lone parents' circumstances can change within an overall IS spell, they best reflect the focus of the analysis.

⁴ The variables 'stcasedt' and 'ndenddt' were used to flag, respectively, the start date and the end date of the NDLP spells.

⁵ 'lpstart' and 'maxn3'.

⁶ Which are pclmstdt, pmaxcmdt.

2.5.3 IB dataset

The IB data extract contained information on 25,151 spells, which includes all IB, SDA or IS (Disability Premium)⁷ spells for the 14,683 benefit claimants who were sent the postal questionnaire and received one of these benefits before the 22 December of 2003⁸.

The correct assumption is that if the end date is missing, the claim is still active, for IB benefits. However, for some claims, it may be found that one IB claim that has a missing end date and still appears to be live, followed by another IB claim that started on a later day. These two spells should be treated as the same spell – it seems that this type of record can occur when a new claim is created rather than just amending an existing live claim. Therefore, the two spells can be combined into a single continuous spell.

2.5.4 JSA dataset

The JSA data extract had 40,048 spells on all JSA claims for the 19,786 benefit claimants who were sent the postal questionnaire and had a JSA claim before the 22 December 2003⁹.

In general the JSA data is treated as more reliable, due to the client having to come in to attend a Jobcentre Plus office every two weeks to sign on. This regular attendance means their information is more likely to be accurate, and up-to-date, than the IB and IS data. This assumption has an important implication when addressing the overlapping problem.

Several records on the IS, JSA and IB data extracts had missing end dates. The initial assumption made was that such claims were still active at the time the record was extracted from the system. This assumption was further refined when addressing the issue of overlapping spells, as discussed in the next section.

2.6 Overlapping claims

In this section the logical problems with the data which result from merging the administrative data sets, described in the previous section, are explored. The most common problems of reconciling the spell data from the different administrative

⁷ Compared to those IS claims in the IS extract, which are just IS claims with the Lone Parent Premium.

⁸ The GMS Guide states that 'the best indicators of claim duration are claim start/entitlement start (where available) and max.claim end'. Accordingly, we attempted to track spells of benefit receipt using the date field 'bclmstdt' as the start date, the date field 'bmaxcmdt' as the end date, and 'cxben1' to indicate benefit (IB, IS, SDA).

⁹ The claim start and end date variables used in the analysis were 'sue' and 'etd' respectively.

data sources are described. By the use of actual real individual case history examples, these complexities are described and exactly what changes have been made to the final reconciled spell data are explained. All of the alterations to the data were done with the explicit guidance and advice of experts from the DWP.

A number of people appear to have overlapping spells. In most cases these are feasible admissible contemporaneous spells since people can be on NDLP at the same time as claiming any benefits such as IS, IB or JSA. A person can also claim both IS and IB at the same time. However, some overlaps are technically impossible, since no one can simultaneously receive IS and JSA or IB and JSA. All together these illegal overlapping spells were slightly less than 1.0 per cent of the total sample of 69,829 claims.

Three major possible reasons for overlaps were identified:

- Recoding of missing values or improbable dates (e.g., 1 September 2025): The standard advice is that in both cases the claim remains active at the date of the extract.
- Random imputation of end dates: Sometimes there may be an overlap as a client finishes a claim on IB or IS and moves onto JSA. This occurs because the IS and IB, DWP computer programmes update the data only once every two weeks (six weeks in the case of the IB database), and so the actual end date of a claim is not known. This means that the programme estimates an end date for claims within a two week interval. So the estimated end date may fall after the date the claimant starts to claim JSA, although they were not, in reality, claiming them both at the same time.
- Simultaneous enrolments: a third reason for overlaps is that branches used to put applications through the IS and JSA channels at the same time so that, if the claimant did not qualify for JSA, they would receive an IS payment instead (so in this case it is likely that the IS claim is the 'real' one). This is something to bear in mind when information on JSA spells earlier than the participation date are used as conditioning information or as a benchmark to compare post participation history.

Figure 2.2 illustrates three examples of overlapping spells; The first is the overlapping of IS claims. As we can be seen, the first IS claims ends (16 November 2001) after the beginning of the second IS claim (26 October 2001). This is assumed to be an example of imputation of a random end date of IS claims (within a two week period). The second example is the overlapping of the NDLP and the IS claims of these individuals. This is a perfectly acceptable overlap and does not need any amendment. The third and last overlap is the one between the JSA and IS claims, where the IS claim ends 4 October 2002 while the JSA claims starts on the 7 September 2002. This is also assumed to be due to the imputation of a random end date to the IS claim.

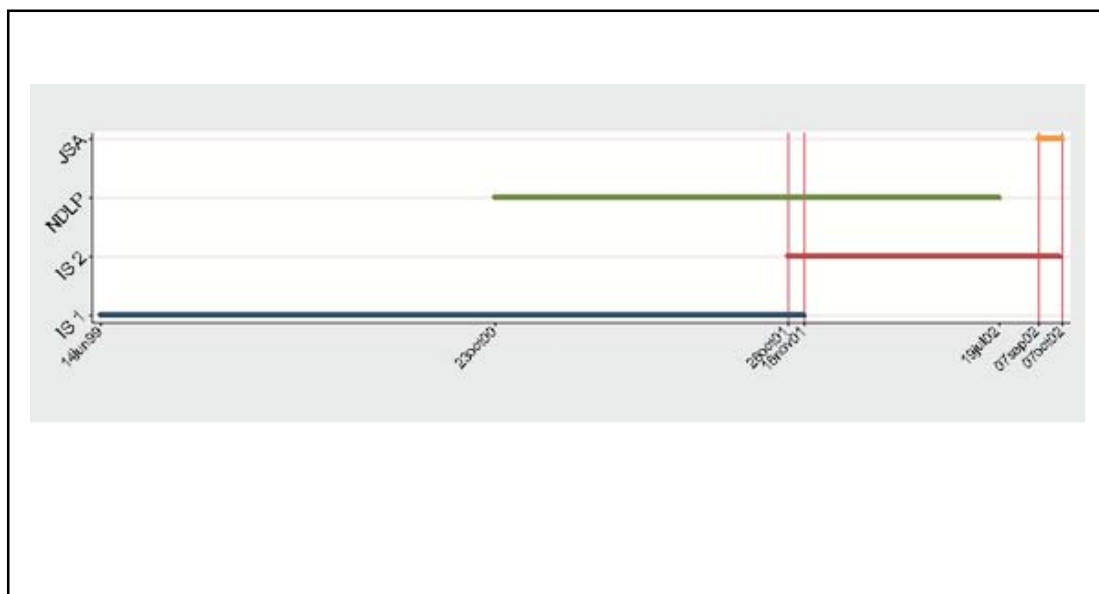
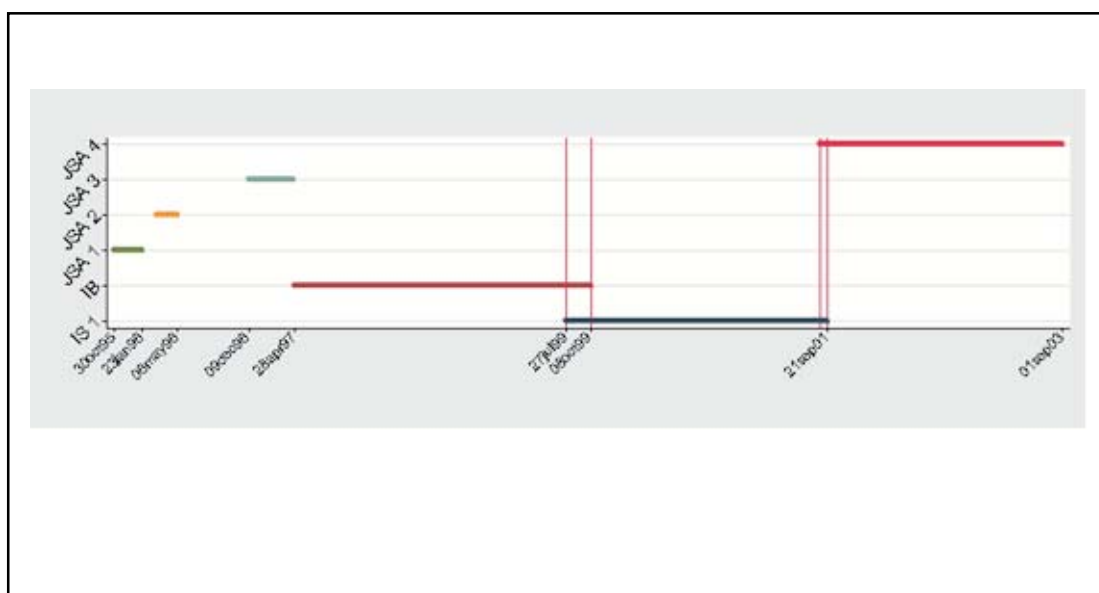
Figure 2.2 Overlapping spells

Figure 2.3 illustrates another example of overlapping claims. The first is the overlapping of IS and IB claims. As can be seen, the IB claims ends (8 October 1999) after the beginning of the IS claim (27 July 1999). This is a perfectly acceptable overlap and does not need any amendment, since the same individual can claim both IS and IB at the same time. The second example is the overlapping of the IS claim with the fourth spell of JSA, where the IS claim ends 21 September 2001 while the JSA claims starts on 28 August 2002. As noted above, the JSA information is treated as more reliable, subsequently in cases with overlapping claims on JSA and IS or IB, the JSA claim took priority over the IS or IB claim. In this case the day before the beginning of the JSA claim (27 August 2001) was used as the end date of the IS claim.

Figure 2.3 Second example of overlapping spells

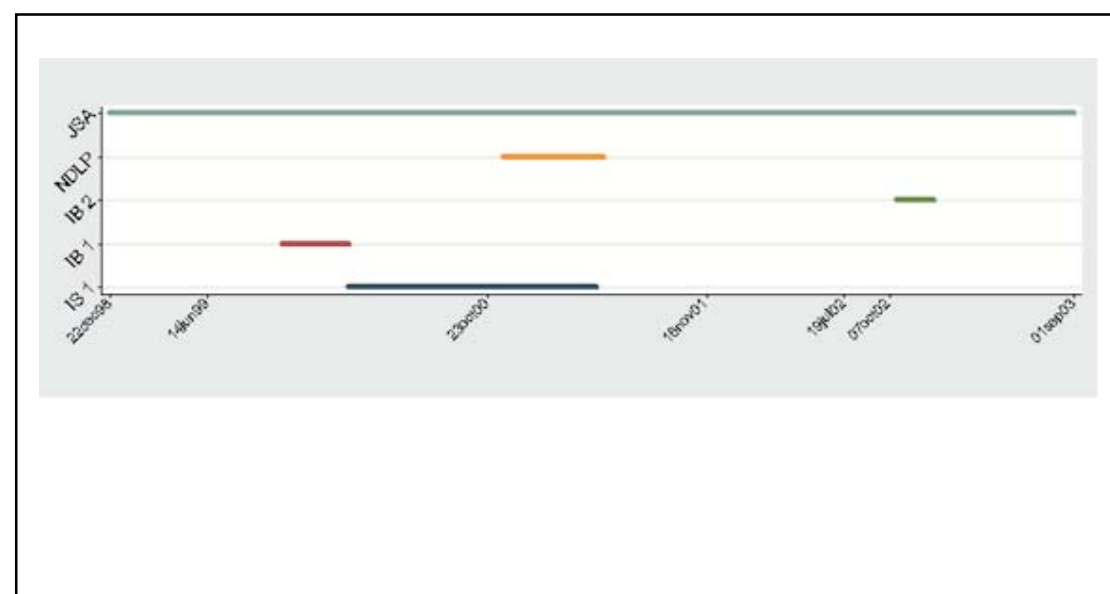
In order to overcome the problems of inadmissible spell overlaps, the raw spell data was modified in a number of ways:

- For IS claims that overlapped any other claim for a period of 18 days or less, the end date of the overlapped claim was pushed back to the day before the start date of the subsequent claim. This was done because it is known that the end dates of the IS claims are randomly imputed within approximately a two week interval.
- In cases of two IB claims, the first of which had a missing end date and still appeared to be active, and the second of which started on a later day, the two spells were combined into a single spell with the start date of the earlier spell and the end date of the later spell.
- In cases with overlapping claims on JSA and IS or IB, the JSA claim takes priority over the IS or IB claim. As mentioned earlier, JSA clients have to come in to the Jobcentre Plus office every two weeks to sign on, so their information is most likely to be more accurate and up-to-date than the IB and IS data.

The application of these simple rules resolves most cases of overlapping spells, leaving only fewer than 700 out of the total sample of 69,829 spells, or about 0.01 per cent.

Figure 2.4 illustrates one of these examples that the current procedures were not able to solve. The main reason is that the JSA spell runs from 22 December 1998 to 1 September 2003, and overlaps with all the other programmes, such as IB and IS. If the day before the beginning of the first IB spell (25 October 1999) was used as the end date of the JSA spell, the length of the JSA spell would have to be dramatically changed.

Figure 2.4 Third example of overlapping spell



3 Basic notation, the matching methodology and the NatCen model

Summary

The main objective of this chapter is to clarify the econometric model implicit in the NatCen analysis in order to make clear the precise assumptions required for their model to generate consistent estimates.

These assumptions include:

- the relatively limited set of regressors included in the New Deal for Lone Parents (NDLP) participation probability model suffices to control for non-random selection into the programme. In particular, they assume eliminating selection bias does not require matching on local labour markets or matching on detailed Income Support (IS) and employment histories;
- non-response to the postal survey and non-response to the face-to-face interview depend only on the small set of observables used to construct the sampling weights. In particular, they assume that non-response is unrelated to outcomes conditional on this small set of observables;
- participation in NDLP is what causes a lone parent to find employment rather than being offered a part-time job and possibly having a Work Focused Interview (WFI) and registering for NDLP to simply get the 'better off calculation' in order to explore whether the job is worth taking.

It is important to note that the 'full' characterisation of the underlying econometric model which would capture the NatCen data is in fact quite complicated but the version of the impact assessment model used by NatCen is relatively simple and relies on a number of strong and potentially controversial assumptions

See Appendices H-J for more in-depth explanation of the econometric model implicit in the NatCen analysis.

4 Evaluating the NatCen report

Summary

This chapter explores the robustness of the NatCen results to changes in the matching algorithm employed and in the specification of the propensity score model, conditional on using only variables from the postal survey.

Among the steps taken in this analysis, it is worth highlighting that:

- the NatCen matching was repeated using routines written in Stata rather than in BLAISE;
- a balancing test was performed using the NatCen propensity scores;
- matching was examined using only matched pairs in which both people respond;
- rematching was carried out using the face-to-face interview sample and a variety of matching methods including nearest neighbour with and without replacement and kernel matching with various bandwidths; and
- rematching was carried out using the postal interview sample and alternative propensity score specifications that employ only variables available on the postal survey.

The main finding of this chapter is that although NatCen adopted a number of approaches that are unconventional relative to the econometric matching literature, the estimates obtained are relatively insensitive to these dimensions.

See Appendices H-J for replication of the propensity score estimation performed by NatCen.

5 The problems of non-response, sample weighting and the window of participation

Summary

This chapter looks at how NatCen dealt with some of the statistical problems associated with non-response of both the postal questionnaire and the face-to-face interview, and explores a few alternative weighting procedures.

The weighted estimates that adjust for selective non-response on observable characteristics make very little difference to the estimates. For example, the weighting changes the impact estimate on the 'TONISA' variable from 34.9 weeks to 41.7 weeks (Table J.20). Despite these findings, there remains the possibility that an analysis that incorporated selective non-response based on unobserved factors correlated with outcomes might tell a different story.

Another important issue explored in this chapter is the fact that the window of time available for an individual to participate in the New Deal for Lone Parents (NDLP) after returning the postal questionnaire varies among the individuals in the sample. Taken together, the evidence presented suggests that the precise point at which the postal questionnaire is returned during the window does make a difference in the overall impact assessment of NDLP.

In this chapter we also determine the importance of conditioning on local labour markets, or variables at the local labour market level. This involves matching the evaluation data to administrative data that contain these variables. We found that adding these local labour market data did not change the results much.

5.1 Introduction

The previous chapter presented the basic estimates of the impact or treatment effect of the NDLP programme. These estimates ignore several key conceptual and statistical problems. The potential for non-response bias resulting from selective non-response to the postal and face-to-face interview samples stands out as the most important among these problems. The fact that the window of time available for an individual to participate in the NDLP after returning their postal questionnaire varies among the individuals in the sample, represents a second important potential problem. This chapter empirically addresses both these potential problems, and provides preliminary impact estimates that attempt to take account of them. This chapter also examines the sensitivity of the impact estimates to the use of supplementary exogenous conditioning data relating to local labour markets.

5.2 Non-response and weighting issues

Chapter 3 of this report set out the most general model of the evaluation of the NDLP. An important component to this evaluation problem, given the data available to NatCen, relates to non-response to the original postal questionnaire and then, subsequently, non-response to the face-to-face interview. This section explores both types of non-response in a simple way. As it is unclear *a priori* whether these two non-response problems are interrelated, a start is made by taking the simplest approach and assuming the two non-response processes are independent. The limitations of this approach are realised, and it is suggested that the issue of non-response and its effect on the representative nature of the sample used by NatCen is a major issue worthy of further study. Time does not permit this avenue to be explored further in this report, but in the authors' view, it remains one of the most important potential limitations of the NatCen analysis.

5.2.1 The NatCen weighting procedure and calculating new weights for the stata

As previously described (Section 2.3), the final sample of cases for the evaluation was obtained in stages. Weights were required for the three sources of observations: the original Wave 1 sample, the replacement Wave 1 sample and the Wave 2 Booster sample. The selection weights for each stage were calculated separately by NatCen as follows: Full details of the computation of their weights is provided in Phillips et al (2003). Here a recap of their procedure is required in order to explain how it was thought appropriate to weight the stata created by sampling on duration of claim and age of youngest child.

Original Wave 1 sample

For the original sample at Wave 1, 130 Primary Sampling Units (PSUs) were sampled with probability proportional to the number of estimated cases from a sampling frame of all eligible PSUs. Within each selected PSU, a fixed number of cases (or all cases) were sampled within categories defined by duration on benefit and the age of the youngest child.

The selection probability for each sampling cell (defined by duration on benefit, age of youngest child and PSU) was calculated by dividing the issued number of cases in that cell by the eligible number of cases and multiplying by the probability of selection of the PSU. The sampling weights were obtained as the inverse of the sampling probabilities.

Replacement Wave 1 sample

For the replacement sample at Wave 1, a further 34 PSUs were sampled proportional to the recalculated estimates of the number of eligible cases from all eligible PSUs (excluding those that had already been selected for the original sample). Within each selected PSU, a fixed number of cases (or all cases) were sampled within categories defined by duration on benefit and the age of the youngest child. The sampling weights were obtained as for the original sample, but based on the recalculated estimates of the eligible number of cases in each PSU.

New cases boost (Wave 2) sample

The Wave 2 sample consisted of new cases sampled from the 139 (eligible) PSUs selected at Wave 1. Within each Wave 1 PSU, a fixed number of cases (or all cases) were sampled within categories defined by age of the youngest child.

For the purposes of weighting for unequal selection probabilities, the Wave 2 sample was treated as if it had been obtained for the original and replacement PSUs separately. The selection probability for each cell (defined by PSU and age of youngest child) was calculated by dividing the issued number of cases in that cell by the eligible number of cases and multiplying by the probability of selection of the PSU at Wave 1. For the 'original' PSUs, the probability of selecting the PSU was calculated using the original estimates of the number of eligible cases, whereas for the 'replacement' PSUs the recalculated estimates were used.

Therefore, weights were generated (as the inverse of the selection probabilities) separately for the Wave 2 cases in the PSUs selected for the original Wave 1 sample and for those in the replacement Wave 2 sample.

Selection weights for combined datasets

In order to obtain selection weights for the complete Wave 1 and Wave 2 datasets, the selection weights were first re-scaled so that the weighted sample size for each component (the original sample and the replacement sample at Wave 1 and Wave 2) was equal to the issued sample size. This ensured that the original and replacement components were correctly represented in the datasets.

A further re-scaling was required to combine the Wave 1 and Wave 2 samples, so that the new cases in the sample correctly represented the true proportion of new cases in the population.

Stata - accounting for the different distribution of age of youngest child and duration of present claim

Having selected the PSUs, the next stage involved selecting the 65,000 lone parents, to include lone parents of different benefit durations and with different ages of youngest child. The plan was to select 500 cases from each of the 130 PSUs. To do this:

- banded variables were created for sampling purposes according to the table below;
- sample members were sorted into categories defined by these bands, and within these categories by area;
- a systematic sample was selected within each category using a given sampling fraction and a random start.

This process generated a sample of 64,781 cases. The distribution of the numbers of eligible lone parents selected and the total number of eligible lone parents by duration of claim and age of the youngest child are reported in Tables J.13 and J.19 of the NatCen technical report. A more mysterious part of the NatCen analysis is the use of specific sampling fractions for each of the duration of claim and age of the youngest child categories. The NatCen technical report provides these in Table J.16 and although they are loosely based on trying to boost the cell sizes with low frequency, they are essentially rather arbitrarily determined.

The distribution of NDLP participants and the total eligible population by the duration of their claim and the age of the youngest child is reproduced in Table J.16. Likewise, the same distribution for those in the whole sample is reported in Table J.19. These tables also show the distribution of these samples which come from the Wave 1 or Wave 2 sample and the Booster sample. These distributions are important in the calculation of the appropriate weights for the sampling proportions by age of youngest child and by duration of claim.

The final number of observations on each stata in the sample used in these calculations is detailed in Table J.16.

The last stage in the sampling and weighting procedures adopted by NatCen was to allow for the non-response in the postal survey. This was done using the weights detailed in the Table J.19.

Non-response weighting

The NatCen analyses (see Phillips *et al* (2003) Section 2.4.2) have shown that there is evidence of variability in non-response for different population sub-groups. If

uncorrected, these patterns of non-response could add bias to the survey estimates, and so further weights were calculated to allow for variable response rates.

To generate the non-response weights, the patterns of non-response were first modelled by NatCen using logistic regression. Separate models were produced for the three sample types (main stage, replacement and boost) to allow for likely differences in the profiles of respondents in the three groups.

For each model fitted, the non-response weights were estimated as the reciprocal of the predicted probability of response. The resulting non-response weights were multiplied by the selection weights previously generated and then scaled to produce the composite weights used in the analyses.

5.2.2 Non-response by NDLP participant status

Tables J.14 and J.15 cross-tabulate the two types of non-response involved in this study with participation status in NDLP. Table J.14 shows a 10 percentage point difference on the response rate of the participants and non-participants, with 70 per cent of NDLP participants responding to the postal survey, but only 60 per cent of non-participants. This ten per cent difference suggests the potential for moderate biases due to differential non-response. The Pearson Chi Squared statistics indicates that this is significant at the 99 per cent level. It is worth noting, that almost by definition, the Department will have better contact information for participants than for non-participants, making the extent and importance of this problem even harder to assess. Table J.15 reveals only minor differences in response rates to the face-to-face interview survey between participants and non-participants.

5.2.3 Multivariate analysis of non-response

Multivariate probit models of the two types of non-response will now be examined¹⁰. By estimating a separate model for non-response to each survey, the two non-response processes are implicitly treated as independent. An estimate is made of a number of different specifications using variables drawn from the administrative data. The list of variables included in one or more specifications appears in Table J.17, followed by the specifics of the five specifications estimated.

Table J.18 presents the estimated coefficients from the probit models. Several variables, including sex, age, and interactions between NDLP participation and labour market outcomes have strong, and statistically significant, effects on the probability of non-response to both surveys. These findings, especially the importance of the interactions between NDLP participation and labour market outcomes, suggest the potential for important biases in the impact estimates due to selective non-response correlated with both participant status and the outcomes of interest in the evaluation.

¹⁰ It should be noted that NatCen also weighted for non-response (see Chapter 6 of NatCen (2003b) for details).

In Section 5.3 these probit models are used to generate weights that attempt to correct for non-response in the impact estimates, and present adjusted impact estimates based on the weights in Model 2.

5.3 Impact estimates and non-response

This section presents impact estimates adjusted for differential non-response using weights derived from the multivariate probit Model 2 of non-response presented in the preceding section. This is only one of the several ways of incorporating this information on the model. An alternative to the approach employed here would make use of the weights in the estimation of the propensity scores.

5.3.1 Description of the weighting procedure

The following algorithm was used to generate the weights employed to produce the impact estimates adjusted for differential non-response. This procedure assumes that the decision to respond to the postal survey is independent of the decision to respond to the face-to-face interview survey.

Let $\pi_{1i} = \Pr(\text{Response to the postal survey} \mid X_{1i})$.

Let $\pi_{2i} = \Pr(\text{Response to the face-to-face interview} \mid \text{request to complete the survey}, X_{2i})$.

Create the product $\pi_i = \pi_{1i}\pi_{2i}$.

Take the inverse, and let $\theta_i = 1/\pi_i$.

Sum θ over all the participants to create $S_\theta = \sum_{i \in \{D_i=1\}} \theta_i$.

Create weights $\omega_i = \theta_i / S_\theta$.

Construct the estimated counterfactual for each observation using matching.

Calculate the weighted average of $Y_{1i} - E(Y_0 \mid P(W_i))$, where $E(Y_0 \mid P(W_i))$ represents the estimated counterfactual for observation i .

Note that the weights apply only to the participant observations in calculating the weighted average.

5.3.2 Weighted outcomes

Table 5.2.2.1 presents the weighted estimates that adjust for selective non-response on observable characteristics based on Model 2. These weights do not account for non-response based on unobserved factors that also help determine outcomes; the results from models (3) and (4) in Table J.18 suggests that such factors play an important role here, but an analysis that takes them into account lies beyond the scope of this report. For convenience, Table 5.2.2.1 also presents the corresponding unweighted estimates presented in column (5) of Table J.11 and Table 4.4.4.

Table 5.2.2.1 indicates that weighting to adjust for selective non-response as a function of observable characteristics makes very little difference to the estimates. For example, the weighting changes the impact estimate on the 'TONISA' variable from 40.008 weeks to 41.699 weeks. Despite these findings, there remains the possibility, supported by the evidence from models (3) and (4) in Table J.18, that an analysis that incorporated selective non-response based on unobserved factors correlated with outcomes might tell a different story.

5.3.3 The window problem

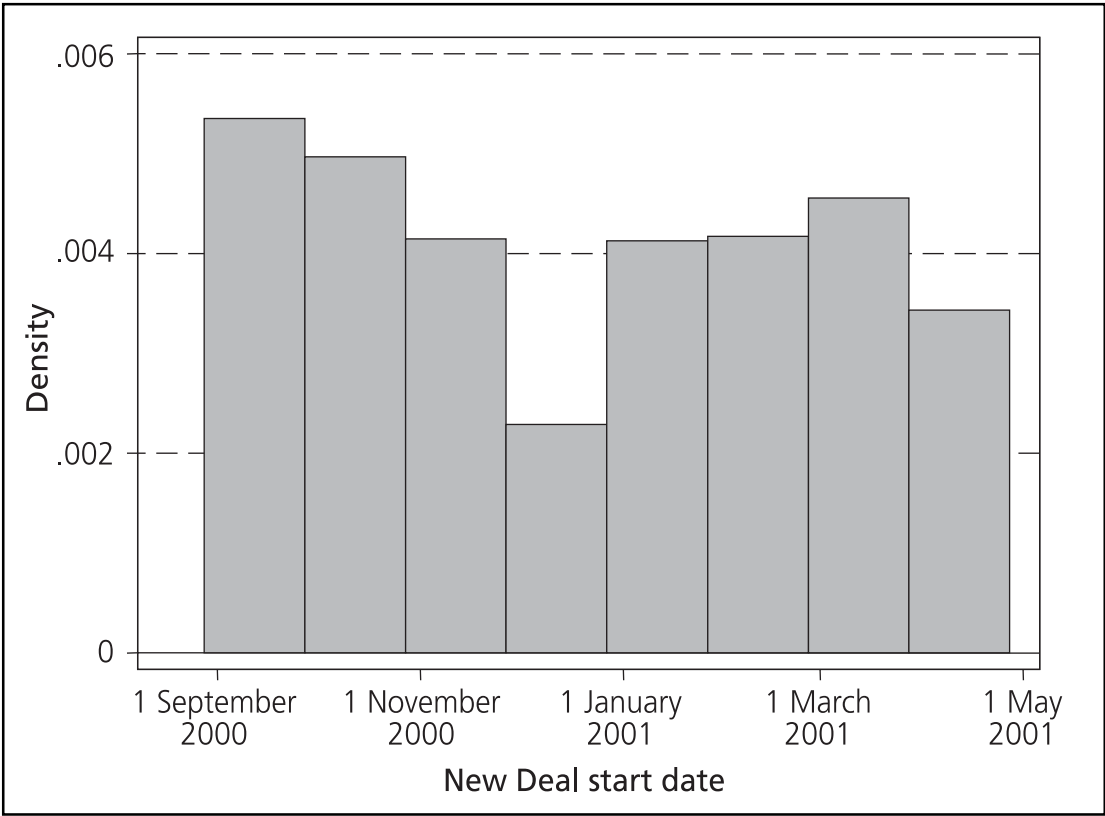
This section considers the 'window' problem – the issue of how best to define the sample in terms of the difference in dates between NDLP participation and return of the postal survey. The analysis of the window issue began by checking the syntax employed by the NatCen to implement the window definitions for the respondents categories, presented in Chapter 2.

The NatCen Group B respondents satisfied two criteria: First, they returned their questionnaires prior to NDLP participation. Second, they participated in the NDLP within the window of time between August 2000 and 28 April 2001. Figure 5.1 presents the distribution of participants by participation date. In their final report (NatCen, 2003a), NatCen shortens the window, and reduces the size of the final participant sample (Group B), by putting the end of the window at the end of February 2001 rather than at the end of April 2001. This has the effect of dropping the participants in the last two bars in Figure 5.1 from the final participant sample.

Regardless of the end date of the interval, because respondents can return their questionnaires at different times within this interval, the length of the window during which they can participate in NDLP, and still become a Group B participant, differs among people. Respondents who return their questionnaires in September 2000 have a much longer period over which to enter the programme than do respondents who return their questionnaires in January or March of 2001.

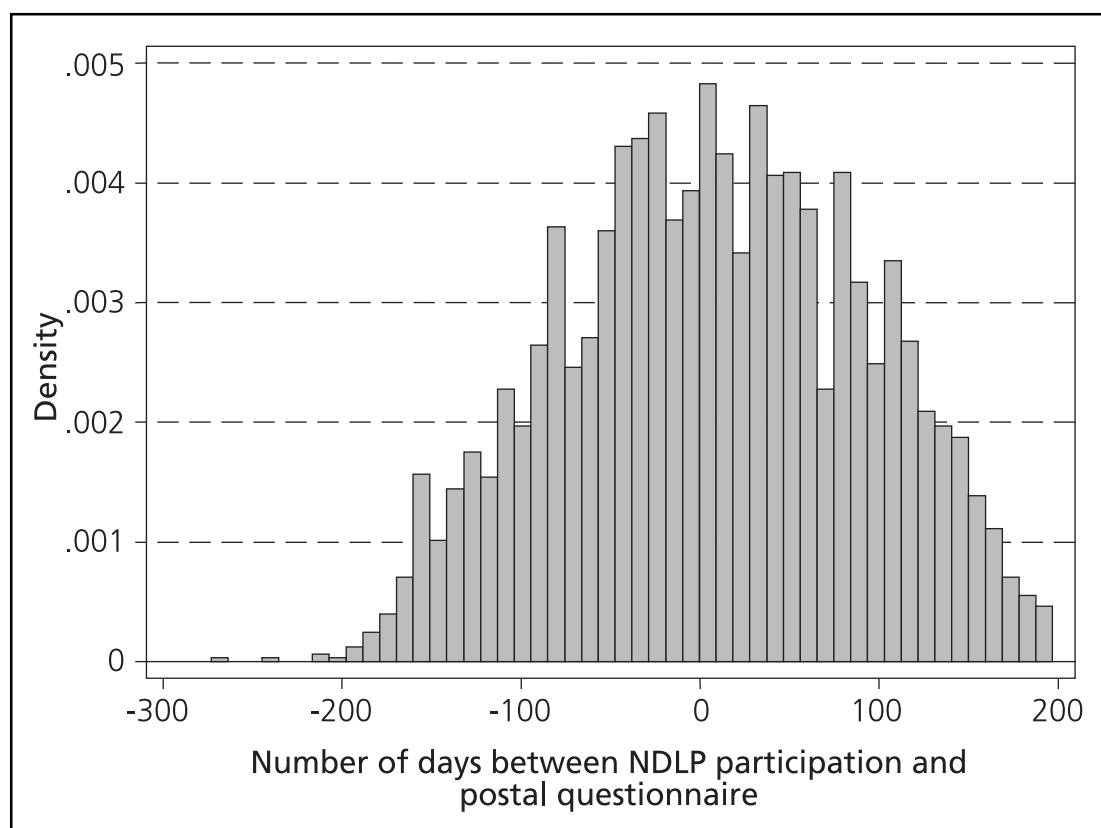
To examine the importance of this issue, impact estimates were constructed separately for groups of participants defined by the elapsed time between the return of the postal survey questionnaire and the start of participation in the NDLP. Figure 5.2 presents the full distribution of participants according to the number of days between NDLP participation and the return of the postal questionnaire. Note that positive values for the number of days correspond to persons who started in NDLP after returning their postal survey questionnaire.

Figure 5.1 Distribution of participants, by participation date



To construct the impact estimates for sub-samples defined by the number of days between and return of the postal survey questionnaire and NDLP participation, the sample was split into quintiles based on this quantity. The quintiles are calculated on the distribution after they are truncated at a value of -6 days to reflect the minimum interval defined for NatCen for Group B participants. This yields the following intervals for the postal survey respondent sample (and very similar intervals for the face-to-face interview survey respondent sample): [-6, 20], [21, 48], [49, 82], [83, 117] and [118, 197]. Separate estimates are then constructed for the participants within each quintile. The single nearest neighbour matching is used with replacement estimator for two different samples: all postal survey respondents and all face-to-face interview respondents. The estimates appear in Tables J.21, J.22, J.23 and J.24. Tables J.21 and J.22 present estimates for the full samples of postal survey and face-to-face interview survey respondents, respectively. Tables J.23 and J.24 are the corresponding tables when the samples exclude respondents with estimated propensity scores in excess of 0.3. The first column of estimates in each table refers to the full sample, while the remaining five columns present estimates for the five quintiles based on the number of days between NDLP participation and postal survey response just defined.

Figure 5.2 Distribution of participants according to the number of days between NDLP participation and postal questionnaire response



A review of the estimates shows that they do indeed differ in many cases among the quintiles. As omitting individuals with high estimated propensity scores has little effect on the patterns, the discussion is focused on Table J.20 and Table J.21. The largest differences across quintiles in Table J.21 emerge for the 'PART18B' and 'PART18C' variables, which measure IS receipt six months after the sample date and NDLP participation date, respectively, and the 'TONISA' and 'TONISC' variables, which measure the duration of the Income Support (IS) spell relative to the sample and NDLP participation dates, respectively. The estimates indicate less positive or more negative impacts for quintiles with less of a lag between survey response and NDLP participation for the incidence variables and for the IS spell duration variables. Although the point estimates differ, the same basic pattern across quintiles appears for the postal questionnaire respondents in Table J.22, particularly for the 'PART18C' and 'TONISC' variables. Taken together, the evidence presented in this section suggests that the window definition makes a substantive difference to the resulting impact estimates.

5.4 Using supplementary exogenous conditioning data relating to local labour markets

One important area for further research work involves the use of supplementary exogenous conditioning data relating to local labour markets. Since the local authority of each person in the administrative data is known, available data by Local Authority Districts (LADs) on unemployment (International Labour Organisation (ILO) definition) and deprivation is used. These data sets are available through the Quarterly Labour Force Survey (QLFS) and the NOMIS respectively and were merged into the full data. Such data provides valuable information about the locality where each lone parent lives and also proxy for deprivation and employment prospects in their location. Models estimated without conditioning on this kind of background information, assume that these variables do not affect both NDLP participation and outcomes in the absence of NDLP participation, conditional on the variables already included in the regressor set. Given findings in the literature regarding the importance of local labour markets for this type of evaluation, this assumption is a strong one.

The merge matches the postal questionnaire information with the unemployment rate of the LAD from 2001. Only 39,907 out of the original 40,964 had a correct LAD code, and could be used in the analysis. With respect to the deprivation information, the Department of the Environment, Transport and the Regions, Indices of Deprivation 2000 were used (unfortunately these indices have only been calculated to England, thus when using them, we had to drop the 7,236 postal questionnaire respondents from Wales and Scotland from the analysis).

The main deprivation variables used in the analysis were the Income Scale (number of people who are income deprived) and Employment Scale (number of people who are employment deprived). These two measures are designed to give an indication of the sheer numbers of people experiencing income deprivation and employment deprivation at district level.

For simplicity, the same subset of the outcome variables examined in our previous report are considered, excluding those generated from the face-to-face interview (Table J.25).

Table J.26 shows that although the inclusions of regional factors do make a difference on the impact, this change is relatively small.

6 Alternative impact estimations

Summary

This chapter considers a variety of alternative ways to estimate the impact of the New Deal for Lone Parents (NDLP). Outcome measures that capture not only exits from the initial Income Support (IS) spell in progress at the start of the participation window but also possible returns to IS later on among both NDLP participants and non-participants are considered. These estimates provide a deeper, and longer run, look at the impacts of the NDLP, and are complementary to the lessons provided by the NatCen report.

The symmetric matching estimator of Heckman and Robb (1985), which uses the pre-programme outcomes of NDLP participants to construct the estimated counterfactual, rather than the outcomes of the non-participants, is also considered. Although these estimates were rejected as relatively implausible, examining multiple lines of evidence gives us a deeper understanding of the data and of the underlying behaviour.

In the most important sections in this chapter, matching estimators that rely heavily on detailed information about patterns of benefit receipt in the pre-programme period are examined. The analysis began with simple cell matching on benefit history strings constructed along the lines of Card and Sullivan (1988). These strings allow the capture of both the timing and the intensity of pre-programme benefit receipt in a way that the usual alternative methods of coding this information do not. Estimates are also reported from propensity score matching based on models of participation that include both dummy variables for the benefit history strings and other variables related to pre-programme benefit receipt. Including this information facilitates a much better job of balancing the pre-programme benefit receipt patterns of the participants and matched non-participants, which, in turn, heightens the credibility of the impact estimates obtained from matching.

Continued

Perhaps surprisingly, the estimates that incorporate the extensive pre-programme benefit receipt information are not all that different than those obtained using the NatCen propensity score model, though they do tend to indicate somewhat smaller impacts. Important heterogeneity existed in mean impacts between participants more or less continuously on benefit in the pre-programme period and all other participants. The former group is found to have much larger impacts than the latter, perhaps because in the former group there was incomplete information on benefit receipt in the period prior to June 1999 necessary to effectively sort among them and thereby credibly eliminate all of the bias using matching. Evidence of heterogeneous impacts by region and by age of the youngest child were sought but strong evidence did not exist in either case.

In Section 6.2.4, the preferred estimate is presented, obtained using the full administrative sample on the 'flow' (individuals who were not on benefit for more or less the entire pre-programme period back to June 1999) and a rich propensity score specification including a wealth of information on pre-programme benefit receipt. The preferred estimate is 13.7 percentage points.

Finally, the non-parametric regression estimates are presented showing how the average treated and untreated outcomes vary with the probability of participation. This analysis indicates that individuals with the greater probability of participation have the highest mean impacts, which has implications for the targeting of NDLP within the eligible population.

6.1 Introduction

This chapter examines the NDLP impact from two different angles: Section 6.2 presents estimates of the impact of NDLP participation over time, and analyses the sensitivity of those estimates using several alternative non-parametric and semi-parametric methods. Section 6.3 examines plots of the outcomes with and without NDLP participation conditional on the propensity score, and discusses their implications.

6.2 Impacts measured over time

This section uses administrative data to compute basic non-parametric estimates of the impact of NDLP participation by comparing the proportion of the NDLP participants who are benefit recipients (i.e. Income Support (IS), Jobseeker's Allowance (JSA), or Incapacity Benefit (IB)) at a given moment in time with the proportion of non-participants who are benefit recipients in the same time period¹¹.

¹¹ The precise technical conditions under which dynamic estimates of the size of the treatment effect such as those presented here are valid, and the exact nature of the parameter to which they correspond are discussed by, among others, Abbring and Van den Berg (2003), Sianesi (2004), and Frederickson and Johannsen (2003a, 2003b).

Looking at benefit participation rates over time rather than at variables related to exit from the current spell of IS receipt has the advantage of taking into account the possible recidivism among IS participants¹². 'Recidivism', in this context refers to NDLP participants who were on IS and left it after joining NDLP but then returned to IS or another benefit at some later month. Such individuals have not only left the programme but have re-entered the 'out of work' state which NDLP was designed to help them leave. Recidivism will occur if some IS recipients find that work does not pay for them and leave jobs that they obtain through NDLP. Outcome measures that look at lengths of spells of IS, or any other benefit, receipt in progress at the time of NDLP participation or of sampling, explicitly ignore possible future spells, as do the life tables in the NatCen (2003a) report related to these measures¹³. In formal terms, these measures treat the first state following IS exit as an absorbing state, from which no one ever returns to benefit. Measures that look at whether an individual ever left benefit within a particular time frame similarly ignore the potential for returns to benefit. Measures that look at outcomes in only the first six months after participation are likely to allow too little time for recidivism and not enough time for non-participants to find jobs on their own, thus substantially overstating the long-run (or even the medium-run) impact of the programme.

This section examines the proportion of both the NatCen participants and non-participants who are benefit recipients. The chosen unit of time adopted for this analysis is weeks, although the data on spells are reported in terms of precise daily dated events. The problem with the daily information is that a substantial fraction of the spell end dates are imputed. In addition, the administrative units used in the consideration of benefits are weeks. On the other side, performing the analysis in terms of claimant weeks means that there is no need to translate the weekly data into monthly start and end dates. As the claim start and end dates are recorded in days, the choice of weeks creates a few falsely overlapping claims. In order to solve this problem, a truncating criterion is necessary; the rule followed is to recode the claim with the smallest number of working days (≤ 2) in the overlapping week to end in the previous week. The total period of the analysis goes from 1990 to 2003; however, most of the focus of this section will be from 1997 onwards.

Before moving on to the analysis of these figures, it is important to reiterate that the definitions of NDLP participant and non-participant used throughout this report are the same ones used in the NatCen study. Their definition of an NDLP participant is anyone in the eligible sample who started NDLP on any date on or after August 2000 up to the end of April 2001.

¹² This issue is further explored in Section 7.4.

¹³ It must be acknowledged that the Department for Work and Pensions explicitly asked NatCen to investigate additionality in this measure, because their standard report format presents gross outcomes in terms of job entries.

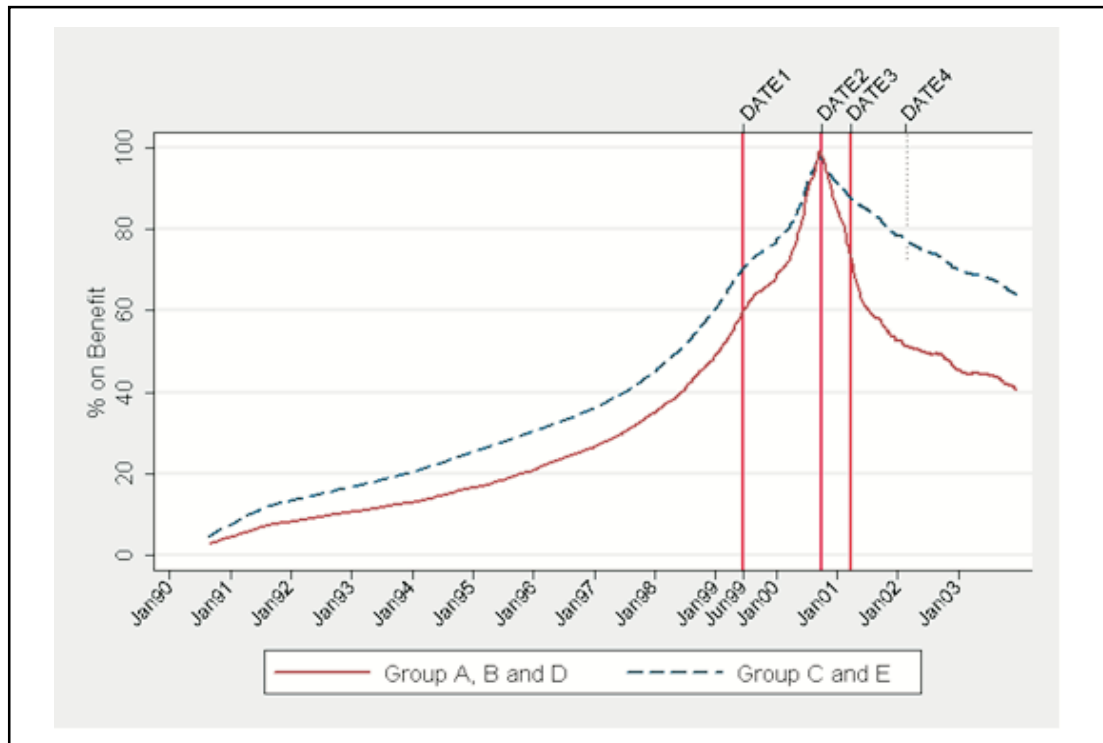
Figure 6.1 shows the proportion of benefit recipients from 1990 through the 53rd week of 2003, for the full samples of NDLP participants and non-participants. Before discussing the substance of the figure, it is important to flag four dates:

- DATE 1 is the first week of June 1999. This is an important date relating to data quality as this is the first month in which the database provides accurate figures for both the stock and flow of individuals on benefit¹⁴.
- DATE 2 and DATE 3 represent the postal questionnaire interval from October 2000 to April 2001. The proportion of the sample on benefit reaches 100 per cent for both NDLP participants and non-participants at the beginning of October due to the eligibility criteria for the postal questionnaire used by NatCen.
- DATE 4 is the first week of February of 2002, which is the end date for the NatCen analysis.

The Generalised Matching Service (GMS) dataset, from which the data is drawn, contains information on IS and IB receipt from June 1999 onwards. In addition, it includes information on JSA receipt for the same period, as well as information on the duration of all claims in progress as of June 1999. That is, JSA claims that start before June 1999 and continue at least until June 1999 are included, while JSA claims that start and end before June 1999 are not. In general, the Department for Work and Pensions (DWP) does not make use of the information on JSA receipt prior to June 1999 due to the systematic measurement error resulting from the omission of spells ending prior to June 1999, and recommends that others not use this information as well. However, in this analysis the interest is in comparing the NDLP participant and non-participant groups and in using the information on this early JSA receipt to construct conditioning variables. As both groups are affected in the same manner by the partial lack of information prior to 28 June 1999, analyses using these data will still provide some insight.

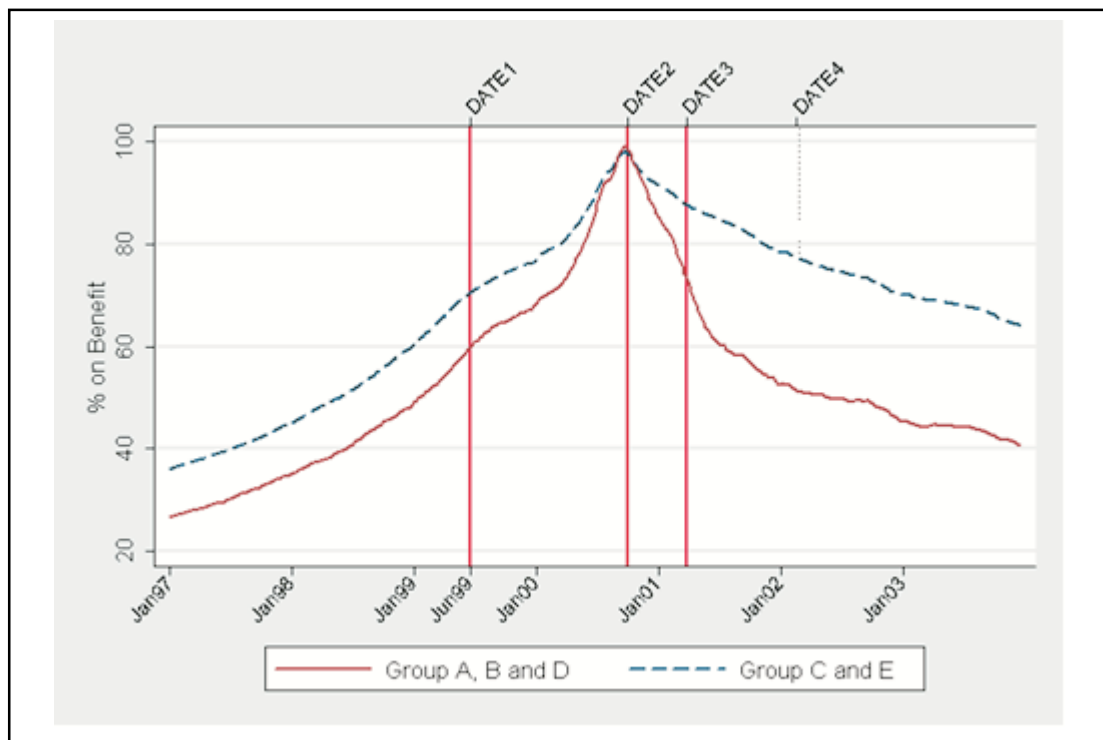
¹⁴ The only reliable benefit receipt information available before June 1999 comes from the JSA data base; for all of the other benefits (i.e. IS and IB), complete data are available only after June 1999.

Figure 6.1 Weekly proportion of participants and non-participants on benefit – 1990-2003 (Sample I – Full)



For ease of viewing, Figure 6.2 presents only the portion of Figure 6.1 corresponding to the period from 1997 onwards. Figure 6.9 presents the same information, but using only the respondents to the postal survey.

Figure 6.2 Weekly proportion of participants and non-participants on benefit – 1997-2003 (Sample I – Full)



Figures 6.2 and 6.3 display three important patterns: First, by construction, everyone in both figures receives benefits on DATE 2 in October 1999. Second, benefit receipt rates for both participants and non-participants fall fairly dramatically in the period after October 1999, with the rate for the NDLP participants falling substantially more than that for the NDLP non-participants. This crude comparison suggests that NDLP is highly effective at getting individuals off benefit, despite any recidivism that may be present. Of course, this simple non-parametric comparison of rates does nothing to account for possible non-random selection of more motivated or more employable individuals into NDLP from among the IS recipient population. Third, it can be seen that the likelihood of benefit receipt not only differs between NDLP participants and non-participants after the participation window, it differs before it as well, with future NDLP participants much less likely to be on benefit throughout the period prior to October 1999. This pattern strongly suggests the presence of non-random selection on factors that affect outcomes both before and after October 1999, with the implication that the simple non-parametric comparison of NDLP participants and non-participants presented in these figures overstates the causal effect of NDLP participation on benefit receipt.

It is important to highlight the relationship of Figure 6.2 with Table 7.2.1 from the NatCen report (NatCen, 2003a). In this figure, the interval between April 2001 and February 2002 represents the nine month period considered in the NatCen report. Within that interval there are three differences worth noting: First, NatCen defines benefit receipt solely as IS receipt, while this analysis defines the benefit receipt variable as any of IS, JSA or IB receipt. It is believed that this wider definition better captures the dependency that NDLP is designed to reduce. Second, the NatCen measure is the cumulative percentage of participants having leaving IS, while this analysis uses the fraction on benefit in each week. Thus, as noted earlier, this measure takes into account the possible recidivism among benefit recipients by counting individuals who return to benefit receipt after a spell of non-receipt. Third, this analysis presents information on the four years prior to the participation window. As is discussed in the remainder of this sub-section, this information provides a much deeper understanding of the similarities and differences between these two groups. Finally, it should be kept in mind that in this analysis, as in the NatCen analysis, individuals in the non-participant group, who by definition did not participate in NDLP during the participation window, can and do participate in NDLP later on.

Figure 6.3 Weekly proportion of participants and non-participants on benefit – 1997-2003 (Sample III – Matched)

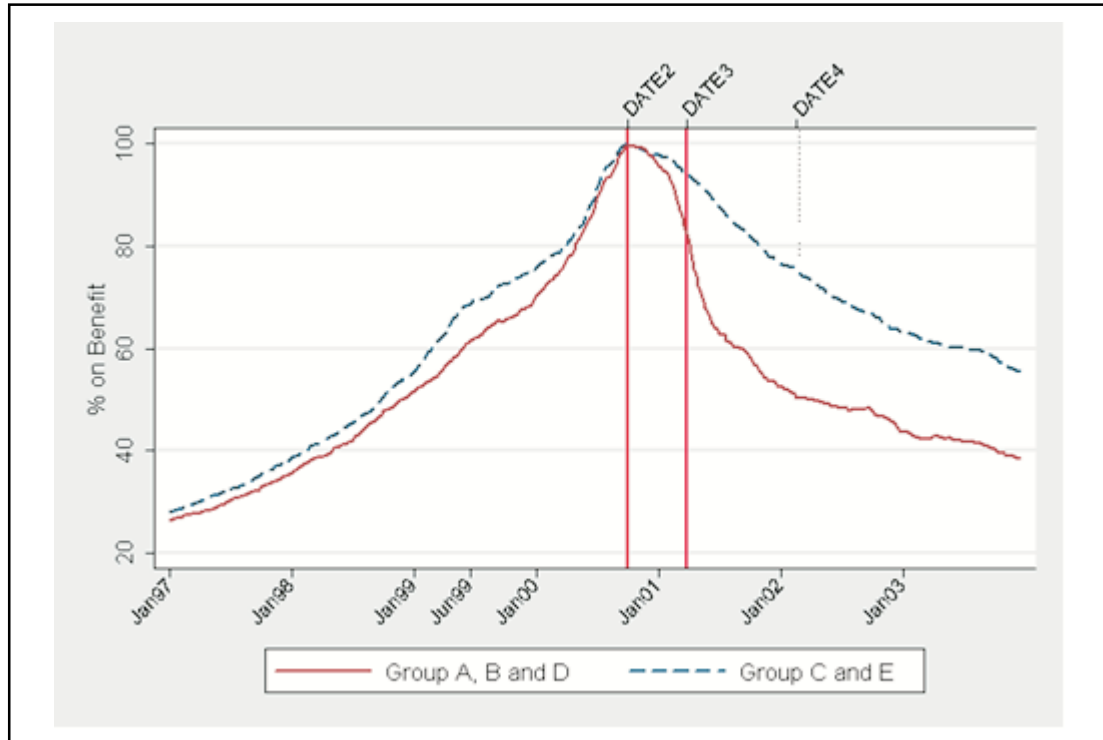


Figure 6.2 presents the same outcome variable over the same time period as Figure 6.3, but for the matched sample consisting of the 3,596 individuals who responded to the postal questionnaire and were selected for face-to-face interviews by NatCen using their propensity score matching scheme. As discussed in detail in Chapter 5, the matching takes care of any non-random selection into the NDLP that depends on the observable characteristics included in the NatCen propensity score model. Figure 6.7 presents a similar graph but using the full face-to-face interview sample (including respondents and non-respondents), rather than just the matched sample. This figure illustrates one of the main advantages of using administrative, namely, no non-response problems. Since Figure 6.7 tells the same qualitative story as Figure 6.2, there is no need to discuss it in detail here.

Three important results emerge from Figure 6.2: First, the conditioning on the NatCen propensity score reduces the differences between the participants and the non-participants both before and after the participation window. In particular, the difference in the period prior to October 1999 after matching, shown in Figure 6.2, is typically less than half of what it was prior to matching, in Figure 6.3.

Second, conditioning on the observables factors included in the NatCen propensity score matching does not completely remove the difference in rates of benefit receipt in the period prior to October 1999. Comparing Figure 6.3 and Figure 6.2 shows that although the difference between the two groups is smaller in the latter, it has not been fully eliminated.

The third and last result is that the difference in benefit receipt rates between NDLP participants and non-participants does not dissipate from February 2002 to December 2003. To the extent that this difference represents the causal effect of NDLP participation, that effect persists for some time following participation.

The remaining pre-programme difference revealed in the figures for the matched samples can be thought about in four different ways: First, it can be ignored, and it can be assumed that even if the matching does not fully account for non-random selection in the period prior to June 1999, it does so afterwards. In this case, the mean difference in benefit receipt in the 'after' period consistently estimates the causal effect of NDLP participation. Such estimates will be presented shortly.

Second, the pre-programme period difference can be treated as a balancing test of the sort described in Chapter 4. In this case, the NatCen propensity score model fails the balancing test because it fails to balance pre-programme benefit receipt, a variable that surely affects both NDLP participation and later benefit receipt outcomes in the absence of participation. Random assignment, as in Dolton and O'Neill's (2002) analysis of the RESTART treatment, balances all observed and unobserved covariates. Matching should similarly balance the observable characteristics affecting both participation and outcomes in the absence of participation; if it had in this context, then the lines in the figure for the two groups would be identical prior to the census point of October 2000 when all of the administrative sample was eligible for the NDLP. Alternative matching schemes that improve the balance in the pre-programme period are investigated in Section 6.3.3.

Third, it can be interpreted as a specification test. In particular, it is a version of the 'pre-programme' test employed in Heckman and Hotz (1989). In this test, an estimator is applied to participants and non-participants in the pre-programme period, and the resulting estimate compared to zero. If the estimate obtained from the data does not equal zero, then the estimator is rejected, as in the case here.

Fourth and finally, it can be assumed that, conditional on the observable characteristics embodied in the propensity scores, there remain some fixed unobserved differences in the rate of benefit receipt between NDLP participants and non-participants. If this difference in unobservable characteristics affecting selection into the NDLP is fixed over the period covered by the data, then the difference in the pre-programme period provides a consistent estimate of the difference in the post-programme period. In that case, a consistent estimate can be obtained of the causal effect of NDLP participation by subtracting the difference in benefit receipt rates prior to the participation window from the difference after the end of the participation window. Estimates constructed in this way are presented shortly.

However, what is also evident is that the set of observable factors used on the propensity score matching suggested by the NatCen was not able to fully address the two groups' heterogeneity. Again, the comparison between Figure 6.1 and Figure 6.2 shows that although the difference between the two groups is smaller, in the latter, it has not been fully eliminated. One possible, non-parametric, solution

for this problem could be to simply subtract the average difference between the two groups prior to the treatment, from average difference after the treatment. In this case the average treatment effect would drop from 20.52 per cent to 16.12 per cent (Table J.25 – Sample III).

The third and last result is the fact that the NDLP effect does not dissipate from February 2002 to December 2003, remaining relatively constant throughout the entire period.

There are distinct advantages to the use of detailed administrative data in the evaluation of the Active Labour Market Programme (ALMP). By using weekly outcomes, the problems with the outcomes from the NatCen analysis discussed in the preceding paragraph are avoided. By using the full postal survey sample, any problems with non-response associated with the face-to-face interview survey, are avoided.

Table J.27 shows the NDLP participation effect, illustrated by the figures above. Different impact definitions have been used, based both on non-parametric and semi-parametric non-participant groups. The tables present the average treatment effect, which is the difference between the average proportion of weeks that participants and non-participants spent on benefit after the treatment window, as well as a measure on the remaining heterogeneity between these two groups before the treatment window and the difference between the two.

The assumption here is that the difference between participants and non-participants during the same interval prior to the treatment window is a measure of the unadjusted difference in outcomes between the two groups and should be removed from the treatment effect measure¹⁵. In this case the treatment period goes from the first week of March 2001 to the 53rd week of 2003, or an interval of 146 weeks. The period used to measure the remaining heterogeneity between participants and non-participants is an inverse image of the same interval, starting in April 2000 and moving back 146 weeks, until June 1997.

A significant result is that although the raw treatment effect measure ranges from 18.58 per cent to 26.62 per cent, depending on the sample definition used, the difference between the treatment effect and the remaining heterogeneity – the net treatment effect – only ranges from 11.84 per cent to 15.84 per cent, suggesting a slightly smaller variance. The additional important point is that there seems to be very little evidence of much difference in the size of these net treatment effects

¹⁵ Heckman and Robb (1985) implement a symmetric difference estimator that justifies the estimation of the bias in an after period with the difference in the symmetric before period and Heckman, LaLonde and Smith (1999) develop a symmetric Ashenfelter's dip that would also justify a symmetric differencing.

according to which dataset is used. This is preliminary evidence that it may be the case that similar results relating to the size of treatment effects can be gleaned from the innovative use of register administrative data as can be estimated from the use of costly face-to-face survey data. This finding deserves further careful consideration.

6.2.1 Symmetric differences

This subsection, presents estimates using the symmetric differences estimator of Heckman and Robb (1985; 1986). This estimator uses the outcomes of participants in the period prior to participation as the counterfactual for what they would have experience in the period during and after participation, had they not participated. In this sense, it resembles the standard 'before-after' estimator. However, the symmetric differences estimator uses the outcomes from before period in a very particular way. In graphical terms, it rotates or reflects those outcomes around the participation date and then obtains impact estimates as the vertical distance between the observed outcome of participants and their rotated or reflected pre-programme outcomes in each period. Thus, if period k is the participation at which individuals choose to participate or not, then the observed pre-programme outcome in period $k-1$ becomes the estimated counterfactual outcome for period $k+1$, the observed outcome in period $k-2$ becomes the estimated counterfactual outcome for period $k+2$ and so on.

This simple estimator has two motivations: The first motivation derives from a straightforward theoretical model of participation and outcomes. Suppose that outcomes in each period (which could be a continuous outcome such as earnings, or the latent index underlying a discrete outcome such as employment) evolve over time with an autoregressive error term of order one, denoted $AR(1)$. In this case, in each period the stochastic component (the 'error term', though it is not error here, but rather the correctly observed deviation of the outcome from its expected value) of the outcome is the weighted average of a combination of a new, independently and identically distributed (IID) component and the stochastic component from the previous period. The weight on the stochastic component is the serial correlation. The stochastic components of outcomes in this simple model are serially correlated and they are also mean-reverting. This means that, on average, individuals with outcomes below their expected value in one period will have outcomes below their expected value in the next period, but not as far below. They stay below because of the serial correlation and they get closer to the expected value because of the random component.

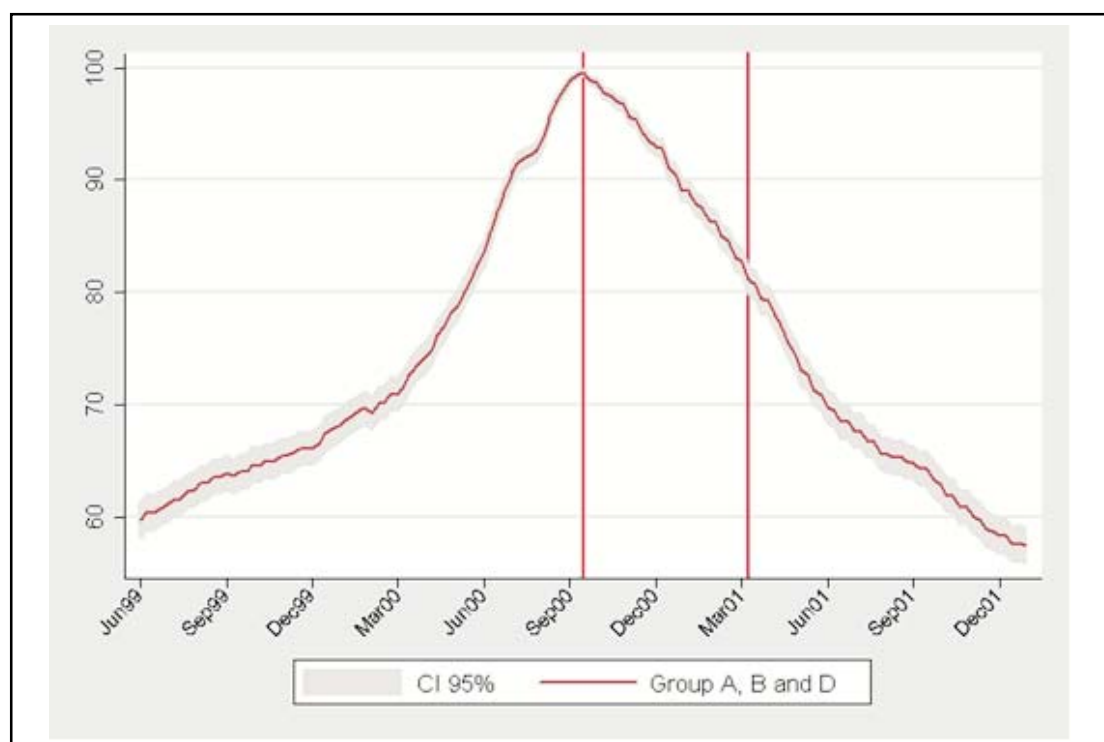
Now combine this model of outcomes with a model of participation that says that individuals participate whenever their outcome gets below some cut-off value c , which is less than their expected value. The combination of this model of participation with the model of outcomes just described has strong implications for what outcomes look like over time for a sample of individuals who participate in a particular period. Specifically, they imply that participants will experience a 'dip' in their outcomes in the periods prior to participation. This dip results from the fact that, in order to reach the cut-off level for participation, participants must have

experienced a series of bad draws on the IID component of their outcomes. In the period after the participation period, the mean untreated outcome experiences a symmetric rise, due to the mean reversion inherent in the AR(1) outcome process. Because the dip and subsequent rise are symmetric relative to the participation period k , the mean pre-programme outcomes of the participants provide unbiased estimates of the outcomes they would have experienced in the absence of participation in the post-participation period.

The second motivation for this estimator is an empirical one. Heckman and Smith (1999) present graphs of the mean earnings over time of the experimental control group from the US National Job Training Partnership Act Study. They find that the control group experiences a dip in mean earnings in the period prior to random assignment and a rise in the period after random assignment. For most groups, the rise is a bit quicker than the dip, though this may result from measurement issues rather than behaviour, due to the alignment of the survey reference period with the random assignment date. Despite this, their evidence is surprisingly supportive of the model, especially for the 12 to 18 months on either side of random assignment.

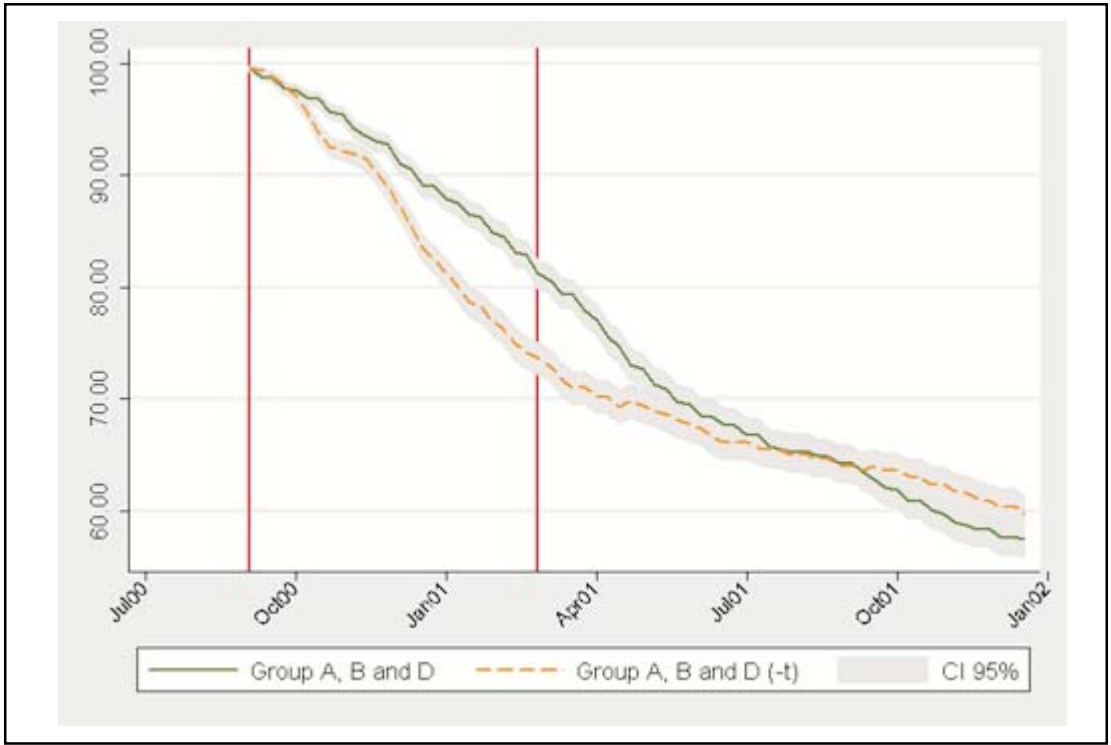
Figure 6.4 shows the proportion on benefit throughout the entire period used in this analysis. For the symmetric difference estimator only data as far back as June 1999, can be used because this is the earliest month in the database that provides accurate figures for both the stock and flow of individuals in benefit. This limitation gives a pre-programme period of 67 weeks, which in turn enables the estimation of the impact of NDLP participation using the symmetric differences estimator for the period up to December 2002.

Figure 6.4 Weekly proportion of participants on benefit 1999-2002 (Sample I –Full)



The symmetric differences estimator is applied to the full sample in Figure 6.5 and to the postal sample in Figure 6.9. Similar findings emerge in both cases. Using this estimator, NDLP participation increases the proportion on benefit for the first few months, before the observed benefit receipt rate of the participants and their estimated counterfactual cross-over. After the two lines cross, the estimator indicates that the NDLP reduces the proportion on benefit for the last few months of the sample period, though this difference is not statistically significant. It is not uncommon in the literature for employment and training programmes to increase the proportion on benefit initially, particularly if they provide services such as classroom training, which often reduce job search in the short run. At the same time, the strength and duration of the effect is somewhat surprising in the NDLP context, given the strong NDLP focus on employment, and given what is found using other econometric estimators.

Figure 6.5 Symmetric estimator of the weekly proportion of participants on benefit 1999-2002 (Sample II – Postal)



6.2.2 Exact matching on IS receipt histories

This subsection considers another simple alternative impact estimation strategy. The substantive motivation for this strategy is poor balance in the proportions on benefit between the participants and matched non-participants using the matches based on the NatCen propensity scores revealed in Figures 6.1.4 and 6.1.5. The econometric motivation is a key early paper in the econometric evaluation literature by Card and Sullivan (1988). This strategy matches participants and non-participants on the basis of their detailed pre-programme benefit histories, and thereby leads to a much better balance on this important factor, which strongly correlates with both participation and, it is suspected, with outcomes in the absence of participation.

The key to this method lies in finding a rich, yet relatively parsimonious, method to summarise the benefit receipt histories. In the spirit of Card and Sullivan (1988), the following approach is adopted: First, the period from June 1999 to September 2000 (the period over which complete data on benefit receipt is held) is broken into six 11 week 'quarters', where the final week just prior to the start of the participation window is omitted. A dummy variable is coded for each quarter that indicates whether or not the individual spent at least half the period on benefit. The six different dummies are then concatenated 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.

There are two standard alternatives to this scheme: a variable or variables indicating the fraction of time on benefit in the pre-programme period and/or a variable measuring the duration of the spell in progress at the start of the NDLP participation window. The method of strings employed here has important advantages relative to both standard alternatives: 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-programme 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 spell in progress at the start of the participation window, the benefit history strings have the advantage of capturing many additional spells during the pre-programme period.

Two important decisions arise in implementing the benefit history strings in this context. The first concerns how finely to partition the pre-programme period. Each additional sub-period doubles the number of possible strings, because it adds an additional dummy variable. Adding additional sub-periods uses up degrees of freedom and also raises the possibility of common support problems due to strings with participants but no non-participants. On the other hand, at least up to some point, adding additional sub-periods increases the plausibility of the conditional independence assumption that justifies the matching.

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 to be coded to one. Setting this value high means that short spells do not count; for example, if the cut-off value is set at 10 of the 11 weeks, then an individual 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 the cut-off value is set 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. The 5.5 week cut-off was chosen as a compromise, keeping in mind that few individuals have more than a couple of spells over the entire pre-programme period and that the vast majority of spells last at least a couple of months.

This analysis consists of four tables: Table J.28 for the full sample, Table J.29 for the postal sample, Table J.29 for the full sample without the 111111 row and Table J.30 for the postal sample without the 111111 row. The first five columns of each table present the benefit history string for that row, the number of non-participant observations with that string, the average proportion on benefit among the non-participants with that string, the number of participant (treated) observations with that string and the average proportion on benefit among participant observations with that string.

Several important findings emerge immediately: By far the most common string among both participants and non-participants is 111111; that is the modal benefit history string in both groups represents more or less continuous benefit receipt. If complete data on earlier periods was available, it would clearly be worthwhile to add additional sub-periods so as to better distinguish among individuals in this group. There will be an attempt to do this in other ways in the next subsection. 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 the 111100 string the new spell of benefit receipt starts in the omitted week before the start of the participation window). These strings represent interrupted spells.

Constructing impact estimates using cell matching on the benefit history strings is reasonably straightforward. For each cell, the cell-specific impact is constructed as the difference in the proportion on benefit between the participants and non-participants in the cell. These differences appear in the column labelled 'TT' in each table. The weight for each cell is then calculated; these weights appear in the column labelled 'WEIGHT'. As the interest lies in the average impact of treatment on the treated, rather than the average treatment effect in the population, the weight for each string consists of the fraction of the participant observations with that string. Each cell-specific treatment effect can be multiplied by its weight and the results appear in the column labelled 'CONTR' (for contribution). Summing these yields, the

cell matching impact estimate of the average impact of NDLP participation on the NDLP participants presented in the lower right corner of each table.

For the full sample, the estimated impact of NDLP participation is a -17.61 percentage point reduction in benefit receipt, while for the postal sample it is -18.66. This suggests a slight upward bias (in magnitude) due to non-response to the postal survey. What is perhaps most surprising is how similar these estimates are to the cross-sectional estimates from the matched sample using the NatCen propensity scores presented in Table J.25. Controlling for only pre-programme benefit history strings gives very close to the same estimate as controlling for the wealth of variables from the postal survey. This shows that the benefit histories, when properly coded, contain a substantial amount of information about both the probability of participation and about outcomes in the absence of NDLP participation. Put differently, the benefit history strings appear to remove a substantial amount of selection bias. This in turn suggests that relying on administrative data may represent a viable, and relatively inexpensive, alternative to using survey data, with its attendant problems of selective non-response.

A comparison of the impact estimates on the full samples with the corresponding estimates for the samples with the 111111 individuals removed shows that the estimated impacts for the individuals on benefit more or less continuously have a much larger impact, on average, than individuals with the remaining values of the benefit history strings. Put less formally, the stock has a larger impact than the flow. This difference has two possible sources: It could be that there was simply a failure to distinguish enough among the individuals in the 111111, 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 much 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. This latter view has important implications for the future impacts that can be expected from NDLP.

6.2.3 New propensity score specifications

This section explores the effect of new propensity score specifications, giving special emphasis to the variables available at the administrative dataset.

All the counterfactuals were constructed with the nearest neighbour matching with replacement algorithm on PSMATCH2¹⁶. It is important to remind the reader that the NatCen matching algorithm did not allow for replacement.

¹⁶ PSMATCH2 is a Stata ado file written by Edwin Leuven and Barbara Sianesi (2003). (available for download from <http://ideas.repec.org/c/boc/bocode/s432001.html>). Note that PSMATCH2 was not available at the time of the analysis of Chapter 4 and 5.

Three main propensity score specifications were used:

- 1 The first only took into account the gender, age, and region. All three variables were categorical. With age being divided on 10 categorical variables one for each five years interval between 16 to 64 inclusive, and region being a 12 categories, 10 for England, plus Wales and Scotland.
- 2 The second specification also includes three extra administrative variables, namely: the number of children, age of the youngest child and disability.
- 3 The third, and last, specification takes into account the pre-programme history explored on Section 6.2.2. Three new sets of variables were used on this dimension. The first one was the pro-programme regimes, created by the Card and Sullivan (1988) cell matches; in total 45 dummies were created¹⁷, one for each regime. As has already been noted in the previous section, over 50 per cent of participants and non-participants were found to be in the stock category (i.e. 111111). Moreover, the NDLP effect on the stock is remarkably different from the flow. In order to take this fact into account, and better distinguish the pre-programme history of the stock from the flow, the third specification also included variables that could capture the most recent and the oldest pre-programme history. For the former, a set of dummy variables for six weeks prior to the participation window was used; for the latter, a continuous variable which denoted the proportion¹⁸ of time that each individual was on benefit prior to June, 1999, was included.

Table J.32 shows the p-values of the joint significance of the main coefficients used in the analysis. As can be seen, most categories were jointly significant, with the exception of the Card & Sullivan (1988) cell match dummies. This result is not particularly surprising, given the probable high co-linearity of these variables with the other proxies for the prior benefit history of the population. For the descriptive statistics of the variables used on the matching algorithm as well as a table with the coefficients of the three specifications using the administrative dataset, please see the Appendix. All the analysis with the administrative data was carried out with both the full set of participants (Groups A, B and D from Table J.9) and only the Group B participants.

¹⁷

All categories with less than 20 observations per cell were pooled together under one single category (i.e. 222222). In total, there were 19 cases that fitted this criteria, representing 68 individuals, or less than 0.01% of the full sample.

¹⁸

An alternative specification using a polynomial of the number of weeks on benefit was also tried, but the proportion of time on benefit, had a better effect on balancing the non-participant group on the pre-treatment period.

Matching by Stata

Another important methodological difference from the present work and the NatCen report is how the survey design was incorporated in the analysis. As described in Section 2.3 the survey sample was stratified by including lone parents of different benefit duration and with different ages of the youngest child. In this report, a choice was made to estimate separate point estimates for each stratum, forcing NDLP participants to be matched only to non-participants in the same stratum. This yielded 24 different point estimates which were subsequently weighted according to the proportion of NDLP participants on each stratum (in the case of the average treatment effect on the treated).

Figure 6.8 and Table J.32 illustrate the effect of the use of the stata information. This figure plots the weighted proportion of benefit receipt of both participants and non-participants. As can be noted the simple use of the stata removes 50% of the pre-programme heterogeneity (Table J.25 and Table J.29). This figure should be compared with Figure 6.2, which shows the same information, however, not weighted by stata.

Administrative

Figure J.9 presents the balance test of the propensity score matching which uses only age, gender and region. The NDLP participants are groups A, B and D (in other words all the NDLP participants during the participation window). As can be seen, the pre-programme history between participants and non-participants is far from matched, indicating the use of an insufficient number of covariates.

Figure J.10 shows the balance test of the propensity score matching using the previous specification, plus information on the number of children and disability (both obtained from the administrative dataset). The NDLP participants are groups A, B and D. The inclusion of both children and disability information, does improve the quality of the pre-programme matching, in particular in the weeks that preceded the treatment window.

Figure 6.6 shows the balance test of the propensity score matching using the previous specification, augmented with the Card and Sullivan (1988) pre-programme history and the recent and long history. The NDLP participants are groups A, B and D (in other words all the NDLP participants during the participation window). In this case, it is clear that the quality of the matching improves dramatically, with the pre-programme benefit history being quite similar for both participants and non-participants throughout the entire interval.

Figure J.11 shows the balance test of the propensity score matching using the preferred specification (specification 3), however, in this case the treatment group excludes groups A and D. This figure, shows the same improvement on the pre-programme matching, suggesting a better matching estimator.

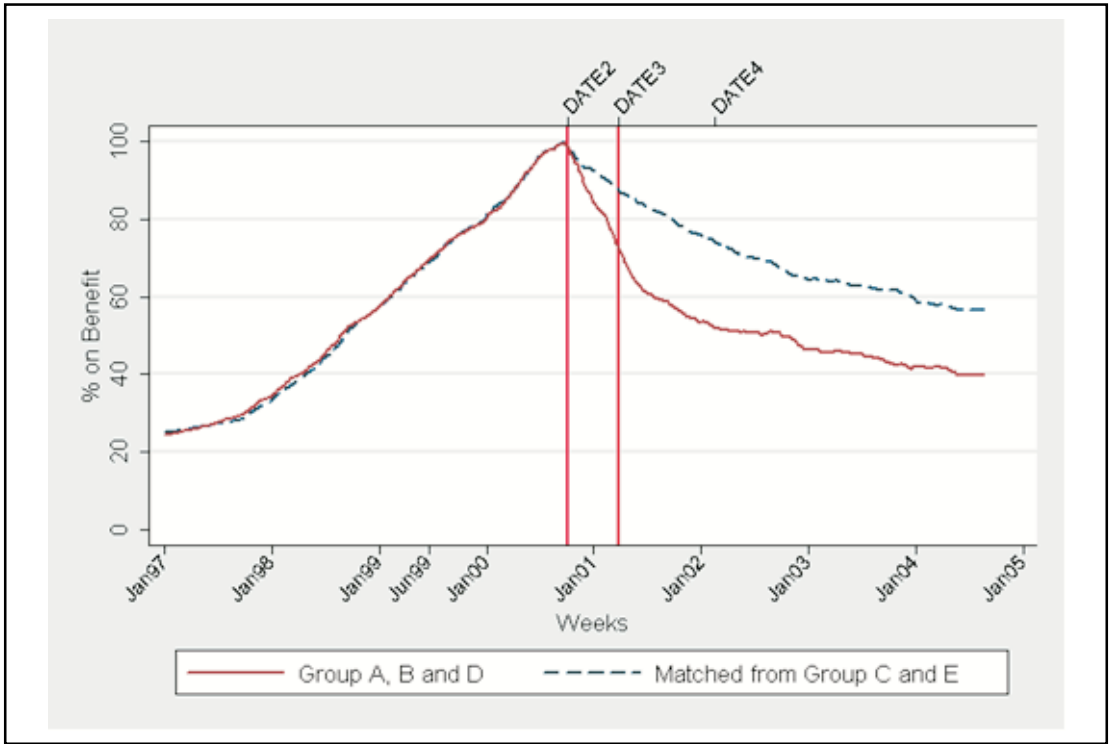
It is important to highlight that all figures generated in this section were generated solely with administrative data.

Postal questionnaire

The postal questionnaire will now be examined in order to evaluate to what extent the use of a different matching algorithm and pre-programme variables would affect the estimates produced by NatCen. A start is made by using the same propensity score specification, but changing the algorithm, allowing for replacement. The same set of pre-programme variables used in the analysis with the administrative data are then included.

As can be seen in both cases (administrative and postal), the quality of the matching estimation improved dramatically as the pre programme history variables were included (Figure J.9, Figure 6.6, and Figure J.11). As already mentioned, these variables were generated by administrative data, at a very low cost, and can be easily generated in future studies.

Figure 6.6 Weekly proportion of participants and non-participants on benefit – 1997-2003 (Sample I – Specification 3)



6.2.4 Flow and stock: heterogeneous treatment effects

One important finding in the cell matching with the benefit history strings in Section 6.2.2 is the significantly different impact estimates on the flow and stock of NDLP participants. In this sub-section, this difference is explored further by repeating the analysis completely separately for the stock and the flow. That is, a ‘hard match’ on being in the stock or the flow is forced, so that only non-participants in the stock are used to estimate the counterfactual for non-participants in the stock and only non-participants in the flow are used to estimate the counterfactual for participants in

the flow. As before, the stock is defined as those individuals who spent more than 50 per cent of the weeks in each of the six 'quarters' prior to the start of the NDLP participation window on benefit and the flow is defined as the complement of the stock. In terms of the benefit history strings defined in Section 6.2.2, the stock consists of individuals with a value of 111111 and the flow consists of everyone else.

Table 6.1 shows the estimated average treatment effects on the treated for both the stock and flow. Estimates using the full sample are presented: the preferred specification was used (model 3) and estimated not only the overall impact, but also at four points in time, namely, three, nine, 24 and 36 months after the end of the participation window.

Table 6.1 Stock/flow heterogeneous average treatment effect on the treated

Sample	Hardmatching by Stata	Description	ATT	Stock	Flow
I	Yes	A, B and D - NNPS with replacement (Specification3)	19.42 (0.17)	20.45 (0.19)	14.24 (0.21)
I	Yes	Three months on IS	21.91 (0.1)	22.49 (0.11)	18.99 (0.17)
I	Yes	Nine months on IS	22.28 (0.16)	23.72 (0.16)	15.04 (0.22)
I	Yes	24 months on IS	18.3 (0.05)	19.13 (0.08)	14.09 (0.2)
I	Yes	36 months on IS	16.51 (0.15)	17.07 (0.1)	13.70 (0.51)

Two important results emerge from this analysis: firstly, the programme impact differs quite substantially from the stock and the flow, and secondly, the effect over time also changes for both groups.

In the view of the authors, the estimates for the flow from the full administrative sample are the preferred estimates (14.24 per cent). There are three reasons for this: First, Model 3 incorporates a wealth of information on factors likely to affect both participation in NDLP and outcomes in the absence of NDLP, in particular detailed information on benefit receipt in the pre-programme period, including the benefit history strings. This model also does a good job of balancing the pre-programme benefit receipt outcomes of the participants and matched non-participants. Second, the full administrative sample is preferred to the postal sample because of concerns about non-random non-response to the postal survey, as discussed in Chapter 4. Third and finally, the estimates for the flow are preferred, both because they are more policy relevant and because it is considered that the conditional independence assumption is more plausible for them, due to the lack of complete benefit history information prior to June 1999 that would allow matching on longer histories for

the stock sample. Thus, the preferred estimate of the mean impact of participation in NDLP on those who participate is 14.24 percentage points.

It is important to highlight the differences between this impact, or additionality estimate and that presented by NatCen. This definition is calculated differently in seven ways for good reasons, the most important being listed first:

- A different time horizon is used. Average impact is estimated over a longer period of time than that used by NatCen, since the dataset used lasts until 23 December 2003 while NatCen analysis ended on 22 February 2002 (see Date in Figure 6.1).
- The impact estimates are averaged over each week for the whole administrative dataset, allowing for recidivism among programme participants. In contrast, the NatCen impact definition calculates (for only postal questionnaire respondents) the cumulative proportion in the off IS state (at Date 4) assuming that those who exit to this state will never return to IS.
- The preferred estimate only relates to the flow (those clients who were not continuously on benefit) while NatCen's include all programme participants.
- A different benefit definition is used. NatCen looked at on-off IS, whilst this report defines benefit as being on either IS, JSA or IB.
- Hardmatching by stata is used to control for the unobserved heterogeneity associated with duration of claim and age of the youngest child.
- Additional variables are used in the propensity score estimation than those used in the NatCen analysis. Specifically, several pre-programme benefit variables are included, such as the Card and Sullivan cell matching indicators and the proportion of time on IS. (Matching algorithm was also used which allows for replacement as opposed to NatCen who use an algorithm which does not allow for replacement).
- In the additionality estimate, pre-programme heterogeneity from the post-programme impact estimate is subtracted.

6.2.5 Impact estimates for different subgroups

Another interesting dimension of analysis is the effect of NDLP on different groups of participants. In this section, a preliminary investigation of this effect from a benefit history and demographic point of view is conducted. First, the NDLP effect on those individuals with the youngest child between 0 and 3, 3 and 5, 5 and 11, and 11 and 16 years old is considered. Then the NDLP impact on individuals with different pre-programme benefit histories, namely, individuals on IS for less than three months, from 3 to 6, 6 to 12, 12 to 24, 24 to 36, and for more than 36 months, is considered.

The exercise presented previously for each one of these groups was made, using the main participant definition, namely, all NDLP participants (Groups A, B and D). The covariates of the preferred specification (Model 3) of Section 6.2.3 were also

explored using only the administrative data. Together, this strategy yielded ten different figures, four for the age of the youngest child and six for the pre-programme benefit history. Figures 6.7 and 6.8 provide the same graphical analysis by the age of the youngest child. Comparing the Figure 6.8, for those with the oldest child between 11 and 16, to the Figure 6.7, relating to that with the child between 0 and 3 years, it can be seen that there is only a modest difference between the size of the NDLP effect, although this effect is clearly operating at a much higher level of being on benefits for those with the oldest child.

Looking at Figure J.17 to J.22, the same analysis by the duration of the claim prior to the programme has been graphed. Here it can be seen that the overall effect of NDLP is largest for those with claims of 6-12 months (Figure J.19) and 24-36 (Figure J.21) and over 36 months (Figure J.22). Why the effect should be smaller for those with claims of 12-24 months (Figure J.20) is not fully understood. In contrast a smaller overall NDLP effect at short durations of less than three months (Figure J.17) and between three and six months would be expected to be found (Figure J.18), as these individuals are often making a claim due to changing circumstances which would not be expected to change very quickly. As has been suggested above in the analysis the focus was on the flow of claimants rather than the stock when the policy was introduced. Equally, for similar reasons the effect on those of short durations would not necessarily be taken to be representative of the total effect of NDLP.

Table 6.2 summarises the main findings of this sub-section. One overall comment is that outcomes tend to differ more due to the pre-programme benefit history than to the age of the youngest child. Participants who were in longer benefit spells had a substantially higher probability of leaving benefit than their counterparts elsewhere.

Table 6.2 Subgroups total average treatment effect on the treated

Sample	Hardmatching by Stata	Description	Treatment Hetero		Diff
I – Flow	Yes	Children at the age of 0 and 3	16.73 (0.31)	-3.05 (0.18)	19.78 (2.21)
I – Flow	Yes	Children at the age of 3 and 5	17.71 (0.26)	-1.40 (0.14)	19.11 (1.56)
I – Flow	Yes	Children at the age of 5 and 11	14.31 (0.18)	-1.66 (0.10)	15.97 (0.77)
I – Flow	Yes	Children at the age of 11 and 16	9.92 (0.33)	-2.22 (0.27)	12.14 (3.13)
I – Stock	Yes	Children at the age of 0 and 3	17.03 (0.28)	-0.37 (0.06)	17.39 (1.47)
I – Stock	Yes	Children at the age of 3 and 5	18.75 (0.16)	0.32 (0.09)	18.42 (0.57)
I – Stock	Yes	Children at the age of 5 and 11	21.26 (0.18)	0.20 (0.07)	21.06 (0.62)
I – Stock	Yes	Children at the age of 11 and 16	23.67 (0.30)	-0.89 (0.14)	24.56 (1.87)
I	Yes	Children at the age of 0 and 3	16.96 (0.25)	-0.98 (0.06)	17.95 (1.15)
I	Yes	Children at the age of 3 and 5	18.58 (0.16)	0.11 (0.08)	18.47 (0.54)
I	Yes	Children at the age of 5 and 11	20.33 (0.17)	-0.04 (0.06)	20.37 (0.59)
I	Yes	Children at the age of 11 and 16	21.23 (0.25)	-1.13 (0.14)	22.37 (1.43)
I	Yes	On IS for less than 3 months	16.07 (0.19)	-0.54 (0.08)	16.62 (0.72)
I	Yes	On IS from 3 to 6 months	13.17 (0.30)	-3.11 (0.21)	16.28 (2.30)
I	Yes	On IS from 6 to 12 months	21.05 (0.21)	-1.13 (0.20)	22.18 (1.42)
I	Yes	On IS from 12 to 24 months	15.87 (0.45)	0.26 (0.16)	15.61 (3.93)
I	Yes	On IS from 24 to 36 months	18.12 (0.13)	-0.45 (0.08)	18.57 (0.42)
I	Yes	On IS for more than 36 months	25.79 (0.21)	0.03 (0.01)	25.76 (0.80)

With respect to the age of the youngest child, the figures suggest that lone parents with an older child have slightly higher probability of leaving benefit than those with a younger one. The effect amounts to around a five per cent difference in the proportion of those on benefit over three years later.

It must be noted that as the cohort is analysed separately by characteristics, a deterioration of the matching is observed, as can be noted by a more heterogeneous pre-programme benefit history among the participants and non-participants. Thus, the interpretation of these results must be suitably tempered with a higher degree of caution.

Figure 6.7 Weekly proportion of participants and non-participants on benefit – 1997-2003 (Sample I – Specification 3). Child between 0 and 3 years old

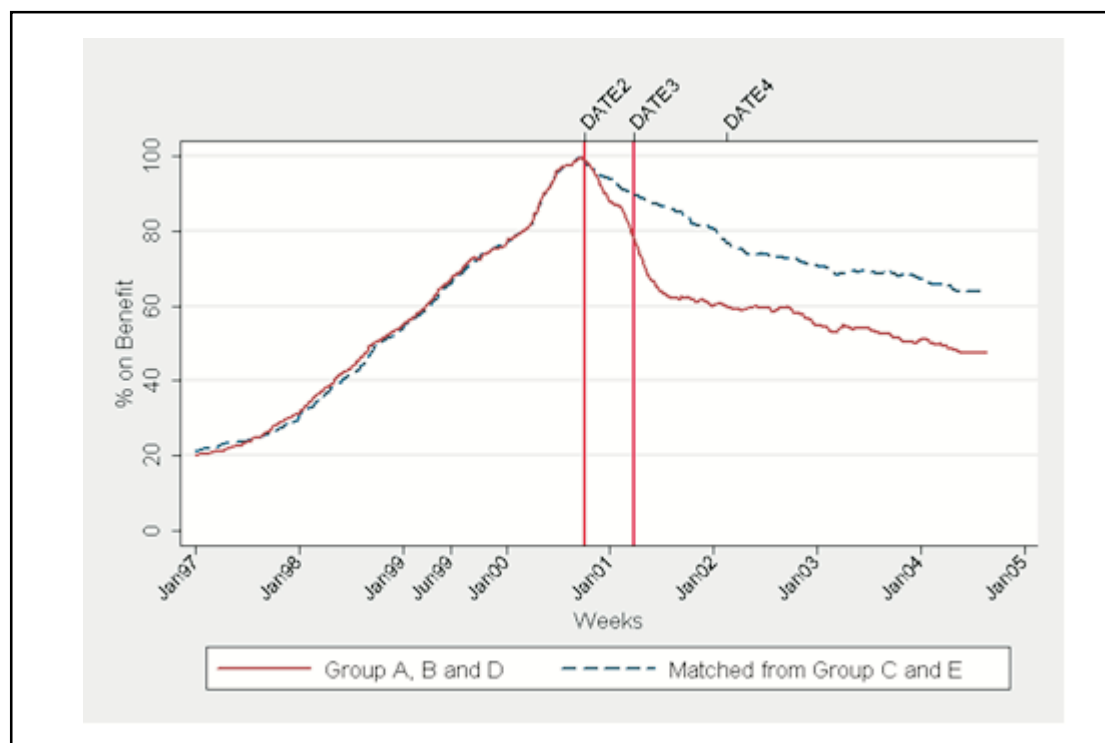
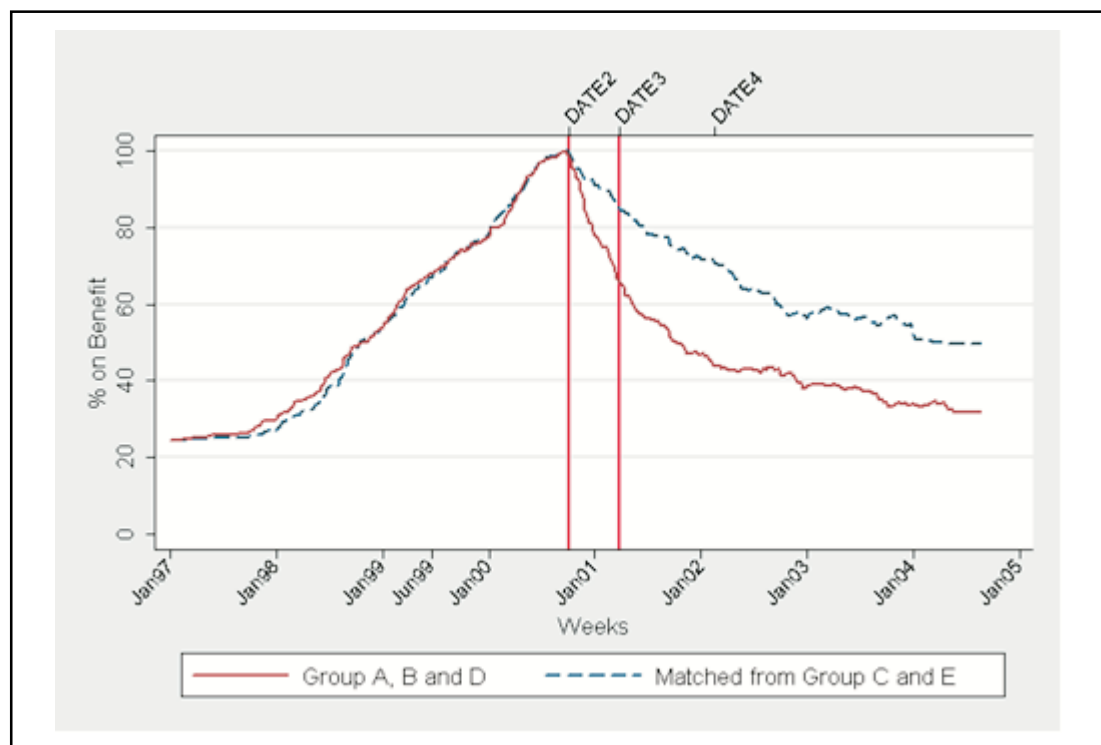


Figure 6.8 Weekly proportion of participants and non-participants on benefit – 1997-2003 (Sample I – Specification 3). Child between 11 and 16 years old



6.3 Graphs of expected outcomes conditional on the propensity score

6.3.1 Motivation

In this section, the expected outcomes conditional on participation and non-participation as a function of the estimated propensity score is analysed. That is, the average fraction of time on benefit for each value of the propensity score is considered separately for NDLP participants and non-participants. This is done by estimating non-parametric regressions of the treated outcome and the untreated outcome on the estimated propensity scores. The estimated non-parametric regression functions indicate the shapes of the relationships between the outcomes and the propensity scores. Put very simply, flexible graphs containing two lines are produced. One line shows how the mean untreated outcome varies with the propensity score and the other line shows how the mean treated outcome varies with the propensity score.

These graphs are of interest for two reasons: First, they illustrate very clearly how the mean impact of treatment varies with the probability of participation in NDLP. This information, like all information on subgroup impacts, is useful in thinking about programme targeting, outreach and so on. In general, the programme should aim to focus its attentions on individuals with larger impacts; the individual characteristics that underlie the estimated propensity scores are known, and this analysis indicates who these individuals are.

Secondly, these graphs shed some light on the economics of the problem, in a manner described in more detail in the next sub-section. By examining these graphs in the contexts of a simple economic model of participation in NDLP, the roles of factors such as opportunity costs and expected programme impacts in determining who does and does not participate in NDLP can be learnt about. For example, if individuals select into the programme based on their expected impacts, then mean impacts would be expected to be higher for those with high probabilities of participation.

The matching estimates of the impact of NDLP presented elsewhere in this report relate to these non-parametric regression functions in an important way. Most common matching estimators, including the nearest neighbour and kernel matching estimators that are employed, use predicted values from a non-parametric regression of the untreated outcome on the estimated propensity score to estimate the expected untreated outcome for the values of the estimated propensity score corresponding to each of the treated observations. The mean of the differences between the outcome of each treated observation and its expected counterfactual outcome in the untreated state constitutes the matching estimate of the impact of treatment on the treated reported throughout this report. The impact estimates implicit in the graphs arise in almost the same way; the only difference is that the treated outcomes are smoothed via non-parametric regression as well as the untreated outcomes. Asymptotically – that is, as the sample gets arbitrarily large – both procedures should yield the same estimates.

6.3.2 Conceptual framework

To interpret the graphs that we present in the next sub-section, it helps to think briefly about the participation process. In particular, it helps to think about why $P(W = w_1)$ might differ from $P(W = w_2)$, where w_1 and w_2 denote two different values of the covariates W , given that the CIA holds when conditioning on W .

In the standard model of participation, e.g., Heckman and Robb (1985), three factors loom large. They are: differences in opportunity costs, differences in direct costs, and differences in expected benefits from the programme. Suppose first that individuals with $W = w_1$ differ from individuals with $W = w_2$ only in having lower mean opportunity costs. In this case, $P(W = w_1) > P(W = w_2)$ and $E(Y_0 | W = w_1) < E(Y_0 | W = w_2)$. More generally, if differences in opportunity costs drive participation, then we would expect $E(Y_0 | P(W))$ would be expected to decrease with $P(W)$. Suppose that individuals with $W = w_1$ differ from individuals with $W = w_2$ in having lower mean direct costs, but not lower opportunity costs. In this case, $E(Y_0 | P(W_1)) = E(Y_0 | P(W_2))$ but $P(W_1) < P(W_2)$ and the function will be flat, at least around these two values.

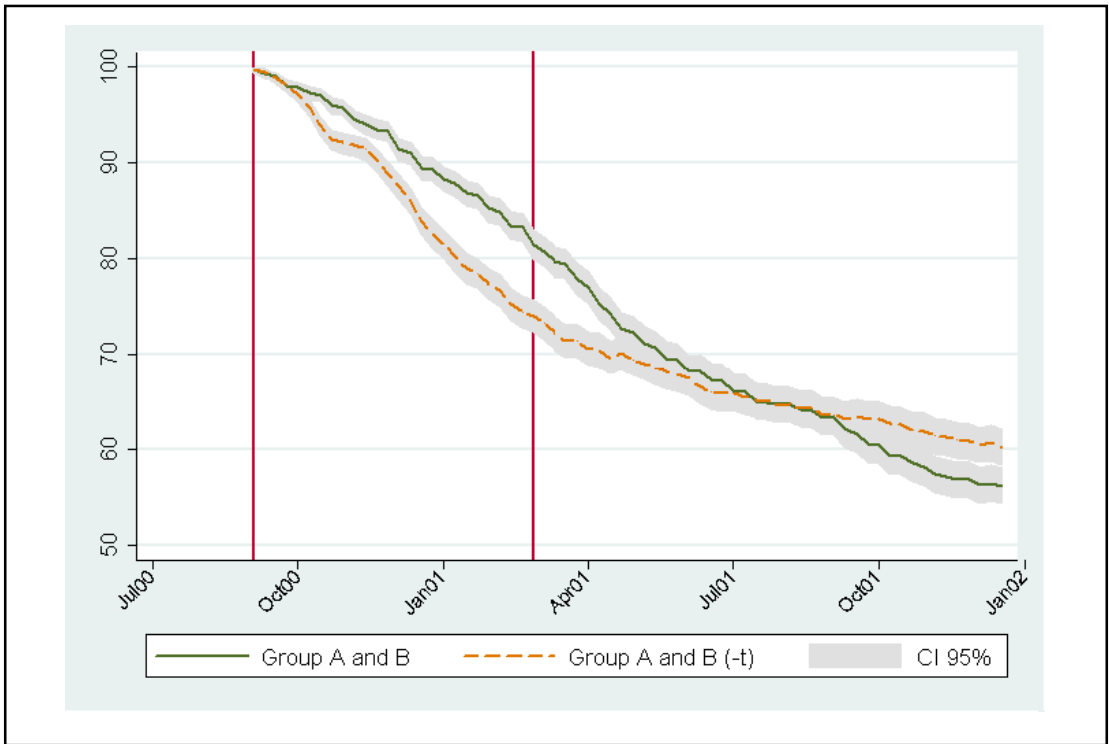
Finally, suppose that neither direct costs nor opportunity costs vary with W , but that programme impacts do. Suppose further that someone, either the potential participants or their caseworkers, have some idea of how the impacts vary, so that they can base participation decisions upon them. In this case, groups with higher mean impacts should have higher probabilities of participation. This implies that

$E(Y_1|P(W)) - E(Y_0|P(W))$ should increase with $P(W)$. This contrasts with a common effect world, in which the mean impact should not depend on $P(W)$ and the $E(Y_1|P(W))$ and $E(Y_0|P(W))$ lines should run parallel to one another.

6.3.3 Evidence

Figure 6.9 presents the two non-parametric regression functions estimated using the full sample data hard-matched by stata and weighted. The dependent variable is the fraction of time between the close of the participation window and June 2003 that the respondent spends on benefit. The upper line represents - the expected outcome in the untreated state conditional on the propensity score - while the lower line represents - the expected outcome in the treated state conditional on the propensity score. The estimation was hard-matched by stata and weighted accordingly.

Figure 6.9 Participants and non-participants LOWESS non-parametric regression (Sample I –Full)



Now consider the non-parametric regression lines in light of the discussion in the preceding sub-section. First, it is found that $E(Y_0|P(W))$ declines with $P(W)$. Because the outcome variable represents time on benefit (rather than, say, time employed or earnings), this means that the opportunity cost of participation increases with $P(W)$. Thus, it must be the case that the direct costs of participation decrease with $P(W)$ in a way that more than offsets the increasing opportunity cost.

Comparing the regression lines for the treated and untreated outcomes shows reductions in time on benefit that increase as the probability of participation increase. Although formal tests of differences in impacts are not presented by propensity score value, the regression lines in Figure 6.9 suggest important heterogeneity in impacts along this dimension. They also suggest that the NDLP may be better serving, at the margin, those lone parents on IS who would benefit most from participation.

7 Further econometric problems and issues

Summary

This chapter considers conceptual and econometric issues related to the evaluation of the New Deal for Lone Parents (NDLP) presented in this report and in the NatCen report, but beyond the scope of each. These issues include using matching to estimate programme impacts in stratified samples, potentially differing treatment effects from long and short spells of NDLP participation, taking account of the timing of participation in NDLP in producing and interpreting impact estimates, taking account of repeat participation in NDLP in producing and interpreting impact estimates, general equilibrium effects of NDLP and the role of contemporaneous Active Labour Market Programme (ALMP) in affecting the mean impacts of the NDLP. All of these issues affect the interpretation of the estimates presented here; and several suggest potential biases (in different directions) in these estimates. The aim in raising these issues is to give a sense of the state of literature in this area, to highlight the potential for biases unrelated to traditional selection bias, and to suggest useful avenues for future research.

7.1 Introduction

This chapter briefly considers five additional conceptual and econometric issues that arise in the context of the evaluation of the NDLP. The first issue is the potential effects of stratified sampling on the estimates of programme impact based on matching methods presented in this report. The second issue is the duration of NDLP participation, which is quite heterogeneous across participants. The third issue is the heterogeneous timing of participation in NDLP relative to the start of individual spells on benefit, which leads to the empirical phenomenon of non-participants in the evaluation later becoming participants. The fourth issue concerns repeat participation (or recidivism) by NDLP participants, and its effects on the construction and interpretation of estimates of the impact of NDLP participation. The final issue

concerns general equilibrium effects of the programme. A full treatment of any of these issues is beyond the scope of this report, but raising them helps to put the analyses in the remainder of the report into a broader context and suggests possible conceptual issues and lingering biases related to the existing estimates.

Of more general concern is the third issue: the dynamic nature of NDLP participation. Individuals on benefit participate at different points in their spells on benefit. As a result, a comparison group consisting of individuals who do not participate as of some date will inevitably include individuals who later become participants. The fact that the comparison group includes some future participants complicates the interpretation of the impact estimates and suggests the value of alternative, more complex, econometric estimation methods that explicitly account for this phenomenon.

7.2 Matching and stratified sampling

The data collected for many evaluations, including this one, often consist of a stratified sample. A stratified sample is one in which certain groups defined by some observable characteristics are over-sampled relative to their proportion of the underlying population of interest. Stratification has several motivations; the most common is a desire to have a sufficient sample size to allow inferences about subpopulations that represent a small fraction of the overall population but are nonetheless of policy interest. The current understanding is that the full NDLP evaluation sample is stratified on both the age of the youngest child of the Income Support (IS) recipient and the duration of the IS spell in progress as of the start of the participation window.

Stratified sampling has important implications for the estimates obtained from matching. These implications are not fully recognised in either the NatCen report or in the current version of this report. Failing to correctly account for stratified sampling in constructing impact estimates may result in strongly biased impact estimates. The potential for such bias increases in cases with heterogeneous treatment effects that vary systematically with the stratification variables. The best way to see how all this works is with a simple example, which now follows.

Consider a population consisting of two groups, which are named group 1 and group 2. Suppose that the population consists of $1/3$ members of group 1 and $2/3$ members of group 2, but that the sample is evenly split between the two groups. Put differently, suppose that group 1 has been over-sampled relative to group 1 in its representation in the population. Suppose too that, in the population, both groups have the same probability of participation in the treatment. That is, they have the same propensity score. Reflection indicates that the two groups will also have the same propensity score in the stratified sample. The reason for this is the probability of participation within a particular group (which here is synonymous with a particular stratum) is simply the fraction of the individuals within the stratum who participate; this probability is not affected by the stratified sampling. Of course,

assuming an identical probability of participation for everyone is an extreme simplification, but it allows this conceptual point to be made more easily.

Further assumptions are made about the outcomes of interest with and without treatment in the two groups. For simplicity, a continuous outcome is considered, which can be thought of as earnings in the period following the decision to participate in the treatment or not. Suppose that group 1 individuals would earn 100 without the treatment, while group 2 individuals would earn 200 without the treatment. Further suppose that an extremely effective treatment is being examined, which increases the earnings of everyone who receives it by 20. Thus, the treated outcome for group 1 is 120 and that for group 2 is 220. It is important for the discussion to be clear that this initial world is a common treatment effect world, where everyone in both strata has the same effect from treatment. Note too that because a common treatment effect is assumed, the usual parameters of interest, such as the impact of treatment on the treated and the average treatment effect, are all the same.

Now consider how matching works out in this case. Three different matching estimators (MEs) will now be defined for consideration:

ME1: Ignore the stratification.

ME2: Using the entire sample, estimate separate impacts for the group 1 and group 2 observations. Then re-weight the group-specific estimates to obtain a population estimate using the fraction of each group in the population as the weights.

ME3: Estimate an impact for group 1 using only the group 1 observations and then estimate an impact for group 2 using only the group 2 observations. Then re-weight the group-specific estimates to obtain a population estimate using the fraction of each group in the population as the weights. This is called 'hard-matching' on the variable that defines the strata.

Each of the estimators are now applied, in turn, to the data from the simple example. In each case simple cell matching is carried out. That is, as all of the observations have the same propensity score (because it was assumed that they have the same probability of participation), a comparison is made of either all treated and all untreated observations, as in ME1, all treated units in each group/stratum to all untreated units in the sample, as in ME2, or all treated units in each group/stratum to all of the untreated units in the same group/stratum, as in ME3.

To calculate ME1, it is noted that the mean outcome among all of the treated units in the sample equals $(1/2)(220) + (1/2)(120) = 110 + 60 = 170$. At the same time, the mean outcome among the untreated units equals $(1/2)(200) + (1/2)(100) = 150$. The difference between the two equals $170 - 150 = 20$, which is the mean impact both in the sample and in the population. Thus, in a common effect world, ME1 provides an unbiased estimate of the treatment effect.

To calculate ME2, the mean income of all of the untreated units in the sample is subtracted from the mean outcome of the treated units in group 1. This may seem an odd thing to do in this context, but it arises quite simply in more complex, and more realistic, contexts. For example, suppose a nearest neighbour matching was being done using the full NDLP administrative sample and propensity scores estimated using the full administrative sample. When choosing the nearest neighbours for treated units in a given stratum, untreated units from stata that were over-sampled in the course of the stratified sampling would be over-represented relative to a situation where matching was being done using a random sample. That over-representation is exactly analogous, in the simple example, to using the mean of all of the untreated units to construct the estimated counterfactual for the treated units in each of the two stata.

Continuing with ME2 and plugging in the numbers, the mean outcome of the treated units in group 1 is 120, while the mean of all of the untreated units equals 150, yielding an estimated treatment effect for group 1 of -40. In contrast, the mean for the treated units in group 2 is 220; subtracting the mean of the untreated units yields an estimated treatment effect for group 2 of $220 - 150 = 70$. Weighting the estimated treatment effects for the two groups by their population proportions gives $(1/3)(-40) + (2/3)(70) = (-40/3) + (140/3) = 100/3$ or about 33.3. This estimator does not provide an unbiased estimate of the average treatment effect. The reason is that units from group 2 get overused in constructing the estimated counterfactuals due to the stratified sampling.

Applying ME3 in the common effect world gives an unbiased impact estimate. First, a mean impact is constructed for group 1 by comparing the mean outcomes of the treated and untreated units in that group. This yields $120 - 100 = 20$. Similarly, comparing the outcome of the treated and untreated units in group 2 yields $220 - 200 = 20$. Taking a weighted average of the group-specific impact estimates using the relative proportion of each group in the population yields $(1/3)(20) + (2/3)(20) = 20$. Not surprisingly, hard-matching on the variable defining the stata, which means only matching treated individuals in each stratum with untreated individuals drawn from the same stratum, yields an unbiased estimate.

Now, consider a slightly more complicated world, in which the impact of the treatment is different, on average, for individuals in groups 1 and 2. In particular, suppose that the impact of treatment is 40 in group 1, but remains at 20 in group 2. This is a world of heterogeneous treatment effects that vary by group. As will be shown, when the treatment effect varies for groups that coincide with the stata in the stratified sampling, additional complications arise. In particular, ME1 no longer provides an unbiased estimate of the mean impact of treatment in the population.

In this world, it is still the case that the mean impact of treatment on the treated and the average treatment effect are the same within each group, so that 'the treatment effect' can still be discussed without a modifier. However, a distinction has to be made between the mean effect of treatment in the sample and in the population. In the sample, which equally represents groups 1 and 2, the mean effect of treatment

is 30, because half the individuals (the ones in group 1) get 40 and half the individuals (the ones in group 2) get 20. In contrast, in the population the mean effect of treatment is $(1/3)(40) + (2/3)(20) = 40/3 + 40/3 = 80/3$ or about 26.7. The population mean impact is lower than the sample mean impact because group 2, which has the smaller treatment effect, is under-represented in the sample relative to the population due to the stratified sampling.

When the stratification is ignored and ME1 applied, the mean outcome among the treated units in the sample is given by $(1/2)(140) + (1/2)(220) = 70 + 110 = 180$, where 140 is the treated outcome for group 1 and 220 is the treated outcome for group 2. The mean outcome of the untreated units in the sample is unchanged at 150. Thus, the mean impact provided by estimator 1 is $180 - 150 = 30$. Thus, ME1 provides an unbiased estimate of the mean impact of treatment in the sample, but does not provide an unbiased estimate of the mean impact in the population. This results from the fact that the estimator fails to account for the differing representations of the two groups in the population and in the sample. In a common effect world, this factor does not matter, because as long as an unbiased estimate is obtained for someone, it is obtained for everyone. In a heterogeneous impacts world, it can potentially matter a lot, depending on the nature of the stratified sampling and the extent of the variation in mean impacts among the subgroups defining the stata.

In the heterogeneous treatment effects world, ME2 has the advantage relative to ME1 of weighting the group-specific impacts by the population proportions from each group. This is not enough, though, when the underlying group-specific estimates are themselves biased. Going through the mathematics, for group 1 we get $140 - 150 = -10$ for the group specific impact estimate, while for group 2 we again get $220 - 150 = 70$. The weighted average of the two in the heterogeneous treatment effects world is $(1/3)(-10) + (2/3)(70) = (-10/3) + (140/3) = 130/3$ or about 43.3. Once again, the estimator is biased due to the overuse of untreated units from the over-sampled stratum in constructing the group-specific impact estimates.

Finally, consider ME3 in the heterogeneous treatment effects world. The mean outcome of the treated units in group 1 is now 140, while that of the untreated units in remains 100. Thus, the impact estimate for group 1 is 40. Similarly, the mean outcome of the treated units in group 2 is 220 and the mean outcome of the untreated units equals 200, yielding a mean impact of 20. Taking the weighted average of the mean impacts for the two groups using the population weights gives $(1/3)(40) + (2/3)(20) = 40/3 + 40/3 = 80/3$. Thus, ME3 provides an unbiased estimate

of the mean impact of treatment in the population. In a world of heterogeneous treatment effects and stratified sampling, among the estimators considered here, only estimator ME3 provides an unbiased estimate of the parameter of interest.¹⁹

The discussion here has two main implications for the analysis of the impact of the NDLP in this report: First, if the world is considered a 'common effects' world, or at least a world in which the effect of NDLP participation does not vary, on average, among the groups defined by the stratification variables of age of youngest child and duration of IS spell, then the estimator ME2 simply needs to be avoided. In addition, a comparison of estimator ME1 with estimator ME3 is useful because these estimators may differ in finite samples because they implicitly emphasise different variables in the matching process.

Secondly, if the world is considered to be a heterogeneous treatment effects world, then examining the hard matching estimator, estimator ME3, is crucial, for only that estimator provides the correct answer among the estimators described here. It is considered likely that there is at least some heterogeneity in the treatment effects as a function of length of IS spell (which likely proxies for labour market attachment) and by the age of the youngest child (which reflects the opportunity cost of the lone parent's time, among other things). If there is substantial heterogeneity in treatment effects among the stata defined by these variables, the impact estimates resulting from applying estimator ME3 could differ substantially from those obtained by applying ME1.

7.3 The duration of NDLP participation

7.3.1 Conceptual issues related to treatment duration

It is shown in this section that the duration of NDLP participation varies widely among participants. Put differently, individual participants receive different 'doses' of the NDLP treatment. If it is considered that more of the NDLP treatment has a larger effect than less of it, then it might be interesting to estimate the differential treatment effects associated with different durations on NDLP. Doing so is more important than it might at first seem, because, unlike the variation in doses

¹⁹ There are alternatives to hard matching on the variables defining the stata. For example, one could use kernel matching on the full sample but re-weight the untreated observations to take account of the stratification. In this scheme, the weights used in constructing the estimated expected counterfactual outcome for a given treated observation would reflect both the stratified sampling and the distance, as measured by the kernel function, between the propensity score of the treated observation and the untreated observations. The practical advantage of the hard matching estimator is that it is much simpler to implement in Stata. It is conjectured that which estimation scheme minimises the mean squared error depends on the nature of the stratified sampling and its interaction with the variation in the mean impact among the stata.

examined in drug trials, from which the 'dose-response' jargon is drawn, doses in the NDLP are themselves endogenous. Thus, attempting to examine the effects of alternative doses introduces a second selection problem into the analysis. In addition to the problem of selection into NDLP per se, there is the second problem of selection into different doses of NDLP.

The most obvious complication caused by different possible NDLP participation durations is that a longer duration may not simply mean more treatment. In the simplest dose-response literature, the longer the treatment, the more likely one would expect to observe an effect from it. This need not be so in the NDLP context. For example, the initial Work Focused Interview (WFI) meeting might change the world of some lone parent, leading them to get straight into a job and sign off receipt of benefit. In this case, the shortest possible NDLP treatment may be associated with an instantaneous large effect on outcomes, while longer NDLP treatments correspond to individuals who linger on NDLP in an attempt to avoid actually finding work. This example makes it easy to see that longer durations on NDLP may well correspond to smaller, rather than larger, impacts from NDLP participation.

To further illustrate why this is a hard problem, consider in more detail three other special cases. To simplify, suppose in each case that there are only two durations, long and short. In the first case, suppose that unmotivated participants quit NDLP and have a short duration, while ambitious participants stick with NDLP and have a long duration. In this case, a simple comparison of the IS participation rates of those with long and short durations will overstate the effectiveness of long durations, by conflating the duration effect (if any), with the non-random selection of the unmotivated and ambitious into treatments of different length. This additional selection can occur whether or not selection into the programme has been successfully dealt with. In the second special case, the NDLP treatment effect does not vary by duration, but individuals simply stay on NDLP until they leave IS. In this case, even though the treatment effect does not vary by duration, simple comparisons will suggest that those with short treatments leave IS faster than those with long treatments. Finally, suppose that local administrators pick a duration that they prefer for all the clients in their locality, and that this decision depends on the factors unrelated to the potential outcomes of the local participants. In this case, because treatment duration varies exogenously, simple comparisons of the outcomes of NDLP participants with long and short programme durations will provide consistent estimates of the differential treatment effects of long programme durations.

The point of these examples is that it matters critically where the variation in NDLP programme duration comes from. The various econometric methods for estimating treatment effects that vary by treatment duration make, implicitly or explicitly, assumptions about the nature and origin of the variation in treatment duration. For example, consider estimating the impact of NDLP using a simple regression of the outcome of interest on a dummy variable for NDLP participation and some covariates. This method of estimating the impact assumes 'linear selection on

observables', which is to say that it assumes that conditioning on observable characteristics linearly in a regression model takes care of any problem of non-random selection into NDLP. If the single dummy variable for NDLP participation is replaced with two dummies for 'long' and 'short' spells of participation, this implicitly adds the additional assumption that the same set of observable variables takes care of any non-random selection into having a long or short spell of NDLP participation. This is a strong assumption that requires theoretical and/or empirical justification; a set of conditioning variables that do control for non-random selection into a programme may or may not also control for non-random selection into different durations of programme participation.

In a matching context, the 'dose' is usually discretised and then the problem is handled as a multi-treatment matching problem, where no dose is one of the multiple treatments. See the discussions in Imbens (1999), Lechner (2000) and Behrman, Cheng and Todd (2004). In choosing the variables for a matching context, it is important to recognise that they must lead to a conditional independence assumption holding that implicitly solves both of the selection problems mentioned above.

7.3.2 Evidence on the duration of NDLP participation

Figures 7.1 and 7.2 present kernel-smoothed densities of the duration of NDLP spell lengths for participants in the A, B and D groups and in just the B group, respectively. For both participant samples, the data indicate a wide variety of different durations on NDLP among those participating in the NDLP during the time of the study. Both duration distributions are skewed right, with most of the mass between zero and 50 weeks, and then a long tail stretching out over 100 weeks, with a second, minor, peak in the distribution between 120 and 150 weeks. The substantial fraction of participants with relatively long durations (more than 13 or 26 weeks) reinforces the values of looking at longer term outcome measures than those considered in the NatCen report, as is done elsewhere in this document.

7.3.3 Understanding the heterogeneity in the duration of participation

Interpreting estimates of the differential impacts of long and short spells of NDLP participation requires an understanding of the nature of treatment implicit in long and short spells, as well as the nature of the closely related processes of selection into different types of services within NDLP and into participation spells of different durations. Some of the key issues are highlighted here. First, as noted in Lessof et al. (2003), NDLP provides a variety of different types of services to its participants. Some receive just one interview with a caseworker while others receive a long series of interviews. Some participants receive job search assistance, others receive classroom training and still others receive on-the-job training of various sorts. Service type necessarily correlates with the duration of participation because different services have different durations by design. Moreover, there will be heterogeneity in the duration of participation conditional on service type as well. For example, participants

who start a classroom-training course may drop out before completion for a number of reasons (including finding a job). Different classroom training courses may also have different planned lengths.

Secondly, the association between service type and duration also suggests an association between the duration of participation and its intensity. If all participants spent 40 hours a week on NDLP activities during their spell, regardless of service type, then there would be no association between duration and intensity. However, this is unlikely to be the case in practice. It might be imagined that short spells represent a single WFI meeting or perhaps job search assistance, both low intensity services, while longer spells represent more intensive services. In this case, longer durations of participation would be expected to have more than proportionally greater impacts than shorter durations of participation. In contrast, if very long spells represent long sequences of relatively infrequent WFI meetings, then the reverse may be true.

Thirdly, and more prosaically, there may be important issues regarding non-classical measurement error in the duration data. This measurement error may result from caseworkers failing to promptly record in the administrative data when participants stop participating in NDLP. Such measurement errors may affect the relationship between duration and intensity because long spells that result from measurement error will include long periods of what is actually non-participation. Failure to account for this measurement error in an analysis of the impacts of participation spells of different durations will lead to misleading conclusions, most likely by biasing downward the estimated impact of long spells.

7.3.4 Implications for the analyses in this report

Differences in the impact of NDLP on participants according to the duration of their participation does not affect the consistency of the impacts of the **incidence** of NDLP participation presented in this report. Rather, the discussion in this sub-section is intended to remind the reader that estimates of the impact of the incidence of NDLP participation represent a relatively crude measure of programme effectiveness, which leave open a variety of important and interesting evaluation questions, the answers to which would provide useful guidance to those seeking to improve the performance of the NDLP programme.

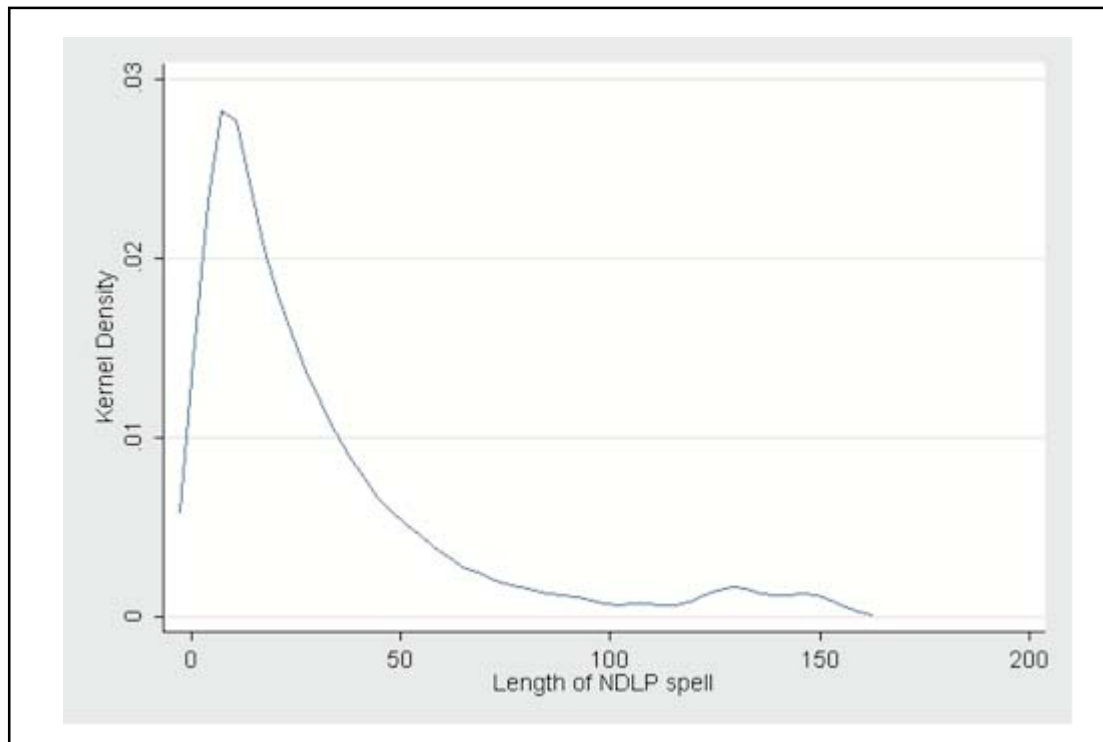
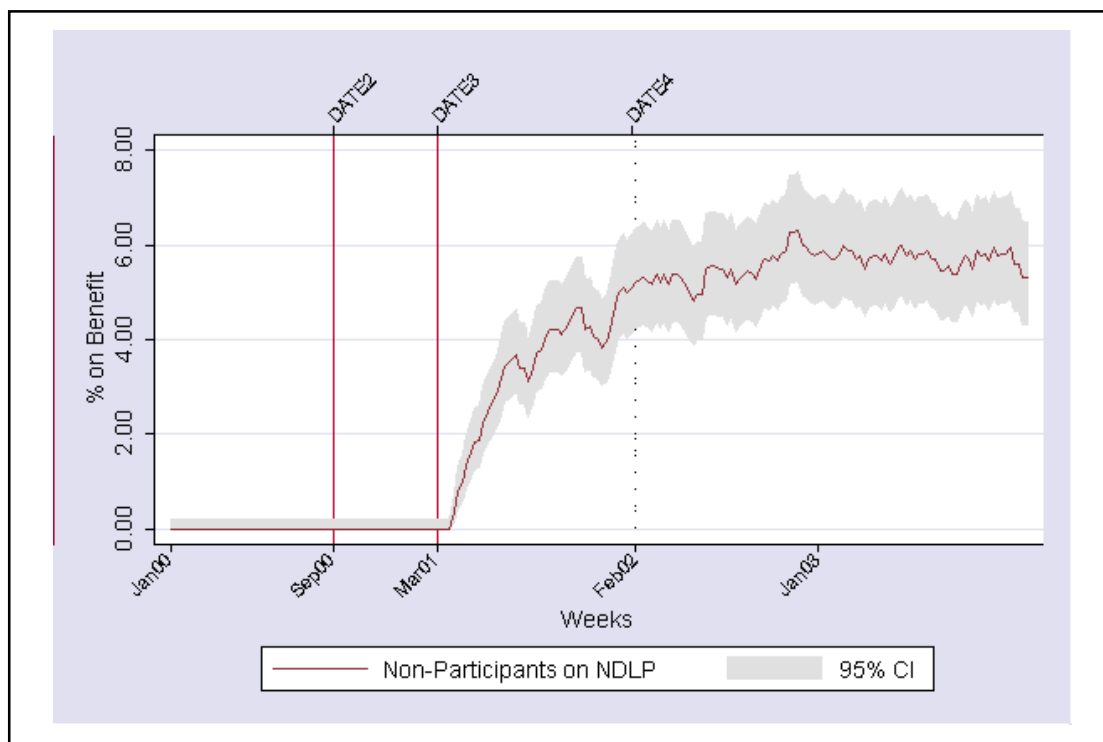
Figure 7.1 Duration of NDLP spell in weeks (Groups A, B and D)**Figure 7.2 Duration of NDLP spell in weeks (Group B)**

Figure 7.3 Proportion of non-participants at April 2001 who subsequently became NDLP participants (Groups C and E)

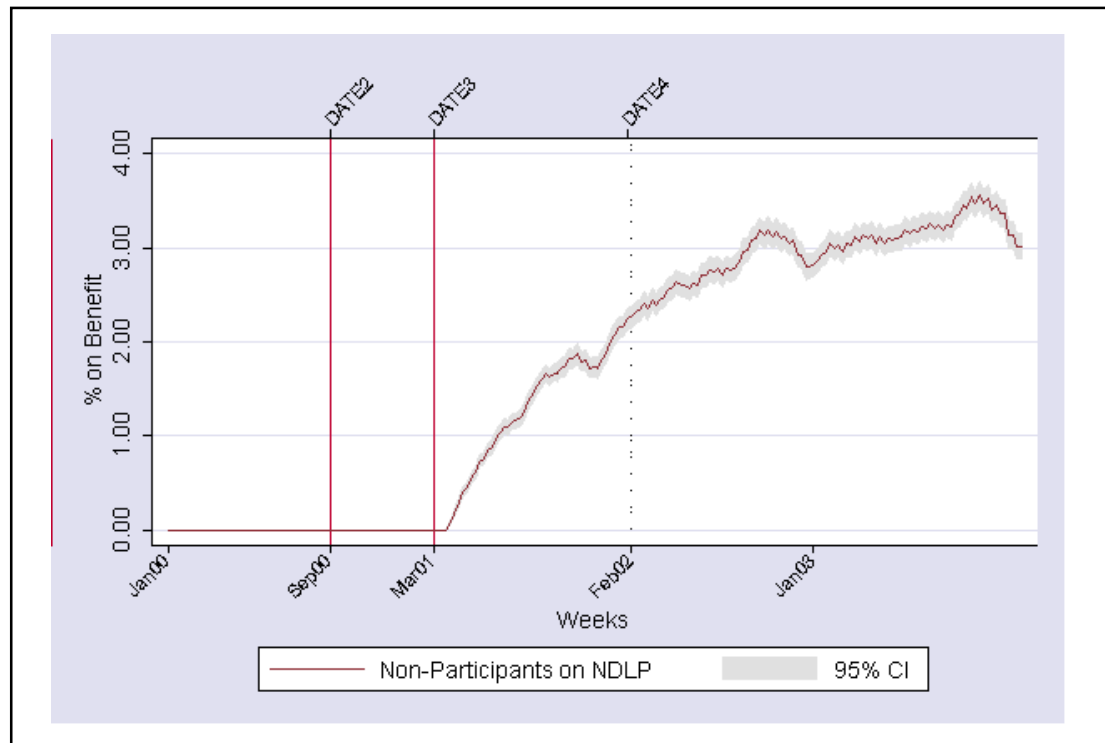


Figure 7.4 Cumulative Proportion of Non-Participants at April 2001 who subsequently became NDLP participants (Group C and E)

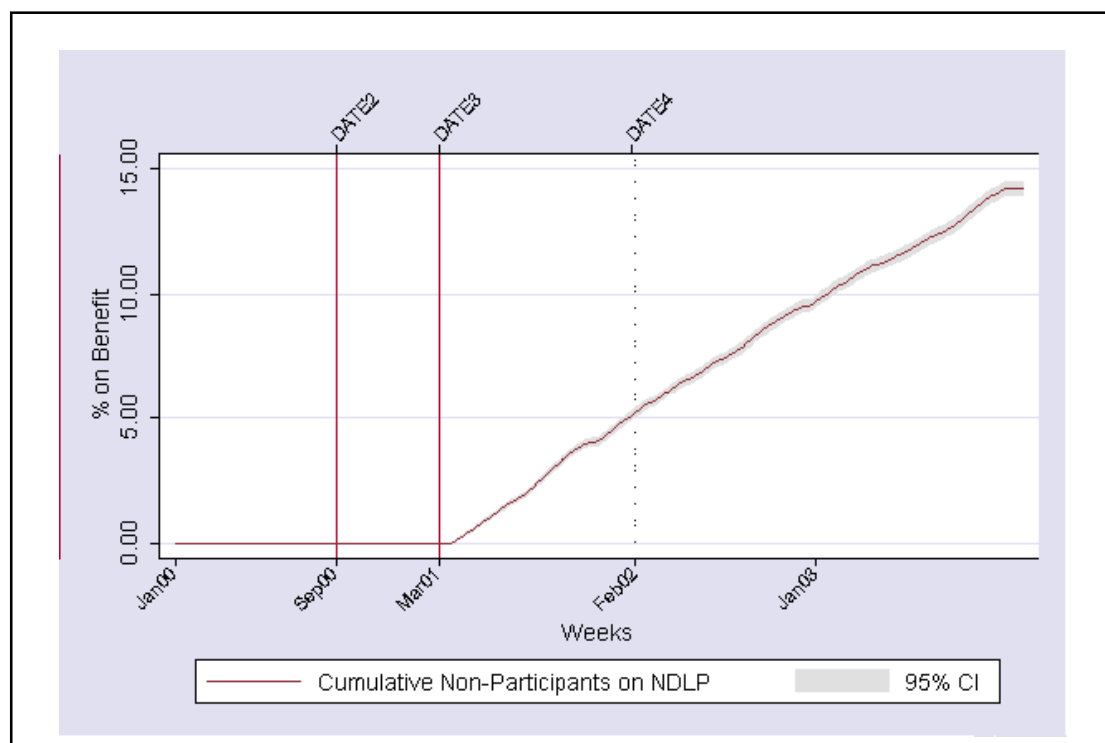


Figure 7.5 Proportion of Participants and Non-Participants on Benefit 1997-2003 (Sample I) Only if Youngest child age ≥ 5 and ≤ 15)

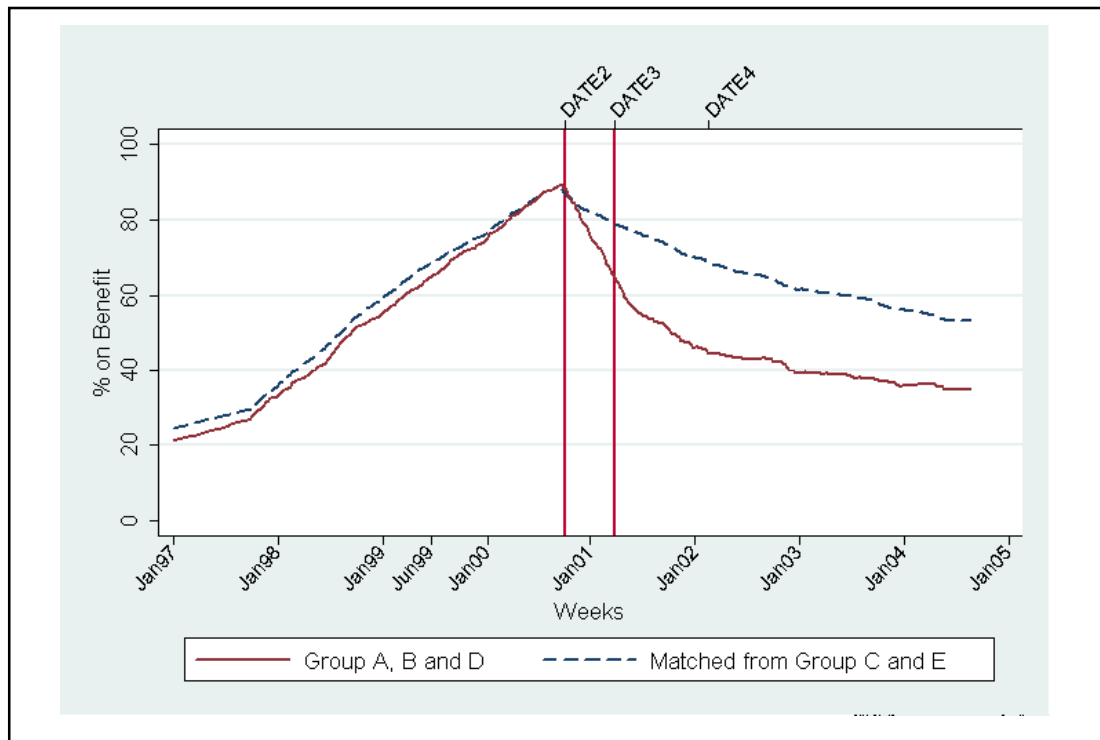
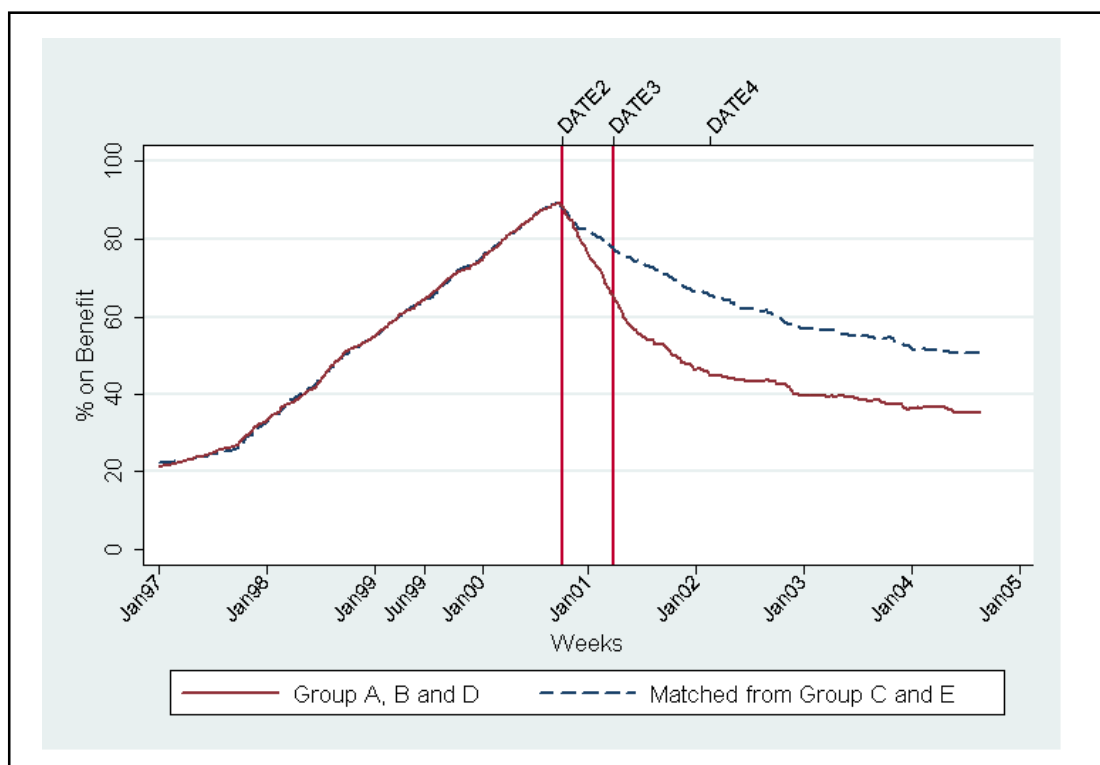


Figure 7.6 Proportion of Participants and Non-Participants on Benefit 1997-2003 (Sample I) C&S, Recent and Long Hist.)



7.4 NDLP participation starts at different times

7.4.1 Dynamic participation: conceptual issues

The standard econometric models for estimating treatment effects, such as those in Heckman and Robb (1985) and Heckman, LaLonde and Smith (1999), assume that the treatment of interest is available in only a single period, $t = k$, and then devise ways to compare those who choose to participate in that period with those who do not, taking account of any non-random selection into the programme as a function of either the untreated outcome (the traditional case) or the impact of the treatment (the more recent interest in the literature). In fact, many programmes, particularly in a European context, have a structure wherein individuals become eligible for a programme by being on some sort of benefit, such as IS or Jobseeker's Allowance (JSA) in the UK context, and then take the treatment, such as NDLP, at some point in their spell, with that point varying among individuals. Even in cases such as the NDYP, where individuals are forced to participate six months into their JSA spell, individuals may choose to participate earlier in the spell, thus raising issues of dynamic participation.

Dynamic participation raises intertwined issues of parameter interpretation and estimation. To see the problem, consider the following, which is based on the model in Frederiksson and Johansson (2003) and, more broadly, on the models in Heckman and Robb (1985) and Heckman, LaLonde and Smith (1999). Ichimura and Todd (1999), Abbring and van den Berg (2003a,b) and Sianesi (2004) also discuss these issues and suggest ways to deal with them econometrically.

Consider a population of individuals who become eligible for some treatment in period 1. Each individual can be in one of three states in each period: employed (and thus ineligible for the treatment), not employed and in treatment. The treatment lasts one period and can be taken only once. Each individual has a probability of finding employment in each period that consists of a constant baseline employment probability common to all and a transitory component that is independent of everything else. At the start of each period, the individual observes their employment probability (but not the associated outcome). With this information in hand, the individual decides whether or not to participate in treatment. Participating in treatment lowers the current period employment probability to zero but results in a permanent increase in the probability in all future periods. In this model, individuals choose to participate in periods when they get a particularly bad draw on their current period employment probability, because it is in such periods that the opportunity cost of the treatment (which is the only cost of treatment in this model) is the lowest.

At this point, it is helpful to define some simple notation, in order to clarify issues of parameter definition and estimation. The notation in Ichimura and Todd (1999) is developed further. Let $Y_0(t)$ denote the untreated outcome in period ' t ' and let $Y_1(t)$ denote the treated outcome in period ' t '. In this example, the outcome is a binary

variable denoting employment, but all the issues raised here apply to continuous outcomes such as earnings and to multinomial outcomes such as employment and participation in alternative benefit programmes such as Incapacity Benefit (IB). Three sets of indicator variables will now be defined. The first indicates participation in period ' t ' and is defined as $I(1), \dots, I(T)$, where period T is the end of life. The second set denotes participation at or before period ' t ', and is defined as $D(1), \dots, D(T)$. The final set denotes treatment availability in a given period, and is defined as $A(1), \dots, A(T)$. Note that $A(t) = 0$ implies that $I(t) = 0$ but does not imply that $D(t) = 0$ for $t > 1$. The observed outcome for each individual is given by:

$$Y(t) = D(t)Y_1(t) + (1 - D(t))Y_0(t).$$

In terms of this notation, two alternative parameters of interest can be defined. The first parameter is the usual one, which is the mean impact of treatment on the treated for individuals treated in period 1, which can be written as:

$$\Delta^1(t) = E(Y(t) | I(1) = 1) - E(Y(t) | I(1) = 0) = E(Y_1(t)) - E(Y(t) | I(1) = 0).$$

The problem with this parameter in a context where individuals can participate in the treatment in different periods is the fuzzy nature of the counterfactual. In terms of the notation $E(Y(t) | I(t) = 0)$ for $t = t' > 1$ will consist of a weighted average of the treated outcome for individuals participating after period 1 but on or before period t' and the untreated outcome for those not participating through period t' . This parameter represents the mean impact of participating in period 1 on those who do so, relative to participating or not at some point in the future. This parameter is difficult to interpret cleanly and difficult to incorporate into a cost-benefit framework.

In contrast, if the interest is in estimating the impact of participation relative to a counterfactual that approximates a world without the programme, then it can be estimated:

$$\Delta^2(t) = E(Y_1(t) - Y_0(t) | I(1) = 1, A(1) = 1, A(2) = \dots = A(T) = 0).$$

This parameter is easy to interpret, answers a policy question of clear and obvious interest – namely what difference does the existence of the programme make to the outcomes of those participating in period 1 – and provides the correct number for a cost-benefit analysis designed to answer the question of whether or not the programme should be kept in place.

If the treatment is available only in period 1, so that $A(1) = 1$ but $A(t) = 0$ for $t > 1$, then the evaluation problem is potentially quite a simple one. For example, if participation depends only on the transitory component of the employment probability, and that transitory component has no serial correlation, simple differences in outcomes between participants and non-participants in later periods provide consistent estimates of $\Delta^2(t)$.

The literature on dynamic participation in treatment remains active, and the issues outlined in this sub-section await a full resolution. The Frederiksson and Johansson (2003) and Sianesi (2004) papers focus on the estimation and interpretation of the parameter $\Delta^1(t)$ and make important progress on this front. The papers by Abbring and van den Berg (2003a,b) frame the issues somewhat differently than is done here. They develop a competing risks model in which the risks are finding work (or leaving benefit) and taking the treatment. Once the treatment is taken, this changes the hazard rate of finding work (or leaving benefit). The estimated treatment effect consists of this change in the exit hazard. This is a clever and attractive method, which also allows for heterogeneity in certain forms. At the same time, it has a couple of important limitations. The first is that it does not capture the effects of the treatment on the timing and durations of subsequent returns to non-employment (or benefit receipt) or on repeat participation in the treatment. As a result, it is difficult to use the treatment effect estimates it provides as an input into a full cost-benefit analysis in contexts where these phenomena matter empirically, as they do in the NDLP context. Second, it requires an assumption of no anticipatory (or deterrent) effects of treatment. Although less relevant to a voluntary programme such as NDLP, the recent literature, such as Black, Smith, Berger and Noel (2003), suggests the importance of such effects in the context of mandatory active labour market programmes.

The analyses in this report, as well as those in the NatCen report, estimate the problematic parameter $\Delta^1(t)$ rather than the easily interpreted parameter $\Delta^2(t)$. This follows from the fact that lone parents not participating during the participation window considered in both evaluations may (and do, as is shown in the next subsection) participate in NDLP later on. In Section 7.4.4 the implications of estimating the problematic parameter for the interpretation of the results of these analyses is shown. Before considering those implications, Section 7.4.2 presents evidence on the empirical importance of participation in NDLP by lone parents not participating during the participation window defined for the evaluation, and on the sensitivity of the impact estimates to omitting these individuals from the analysis during the period of their NDLP participation.

7.4.2 Evidence of later participation by non-participants

This section presents evidence on the extent of later participation in NDLP by individuals who do not participate during the participation window used to define participation in our analyses and in the NatCen report. Figure 7.2 and Figure 7.3 graph the fraction of non-participants during the participation window participating in NDLP in each week after the end of the participation window. Figure 7.2 provides the figures for groups C1 and C2 (the matched non-participants among the postal-survey respondents) while Figure 7.3 provides the proportions for groups C and E (the non-participants in the administrative data).

Figure 7.2 and Figure 7.4 provide the cumulative fraction of non-participants who have participated in NDLP for the same samples in each week after the end of the participation window. Figure 7.2 indicates the number of individuals in the various

samples with spells that start and end at different points relative to the participation window.

This evidence provides two clear messages: First, there is a substantial amount of later participation in NDLP among individuals who do not participate during the participation window. Around 12 per cent of all non-participants during the participation window participate by June 2003. Second, the rates of later participation are even higher for the matched non-participants. Among the matched non-participants, almost 25 per cent have participated by June 2003. This second finding is expected because the matched non-participants are precisely those non-participants whose observable characteristics make them similar to participants. Thus, it would be expected that they participate more later on.

7.4.3 A simple sensitivity analysis

Although a detailed examination of the effects of later participation on the estimated impacts lies well beyond the scope of this report, in this sub-section we briefly reports the results of a simple sensitivity analysis. In particular, Figure 7.4 repeats the standard analysis but omits non-participants who participate in NDLP for those weeks in which they are enrolled in NDLP.

Figure 7.5 does the same thing but also omits the corresponding matched participants, again only for those weeks during which the non-participant (during the participation window) is enrolled in NDLP after the participation window closes. Both figures start with participant sample B and comparison group samples C1 and C2, which together comprise the matched samples of postal survey respondents from the NatCen analysis.

The procedure here is not defensible as an impact estimator, but the finding that the graphs do not change much is an indication that the matched non-participant lone parents who participate in NDLP after the participation window closes for the study are not wildly different in terms of their outcomes than the matched non-participant lone parents who do not participate between the closing of the window and June 2003.

7.4.4 Implications of dynamic participation for the analysis

The evidence in Section 7.4.2 of the high rates of later participation in NDLP by lone parents not participating during the participation window for the evaluation, suggests that $\Delta^1(t)$ and $\Delta^2(t)$ differ substantially in this context. If the NDLP treatment has a positive effect on outcomes, then we it would be expected that $\Delta^1(t) < \Delta^2(t)$. That is, the fact that some non-participants switch to participants after the participation window closes improves their outcomes relative to the world without the programme, and thus biases down the impact estimate down when interpreted as the impact of the existence of the programme on those who participate during the participation window.

Not much can be said *a priori* regarding the likely size of this bias. It is useful, however, to think a bit about the question of the magnitude of the bias in the context of the simple model presented above, generalised to include heterogeneous treatment effects among individuals. It is clear that the extent of the downward bias increases with the fraction of the initial non-participants who later become participants in NDLP. However, the increase may not be directly proportional if those who participate later have, on average, a lower impact from treatment than those who participate during the participation window. Common sense, as well as simple models of programme participation based on opportunity costs, predicts this to be the case. Individuals who expect a large impact from the programme will participate even when the opportunity costs are relatively high. In contrast, individuals who expect a relatively low impact from the programme will participate only in periods when the opportunity cost is especially low. Thus, assuming that opportunity costs vary over time in a way unrelated to the expected gain from participation, those with low expected gains from the programme will wait, on average, longer to participate. Indeed, some individuals with low gains from the programme may never participate, along with the unknown number of lone parents on benefit who anticipate a negative effect of participation. Thus, in sum, the reasoning and evidence here suggest that dynamic participation concerns bias downwards the estimated treatment effect when it is interpreted as measuring the impact of NDLP relative to a world without NDLP, but the extent of that bias remains a topic for further research.

7.4.5 Changes in NDLP eligibility over time

Another issue related to the timing of participation in NDLP is changes in eligibility for the programme over time due to demographic changes. Obviously, individuals can become ineligible for NDLP by moving off benefit and into employment. However, they can also lose their potential eligibility for NDLP by ceasing to be a lone parent due to finding a new partner, a child growing older, losing custody of the child or the death of a child. The second factor is empirically the most important of these and we focus on it here. Because the definition of lone parent is a fixed and known function of the age of the youngest child, lone parents who want to participate in NDLP can time their participation accordingly. This facet of the NDLP participation process has two main implications for the analyses in this report: First, the fact that some individuals eligible for NDLP during the participation window leave eligibility due to the ageing of their children, places an upper boundary on the extent of later NDLP participation by lone parents not participating during the participation window. Second, it is important, as this analysis had to and as NatCen did, to include the age of the youngest child among the matching variables.

To evaluate in a simple way the potential empirical importance of this issue, we present in Figure 7.5 and Figure 7.6 versions of the figure showing the fractions of participants and matched non-participants on benefit restricted to lone parents whose youngest child's age makes them at least demographically eligible for NDLP from June 1999 to June 2003. This requires that the youngest child be from five to 15 years of age in June 1999.

Figure 7.5 uses treated sample B, the NatCen propensity scores and single nearest neighbour matching with replacement. Figure 7.6 uses treated sample B, along with the alternative propensity scores that include the detailed benefit history variables constructed from the administrative data. The figures show that focusing on only those lone parents demographically eligible throughout the period does not change the basic story very much. The estimated impact of treatment on the treated corresponding to these figures is around 0.13; that is, in this subpopulation, NDLP participants spend about 13 percentage points less time on benefit during the follow-up period through June 2003.

7.5 Multiple spells of NDLP participation

7.5.1 Conceptual issues

This section considers lone parents who have multiple spells of participation in NDLP. As with the issue of dynamic participation considered in the preceding section, the existence of substantial repeat participation takes this analysis outside the standard evaluation framework considered in, eg, Heckman and Robb (1985) and Heckman, LaLonde and Smith (1999), in which a programme is offered once at a single point in time.

Repeat participation raises issues in terms of both the definition of the parameter of interest and in terms of econometric estimation of treatment effects. For example, there may be additional parameters of interest, such as the impact of a second spell of NDLP participation (to be compared to the impact of the first spell), or the impact of participating twice compared to participating once (which is not quite the same thing, as the impact of the first spell can include the possibility of participating again). In the case where the impact of the first spell is of interest, it has to be interpreted to include additional spells of participation as part of the potential outcome associated with the initial treatment.

The econometrics are difficult here, in part because it is hard to write down an economic model in which individuals choose to participate more than once. In the typical human capital investment model, a standard result is that you want to 'go to school when you are young', in order to have the maximum amount of time over which to amortise your investment, and then to work steadily and avoid school for the remainder of the life cycle. A simple model of NDLP would yield the same conclusion – that you should participate right at the start of the IS spell (or perhaps shortly thereafter when other search options have been exhausted) and then not participate again. To get multiple spells of participation, it is essential both that the second spell has a positive treatment effect (which it may not if participants have 'seen it all' after their initial spell of participation) and that the opportunity costs of participation changes over time. Thus, there is a requirement for a story where, for example, individuals get and then lose a job (a job raises the opportunity cost of NDLP participation) or a story where, for example, family troubles such as ill health in a child or a co-resident parent raise the opportunity cost and so lead to temporary withdrawal from NDLP.

The econometric literature has not developed methods for handling the case of repeat participation within the treatment effects framework. Developing analogues to some of the methods mentioned in Section 7.4 for handling dynamic participation to deal with the case of repeat dynamic participation can be envisaged. Moreover, a full analysis along these lines would require time-varying data on family status and health. Although a full analysis of repeat participation is beyond the scope of this report, in the next sub-section evidence on the extent of repeat participation in the NDLP evaluation data is presented and in the following sub-section the implications of repeat participation for the analyses in this report are discussed.

7.5.2 Evidence on multiple spells of NDLP participation

Table J.36 shows the number of spells of NDLP participation after the closing of the participation window for both participants and non-participants. The left side of the table presents the values for Sample I (all eligible lone parents) and the right side for Sample III (the NatCen matched sample). In each case, the column for non-participants indicates the number of spells of NDLP participation after the close of the participation window. A non-trivial fraction (0.1425) of the non-participants in Sample I who participate after the participation window closes have more than one spell of non-participation. For the participants, any additional spell after the participation window closes indicates multiple spells of NDLP participation. Thus, in Sample I, 1,465 participants repeat in NDLP prior to June 2003. The corresponding number in Sample III is 680.

Table J.37. presents spell counts for the same samples, but with the count starting at the start of the NDLP programme on 23 October 1998. The values in this table reveal that among participants during the participation window, 27.72 per cent (in Sample I) have multiple spells of NDLP participation. This is a substantial amount of repeat participation among participants, well in excess of the rates found for the US Job Training Partnership Act (JTPA) programme in Baj and Trott (1993). As an aside, Figure 7.2 indicates that a handful of participants also had spells that both began and ended **prior to the start** of the participation window for the evaluation. Thus, not all of the multiple spells occur after the close of the participation window.

7.5.3 Implications of multiple spells of NDLP participation

The focus here is on multiple spells of NDLP participation among individuals who participate during the window; parallel issues arise from multiple spells among non-participants who participate in NDLP after the window closes. The presence of multiple NDLP spells changes the nature of the treated state. That state now includes not just participation during the participation window defined for the study, but also the possibility (and the reality in this case) of repeat participation in later periods.

Thus, even putting aside the issues address in Section 7.4, the estimates no longer correspond to the impact of the programme being available in a single period on those who participate during that period. Now the treatment being evaluated consists of participation during the window relative to non-participation during the

window, where both of these states potentially include participation, and repeat participation, in future periods. Although potentially informative, this comparison does not correspond directly to what should be included in a social cost-benefit analysis of whether or not to keep the programme as it is or eliminate it.

In thinking about attempting to separately estimate the impact of the second (or third or subsequent) spell of NDLP participation, it is not clear *a priori* whether repeat spells would be expected to have a larger or smaller impact on the labour market outcomes of participants than initial spells. One argument would be that in a second spell, the participant will be better matched to particular services within the NDLP, or at least will benefit in some other way from having already experienced the programme once. On the other hand, suppose that individual impacts have a fixed component, which is common to every spell of participation, and a random component, which varies across spells. Individuals with good fixed components will rarely participate twice, because the large impact they receive the first time will get them off benefit. In contrast, individuals with a low fixed component may have to participate multiple times before they get a good enough draw on the random component to get them off benefit. Similarly, if individuals who repeat NDLP have exceptionally low opportunity costs, perhaps due to very low skills, then realised spells of repeat participation may have lower impacts, on average than initial NDLP participation spells.

7.6 General equilibrium effects

7.6.1 Conceptual issues

Another issue that is not addressed in the analyses in this report, but which could potentially bias the estimates it contains is that of general equilibrium effects. General equilibrium effects occur when programmes have effects even on individuals who do not participate in them. Such effects are not captured by the standard partial equilibrium econometric evaluation methods outlined in, eg, Heckman and Robb (1985) or Heckman, LaLonde and Smith (1999). Indeed, in many cases, general equilibrium effects will cause biases in estimates obtained from standard partial equilibrium estimators.

These effects can take a number of forms, of which two are focused on here: The first is displacement. In a model of job search, such as the ones in Davidson and Woodbury (1993) and Lise, Seitz and Smith (2004), when a programme enhances the search effort of one group, the result is that members of other groups with whom they compete in the labour market may find jobs more slowly. Jobs that non-participants would have found and taken in the absence of the programme now get found and taken by eager programme participants who get to them first. The second form of general equilibrium effects consists of changes in equilibrium prices. A programme that, for example, trains lots of lone mothers to be hair stylists might decrease the equilibrium price of hair styling by substantially increasing the supply. A good example in the literature is the paper by Heckman, Lochner and Taber

(1998), which considers the general equilibrium effects of a tuition subsidy for university in the United States. They find substantial effects on skill prices; as the fraction of individuals attending university increases, the relative wages of university graduates decline while the relative wages of individuals completing only secondary school increase. The general equilibrium effects in their analysis differ substantially from those estimated using partial equilibrium methods.

7.6.2 Implications for the analyses in this report

General equilibrium effects from NDLP are more likely to take the form of displacement than of changes in equilibrium wages. Lone parents who search harder for work as a result of NDLP participation, may take jobs that would otherwise have been taken by other lone parents not participating in NDLP (or of other low-skill workers who are not lone parents). This may lengthen the spells on benefit of these other individuals. To the extent that NDLP participants displace other lone parents on benefits, they change the nature of the counterfactual estimated by the matched non-participants by making it worse (less employment, more time on benefit) than it would be in the absence of the general equilibrium effects.

To the extent that the NDLP programme has substantively important general equilibrium effects that take this form, the estimates presented in this analysis are biased up as a result. This bias carries over to a social cost-benefit analysis of the NDLP programme, which will be too positive if NDLP participants displace other lone parents or other low-skill workers besides lone parents. The empirical literature does not provide much guidance regarding the likely size of any displacement effects from NDLP. None of the small handful of programmes for which estimates of displacement effects are available bears that much similarity to the mix of participants and treatment types present in NDLP. However, the empirical literature does suggest that displacement effects can be substantively important, even from programmes whose sole treatment consists of providing financial incentives for job search, as in the US. (Unemployment Insurance bonus experiments analysed in Davidson and Woodbury (1993)).

7.7 Effects of other contemporaneous ALMP

A final consideration over and above general equilibrium effects of the sort described in Section 7.6 concerns the interaction of NDLP with other ALMP in place at the same time in the UK. Such programmes provide the context within which the impacts of NDLP participation are estimated.

It would be expected that the mean impact of NDLP participation would differ in alternative institutional contexts. For example, the New Deal for Young People (NDYP) was initiated prior to the start of the NDLP, and has remained in place during the time that the NDLP has been operational. In a world where both programmes operate, businesses looking to hire someone on benefit, perhaps following

Government-supported classroom training, have their pick of participants in the two programmes. In contrast, in a world with an NDLP but no NDYP, such employers would have to focus on NDLP participants, which would likely lead to a larger mean impact of NDLP (and perhaps greater participation in NDLP as well). Similarly, other recent policies such as the national minimum wage may affect the mean impact of NDLP participation. The economic literature on training sometimes argues that a minimum wage provides an incentive for skill enhancement, in order to raise value of what low skill workers produce above the minimum wage level. Such an effect could lead to increased impacts of NDLP, as low-skill lone parents try to make themselves marketable in a labour market where individuals who produce less than the value of the minimum wage do not get hired.

The literature does not feature much compelling evidence on the nature and extent of such interactions among multiple ALMPs. For the purposes of this study, the primary implication of these issues is simply that readers and policymakers should take care and apply thought when generalising the impact estimates presented here to alternative policy environments.

8 Interpretations and conclusions

Summary

The New Deal for Lone Parents (NDLP) impact evaluation is not straightforward as it is taking place in a complex environment. The programme itself is a voluntary one in which the recipients can still be in receipt of Income Support (IS) and other benefits. It will, therefore, be unclear how to treat NDLP participants who are not actually on IS – are these individuals to be judged to be in receipt of benefit or not because they are clearly getting a state subsidy? The other major problem in evaluation of NDLP is that it has been taking place at the same time as several other New Deal programmes – most importantly the New Deal for Young People (NDYP) – it is, therefore, unclear how big an impact NDLP might have had if this other programme was not running at the same time, as inevitably such competing programmes will also be used by employers who have job vacancies. This report discusses these important complexities.

This research investigated the problems which arise with the consistency of the data which results from merging information from the computers which contain the NDLP, IS, Jobseeker's Allowance (JSA) and Incapacity Benefit (IB) records of lone parents. A number of clients appear to have overlapping spells. In most cases these are feasible admissible contemporaneous spells since people can be on NDLP at the same time as claiming any benefits such as IS, IB or JSA. A set of rules of developed to solve this problem.

The report investigates the 'full' characterisation of the underlying econometric model which would be necessary to analyse the NatCen data. The model is in fact quite complicated but the version used by NatCen is relatively simple and relies on a number of strong, and potentially controversial, simplifying assumptions. We make these assumptions explicit as it is only by recognising them that a realistic appraisal of the treatment effect found by NatCen for the NDLP programme can be put into context.

Continued

The research indicates that although NatCen adopted a number of approaches that are unconventional relative to the econometric matching literature, the estimates obtained are relatively insensitive to these dimensions.

The weighted estimates that adjust for selective non-response on observable characteristics make very little difference to the estimates. Despite these findings, there remains the possibility that an analysis that incorporated selective non-response based on unobserved factors correlated with outcomes, might tell a different story.

Another important issue explored in this research was the fact that the window of time available for an individual to participate in the NDLP after returning his or her postal questionnaire varies among the individuals in the sample. Taken together, the evidence presented suggests that the window definition makes a substantive difference to the resulting impact estimates.

The research also explored the importance of conditioning on local labour markets, or variables at the local labour market level. This involves merging the evaluation data with the administrative data that contain these variables. We found that adding these local labour market data did not change the results much.

The estimate of the impact effect of NDLP in this report conditioned on detailed employment and IS histories using the administrative data. The focus here was on the characterisation of the past employment history over the previous six quarters and the most recent weeks. The techniques suggested by Card and Sullivan to summarise the 64 possible sequences of six previous quarters of labour market state information were followed. These dummy variables proved to be very powerful in satisfactorily matching participants and non-participants prior to the sample date.

The impact effect of NDLP presented substantially different conclusions for those clients who were in the stock of benefit claimants when NDLP was introduced compared to those who flowed onto benefits since NDLP became live. Our preferred impact estimate was computed using the full administrative dataset for lone parent benefit claimants who were new to the flow of benefit claimants and were not on NDLP most of the time in the preceding six quarters.

The sampling problem of drawing a representative sample of individuals who are eligible to receive the NDLP treatment and correcting this sample for any observable differences between the administrative sampling units, the postal questionnaire respondents and those who received a face-to-face interview is non-trivial. The appropriate method for weighting the data based on these sampling problems is outlined.

Continued

In 2002 this report was commissioned by the Department of Work and Pensions (DWP) to examine the accuracy and robustness of the results from the NatCen NDLP evaluation in a number of dimensions, in particular to determine the sensitivity of their estimates to changes in the estimation procedure and outcome variables, and to look at longer-term outcomes.

The econometric evaluation in this report makes use of several, somewhat innovative, features of the NatCen survey sample design. Most importantly, the NatCen data collection strategy involves matching NDLP participants with non-participants using administrative data and information from a postal survey in order to determine which non-participants should receive face-to-face interviews. The econometric literature does not document this prescriptive use of statistical matching.

The analysis in this report makes explicit the statistical and economic assumptions underlying the NatCen analysis, as well as checking the computational details of what the NatCen did and examines the robustness of their computational and estimation methods. By recreating the logic of the NatCen analysis from scratch, the conceptual issues involved in their analysis are clarified and highlight areas that would benefit from further research.

A thorough review of all the data available to assess the impact of the NDLP for the cohort which was eligible to participate in this programme in August 2000 has been conducted. Focusing on the simple outcome measure of the probability of being on benefit we suggest that the NatCen impact of 26 per cent was too high. Our preferred impact estimate of NDLP is 11 per cent, using the full administrative dataset.

The econometric work investigated in this report is very much on the boundary of what is known in applied econometric methodology regarding the evaluation of active labour market policy. The reasons are that this NDLP policy may involve: treatment for different lengths of time for different individuals; non-participants becoming participants at a later stage; and some NDLP recipients being participants on more than one occasion. These issues are explored and the econometric implications of evaluation in this context are spelt out. However, each of these econometric problems deserves much closer scrutiny in further research work.

Further research on the evaluation of the NDLP programme is possible in a more ambitious context. It is possible to link the data made available to us on the benefit and claimant status of each lone parent with what they subsequently earn via Inland Revenue data. If this was possible then a more sophisticated analysis of the outcomes of the NDLP programme may be possible. Specifically it could be possible to perform a cost-benefit analysis of exactly what was paid out to the lone benefit recipient under the NDLP programme and link this to what contribution they subsequently make to the Exchequer in the form of income taxes.

It is important to note that this NDLP evaluation examines a particular cohort of NDLP participants, as such, great care should be taken in generalising this results of this evaluation to any other programme or even the same programme at any other point in time.

At the outset, the readers should appreciate several major difficulties that arise in conducting a non-experimental evaluation of a large-scale on-going programme such as NDLP: Firstly, as has been documented in Heckman and Robb (1985) and Heckman, LaLonde and Smith (1999), all non-experimental evaluation requires strong assumptions to identify treatment effects. Secondly, non-experimental impact evaluations require data that is both rich in terms of covariates and of high quality. This is particularly so for econometric strategies based on matching, such as that employed by NatCen. Thirdly, accurately estimating the impact of one programme requires a relatively stable policy environment in terms of Active Labour Market Policies affecting the same population as the programme under investigation (or populations that compete with it in the labour market). Fourthly, interpretable impact estimates require that the operation of the programme match the causal model underlying the evaluation. If, for example, IS recipients take up NDLP when they find a job, rather than finding a job because they take up NDLP, this condition will fail. Finally, partial equilibrium evaluations of large-scale programmes such as NDLP raise the issue of general equilibrium effects. Increased search effort among NDLP participants caused by the programme may affect the search effort and job-finding probabilities of other groups seeking the same jobs. See, among others, the discussions in Calmfors (1994), Heckman, LaLonde and Smith (1999), and Lise, Seitz and Smith (2002). The NatCen design, which explicitly considers impacts only on programme participants, ignores these general equilibrium considerations. In summary, like any other non-experimental study, the NDLP evaluation undertaken by the NatCen is far from ideal in its construction, which implies both that care needs to be exercised in the interpretation of the NatCen findings and that replications and sensitivity analyses such as are performed in this report will have great value.

How can sense be made of the 14.24 per cent treatment effect of NDLP? There is still a lack of confidence that the estimates may not be an overestimate of the effect of the ATT on NDLP. There are five basic reasons for caution:

- These results relate only to the cohort who were eligible for NDLP in September 2000. Hence, there is priority on testing the robustness of the results on other monthly cohorts of the NDLP using just the administrative data.
- It is the case that the aggregate labour market from 1998-2004 has been a buoyant market with high demand for relatively low-skilled manpower and that this may mean that compared to the 1992-1998 pre-treatment period, there were many more jobs available and the incentive to participate in the labour market was much higher for lone parents in the period after NDLP was introduced.
- The lone parent labour market in the UK was such that there was huge scope for a rising employment rate. Gregg and Harkness (2003) describe how the employment rate of lone parents in the UK was relatively very low. This suggests that there was a lot of scope for the upward movement of this ratio over the short term.

- A fourth argument is that the sample is from a cohort who were relatively near the beginning of the NDLP programme and as such they have caught the full initial impact of the programme. This means that there were a large stock of NDLP participants who were on the margin between working and not working and that the Working Families' Tax Credit (WFTC) reforms and NDLP simply came along at the right time and was sufficient to induce them back into the labour market.
- The final area of caution relates to the econometric problems described in the previous chapter which remain with the empirical work of this project. Specifically, the modelling of multiple spells on NDLP, switching between participant and non-participant groups and the variable length of time on NDLP pose large problems for the assessment of NDLP. However, there is no *a priori* reason why these problems should give rise to a higher estimate of the NDLP rather than a lower impact assessment.

Notwithstanding these caveats emphasis must be given to the fact that the impact estimates reported in this study have been robust to a large array of sensitivity checks and clearly there are substantial grounds for believing that the NDLP has had a significant impact on the take-up and retention of benefits. In the light of these results it is suggested that the DWP can have confidence in the fact that the NDLP has had a substantial impact for this cohort over the 2000-04 period.

8.1 Lessons for economic policy evaluation methodology.

8.1.1 Use of the administrative data

One very important area for further analysis concerns more extensive use of the administrative data relating to the labour market state each person was in prior to the NDLP option and after any participation in the NDLP. The administrative data contain information on each person in the lone parent client group in terms of their post code and their month-on-month claimant status. Little use was made of these data in the NatCen report. This report used the administrative data to compute basic non-parametric impact estimates by simply comparing the proportion of the NDLP participants who are in work in each month with the proportion of non-participants in work each month.

Estimates obtained using the administrative data concurred with those obtained using the survey data. Indeed, the use of the detailed information on the timing of spells on and off of IS in the six quarters prior to the sampling date, proved most valuable when coded into strings as in Card and Sullivan (1988).

The logical conclusion from the use of these data is that it may be possible – under certain circumstances – to replace the collection of expensive survey data with administrative data. Clearly, such a research strategy has advantages in overcoming problems with initial non-response, attrition and also aspects of misreported information. Naturally, there are limitations to using administrative data which

include not having access to key controlling regressors. However, one of the implications of this work is that administrative data should be carefully used in further evaluation work and that there is a good case for adding additional regressor information by attempting to collect this in administrative data. For example, the collection of basic educational and qualification information in the NDLP data would be an excellent way of augmenting the value of these data.

The findings reported might lead to reflection on when administrative data could be used instead of survey data and indeed under what circumstances could the use of survey data be justified.

It is envisaged that survey data would still need to be collected: when the outcome of interest is not like those available in administrative data – specifically qualitative evaluations of NDLP and eliciting individual's preferences and opinions on the working of such a programme would never be possible using administrative data. It is also possible that some surveys will always be necessary to benchmark the evaluations being done by other means using administrative data.

8.2 Future research evaluation of the impact of NDLP

With such a complex project evaluation, it is perhaps inevitable that there is considerable scope for further valuable research. This section provides a brief outline of some of the options for further research that became apparent in the course of this research. Some of these suggestions come directly from discussions with the NatCen and DWP teams involved in funding and coordinating this project.

8.2.1 Prescriptive use of matching to target potential interviewees

One of the distinctive features of this research design was the use of matching to determine who to interview in later surveys. This technique has rarely been used before and its merits and properties should be more thoroughly investigated. A thorough and thoughtful analysis of this approach would suggest ways to determine the optimal size of the target sample to allow for potential non-response and the optimal way to over-sample potential respondents at extremes of the p-score distribution (particularly non-participants with high estimated probabilities of participation) where there may be common support problems.

8.2.2 More careful assessment of the duration nature of the outcome variables in question

The assessment of Active Labour Market Programme (ALMP) which has a possible impact on its participants over a period of time is less straightforward than the assessment of ALMP on a variable which is measured at a particular point in time. The econometric analysis of this type of outcome – like a labour market state – which can vary at any time in the future is, relatively speaking, in its infancy. Early papers, such as Dolton et al. (1994), attempted to assess the impact of ALMP over time using relatively simple duration models. Recent papers by Abbring and van den Berg

(2003a), Eberwein, Ham and LaLonde (2002) and Eberwein (2002) have begun to rigorously set out the assumptions necessary to apply standard duration modelling techniques to the assessment of treatment effects and to lay out other alternative methods for the analysis of programme effects in event history data. The analysis conducted by NatCen used only rudimentary techniques from Life Table analysis to estimate the effects of NDLP over time. It is strongly suggested that the new techniques now being developed ought to be applied to these data.

8.2.3 Long run analysis of the labour market impact of NDLP

Most extant analyses of ALMP consider only the short-term impacts of policy interventions. Relatively little evidence exists on the long-run effects of programmes like NDLP. As the claimant status of the lone parent client group included in NDLP evaluation can be followed up using the administrative data over the next few years, a long-run assessment of NDLP may be possible. Such a long-term assessment should not be underestimated in terms of its policy significance, as part of the philosophy of the NDLP programme is to change lone parents' attitudes towards work and dependency. The impact of NDLP on its participants can only be assessed, in this respect, by taking a long-run look at the same individuals to see if the programme continues to have an effect on their labour market outcomes two to four years after participation. So far, the analysis in this report has examined outcomes up to five years later. In principle, such analysis could be extended indefinitely into the future.

8.2.4 Cost-benefit analysis of the NDLP programme

Ultimately, the overall social value of any ALMP should be assessed by taking full account of not only the estimated treatment effects on the treated in terms of labour market outcomes, but also of other benefits of the programme, the size of any deadweight loss or general equilibrium effects, and the direct costs of the programme, including any excess burden associated with the taxes required to fund it. Such tasks are not easily accomplished. However, it may be possible to further explore these issues given the detailed data available in this evaluation. The method of analysis could be similar to that used in Dolton and O'Neill (2002). Such an analysis would be extremely important from a policy point of view.

The first step in such an assessment would be to work out all the expenditure incurred on NDLP for each participant. Next, we need to know the history of when this individual signs off receipt of benefit or even makes a contribution to the exchequer if they are paying taxes. In principle it is now possible to work out the cost of each person's participation in NDLP but what is not known is if they get a job when they sign off. The latter data will potentially be available for limited five per cent sample of individuals from the Inland Revenue. Such data will facilitate this cost-benefit analysis for the cohort examined in this report.

Ultimately it would be preferable to perform a cost-benefit analysis on the whole programme of participants over the whole lifetime of the programme since its inception (ie. not just those who were eligible in September 2000). This may prove possible at some point in the future.

Appendix A

Logit: Dependent variable NDLP participation using SPSS and STATA

Table A.1 Logit: Dependent variable NDLP participation using SPSS and STATA

Varname	Description	Value	Value labels	Coefficient	S.E.
QWHENWK	When hoping to start work	1	next 6 months	-0.492	0.074
		2	next year	-0.631	0.086
		3	next 3 years	-1.396	0.21
		4	not in next 3 years	-0.951	0.238
QHELP	Level of interest in help or advice	1	none or one	0.357	0.089
		2	two or three	0.816	0.091
		3	four or more	-0.872	0.115
NTIMERES	Deciles of time to respond to questionnaire	1		-0.095	0.094
		2		-0.06	0.09
		3		-0.268	0.094
		4		-0.398	0.106
		5		-0.494	0.112
		6		-0.488	0.104
		7		-0.69	0.118
		8		-0.848	0.121
		9		-0.963	0.124

Continued

Table A.1 Continued

Varname	Description	Value	Value labels	Coefficient	S.E.
QACQUAL	Highest academic qualification	1	grade d or below o level etc	0.208	0.076
		2	grade c or above o level etc	0.349	0.07
		3	a levels etc	0.191	0.116
		4	higher educational qualification	0.631	0.113
		5	missing	0.069	0.119
QEXPSTUD	Experience of studying or training	1	studying/training now	0.007	0.083
		2	start course in next year	-0.353	0.094
		3	start course in next few years	-0.231	0.086
		4	not interested	0.056	0.12
NC91ETHN	deciles of census 91 ethnicity	1		0.173	0.086
		2		0.047	0.089
		3		-0.08	0.093
		4		-0.068	0.091
		5		-0.155	0.096
		6		-0.369	0.104
		7		-0.397	0.108
QDRIVLIC	whether respondent has full driving licence	1	yes	0.224	0.055
QLHEALTH	limited - health condition	1	does apply	0.4	0.087
		2	does not apply	0.419	0.126
QATTMOTH	attitudes towards mother			0.03	0.009
QHOURSP	number of hours respondent prefers to work	1	under 15 hours	0.45	0.115
		2	16 to 29 hours	0.061	0.142
		3	30 hours or more	0.335	0.128
		4	depends on job	0.41	0.131
		5	not decided	0.56	0.21
QYEARSIW	years since last worked	1	one year or less	-0.053	0.081
		2	2 to 4 years	-0.189	0.097
		3	5 to 9 years	-0.306	0.111
		4	10 years or more	-0.535	0.149
		5	working now/ never worked	-0.262	0.099
QHEDUC	help/advice - education or training	1	does apply	0.12	0.077
		2	does not apply	0.378	0.098
QCHILD3	number of children aged <= 3	1	one	-0.111	0.062
		2	two or more	-0.366	0.149

Continued

Table A.1 Continued

Varname	Description	Value	Value labels	Coefficient	S.E.
QCCOSC	child care - after-school club, etc	1	yes	-0.531	0.156
CARER_PR	caring for relative etc	1	yes	-0.518	0.172
QMINWAGE	whether willing to work for minimum wage	1	yes	0.017	0.057
		2	no	0.341	0.105
QLSKILL	limited - not enough skills	1	does apply	-0.102	0.061
		2	does not apply	-0.371	0.111
QSAPPL	look for work - applied for a job	1	yes	-0.167	0.097
		2	no	-0.352	0.114
QTENURE	tenure	1	owner occupier	-0.127	0.17
		2	council or ha	-0.02	0.152
		3	prs	-0.013	0.16
		4	live with parents/ relatives	-0.467	0.19
PRT_HOUR	work part-time	1	yes	0.265	0.118
QTELEPH	whether use telephone at home	1	yes	0.193	0.076
QATTJOB	attitudes towards retaining a job			0.026	0.011
QSCBALO	services contacted - BA local office	1	does apply	0.214	0.099
		2	does not apply	0.379	0.12
QSERVICE	number of services contacted	1	one	0.182	0.187
		2	two	0.209	0.211
		3	three or more	0.205	0.262
		4	all missing	-0.1	0.2
QSCCAB	services contacted - citizens advice bureau	1	does apply	0.106	0.106
		2	does not apply	0.187	0.129
QSCADLIN	services contacted - advice line	1	does apply	0.154	0.117
		2	does not apply	0.168	0.14
QSCNDADV	services contacted - NDLP adviser	1	does apply	-0.741	0.109
		2	does not apply	-0.568	0.127
QSCADV	services contacted - careers adviser	1	does apply	0.081	0.123
		2	does not apply	-0.24	0.145
QLJOBS	limited - not enough jobs	1	does apply	0.156	0.06
		2	does not apply	0.125	0.107

Continued

Table A.1 Continued

Varname	Description	Value	Value labels	Coefficient	S.E.
NDAYSBEN	three categories of total number of days on benefit			-0.047 -0.218	0.064 0.075
QLFINANC	limited - worse off financially	1	does apply	-0.16	0.06
		2	does not apply	0.052	0.089
QLLEAVE	limited - leave children	1	does apply	0.148	0.058
		2	does not apply	-0.076	0.107
QHOURWK	number of hours respondent works in current or previous job	1	under 5 hours	-0.239	0.136
		2	6 to 15 hours	-0.385	0.135
		3	16 to 29 hours	-0.262	0.133
		4	30 hours or more	-0.089	0.16
QSEMPAG	look for work - private employment agency	1	yes	-0.124	0.122
		2	no	0.204	0.138
QCHILD11	number of children aged ≤ 11	1	one	0.124	0.09
		2	two	0.052	0.097
		3	three or more	-0.149	0.127
QSFREL	look for work - friends and relatives	1	yes	0.191	0.078
		2	no	0.048	0.102
DNNONDP	number of non-dependent children	1	one	0.2	0.119
		2	two or more	-0.431	0.294
CONS	Constant			-3.229	0.508

Appendix B

Variable descriptions

Table B.1 Variable descriptions

Name	Description
qwhenwk	when hoping to start work
qhelp	level of interest in help or advice
ntimeres	deciles of time to respond to questionnaire
qacqual	highest academic qualification
qexpstud	experience of studying or training
nc91ethn	deciles of census 91 ethnicity
qdrivlic	whether respondent has full driving licence
qlhealth	limited - health condition
qattmoth	attitudes towards mother working
qhoursp	number of hours respondent prefers to work
qyearsiv	years since last worked
qheduc	help/advice - education or training
qchild3	number of children aged ≤ 3
qccosc	child care - after school club etc
carer_pr	caring for relative etc
qminwage	whether willing to work for minimum wage
qlskill	limited - not enough skills
qsappl	look for work - applied for a job
qtenure	tenure
prt_hour	work part-time
qteleph	whether use telephone at home
qattjob	attitudes towards retaining a job
qscbal	services contacted - ba local office
qservice	number of services contacted

Continued

Table B.1 Continued

Name	Description
qsccab	services contacted - citizens advice bureau
qscadlin	services contacted - advice line
qscndadv	services contacted - ndlp adviser
qscadv	services contacted - careers adviser
qljobs	limited - not enough jobs
ndaysben	three categories of total number of days on
qlfinanc	limited - worse off financially
qlleave	limited - leave children
qhourwk	number of hours respondent works in current
qsempag	look for work - private employment agency
qchild11	number of children aged <= 11
qsfrrel	look for work - friends and relatives
dnnondp	number of non-dependent children
prehist3	Card and Sullivan dummies
a2118	Proportion of time on Benefit on W39_2000
a2117	Proportion of time on Benefit on W38_2000
a2116	Proportion of time on Benefit on W37_2000
a2115	Proportion of time on Benefit on W36_2000
a2114	Proportion of time on Benefit on W35_2000
a2113	Proportion of time on Benefit on W34_2000
proprio	Proportion of time of Benefit prior to June, 1999

Appendix C

Balancing tests for samples with estimated propensity scores of less than 0.3

Table C.1 Balancing tests for samples with estimated propensity scores of less than 0.3

	(1)				(2)				(3)				(4)			
	Matching with replacement		NatCen full matched sample		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs	
	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part
Number of observations	1,550	1,656	1,656	1,656	1,656	1,656	1,656	1,656	1,125	1,171	770	770	770	770	770	770
Number of statistically significant mean differences	10% level	0	0	0	0	0	0	0	24	22	22	22	22	22	11	0
	5% level	0	0	0	0	0	0	0	9	12	12	12	12	12	0	0
Qattmoth	13.04	13.11	13.06	13.11	13.06	13.11	13.06	13.11	13.02	13.01	13.01	13.01	13.01	13.01	12.99	0.92
qattjob	8.67	8.74	8.69	8.74	8.69	8.74	8.69	8.74	8.72	8.69	8.69	8.69	8.69	8.63	8.65	0.88
_lqwhenwk_2	0.26	0.26	0.27	0.26	0.27	0.26	0.27	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.91
_lqwhenwk_3	0.25	0.23	0.26	0.23	0.26	0.23	0.26	0.23	0.28	0.23	0.23	0.23	0.27	0.23	0.23	0.08
_lqwhenwk_4	0.10	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.12	0.11	0.11	0.11	0.13	0.10	0.10	0.09
_lqwhenwk_5	0.11	0.11	0.07	0.11	0.07	0.11	0.07	0.11	0.08	0.12	0.12	0.12	0.08	0.13	0.13	0.00
_lqhelp_2 if	0.22	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.22	0.22	0.22	0.23	0.23	0.23	0.76
_lqhelp_3 if	0.49	0.49	0.50	0.49	0.50	0.49	0.50	0.49	0.49	0.48	0.49	0.48	0.49	0.47	0.47	0.39
_lqhelp_4 if	0.14	0.13	0.12	0.13	0.12	0.13	0.12	0.13	0.12	0.14	0.12	0.14	0.12	0.15	0.15	0.08
_Intimeres_2	0.12	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.14	0.13	0.14	0.13	0.14	0.14	0.26
_Intimeres_3	0.15	0.14	0.13	0.14	0.13	0.14	0.13	0.14	0.13	0.15	0.13	0.15	0.14	0.16	0.16	0.28
_Intimeres_4	0.12	0.12	0.13	0.12	0.13	0.12	0.13	0.12	0.14	0.13	0.14	0.13	0.14	0.12	0.12	0.41
_Intimeres_5	0.09	0.09	0.05	0.09	0.05	0.09	0.05	0.09	0.05	0.08	0.05	0.08	0.05	0.09	0.09	0.01
_Intimeres_6	0.08	0.07	0.09	0.07	0.09	0.07	0.09	0.07	0.09	0.07	0.09	0.07	0.09	0.07	0.07	0.30
_Intimeres_7	0.09	0.09	0.10	0.09	0.10	0.09	0.10	0.09	0.09	0.08	0.09	0.08	0.09	0.09	0.09	0.93
_Intimeres_8	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.83
_Intimeres_9	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.06	0.05	0.06	0.05	0.06	0.06	0.39

Continued

Table C.1 Continued

	(1)				(2)				(3)				(4)			
	Matching with replacement		NatCen full matched sample		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs	
	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part
Number of observations	1,550	1,656	1,656	1,656	1,656	1,656	1,656	1,656	1,125	1,171	770	770	770	770	770	770
Number of statistically significant mean differences	10% level	0	0	0	0	0	0	0	24	22	22	22	22	22	11	0
	5% level	0	0	0	0	0	0	0	9	12	12	12	12	12	0	0
_Intimeres_1	0.05	0.06	0.73	0.07	0.06	0.21	0.06	0.06	0.06	0.05	0.38	0.06	0.06	0.04	0.12	0.12
_Iqacqual_1	0.20	0.19	0.45	0.20	0.19	0.26	0.20	0.19	0.20	0.19	0.71	0.19	0.19	0.20	0.70	0.70
_Iqacqual_2	0.29	0.30	0.55	0.29	0.30	0.73	0.30	0.30	0.30	0.31	0.65	0.30	0.30	0.30	0.87	0.87
_Iqacqual_3	0.06	0.07	0.56	0.05	0.07	0.06	0.05	0.05	0.05	0.07	0.10	0.06	0.06	0.07	0.17	0.17
_Iqacqual_4	0.07	0.07	0.97	0.06	0.07	0.29	0.06	0.06	0.06	0.07	0.73	0.06	0.06	0.07	0.54	0.54
_Iqacqual_5	0.06	0.05	0.60	0.05	0.05	0.95	0.05	0.06	0.06	0.05	0.49	0.06	0.06	0.04	0.19	0.19
_Iqexpstud_2	0.34	0.35	0.48	0.37	0.35	0.33	0.35	0.33	0.35	0.34	0.63	0.34	0.34	0.35	0.79	0.79
_Iqexpstud_3	0.16	0.16	0.70	0.16	0.16	0.52	0.16	0.16	0.17	0.16	0.55	0.19	0.19	0.16	0.12	0.12
_Iqexpstud_4	0.25	0.25	0.87	0.24	0.25	0.46	0.24	0.25	0.23	0.25	0.25	0.23	0.23	0.26	0.17	0.17
_Iqexpstud_5	0.07	0.07	0.64	0.07	0.07	0.74	0.07	0.07	0.06	0.07	0.45	0.07	0.07	0.07	0.48	0.48
_Inc91ethn_4	0.10	0.12	0.11	0.13	0.12	0.51	0.13	0.12	0.13	0.13	0.85	0.13	0.13	0.13	0.76	0.76
_Inc91ethn_5	0.11	0.10	0.46	0.11	0.10	0.42	0.11	0.10	0.11	0.11	0.83	0.11	0.11	0.12	0.75	0.75
_Inc91ethn_6	0.09	0.10	0.39	0.09	0.10	0.69	0.09	0.10	0.09	0.09	0.52	0.10	0.10	0.09	0.43	0.43
_Inc91ethn_7	0.10	0.10	0.89	0.11	0.10	0.35	0.11	0.10	0.11	0.10	0.38	0.11	0.11	0.11	0.93	0.93
_Inc91ethn_8	0.10	0.09	0.24	0.09	0.09	0.82	0.09	0.09	0.10	0.09	0.46	0.09	0.09	0.09	0.93	0.93
_Inc91ethn_9	0.09	0.08	0.29	0.07	0.08	0.29	0.07	0.08	0.06	0.06	0.57	0.06	0.06	0.06	0.92	0.92
_Inc91ethn_1	0.07	0.07	0.98	0.06	0.07	0.41	0.06	0.07	0.05	0.06	0.18	0.05	0.05	0.06	0.74	0.74

Continued

Table C.1 Continued

	(1)				(2)				(3)				(4)			
	Matching with replacement		NatCen full matched sample		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs	
	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part
Number of observations	1,550	1,656	1,656	1,656	1,656	1,656	1,656	1,656	1,125	1,171	770	770	770	770	770	770
Number of statistically significant mean differences	10% level	5% level	0	0	24	9	24	9	22	12	22	12	22	12	22	12
_lqdrivlic_1	0.48	0.47	0.56	0.44	0.47	0.13	0.43	0.47	0.43	0.47	0.43	0.47	0.43	0.47	0.47	0.18
_lqlhealth_2	0.78	0.76	0.13	0.77	0.76	0.67	0.78	0.76	0.78	0.76	0.77	0.75	0.77	0.75	0.37	0.37
_lqlhealth_3	0.12	0.13	0.19	0.14	0.13	0.60	0.13	0.13	0.13	0.13	0.12	0.13	0.12	0.13	0.13	0.59
_lqhoursp_2	0.35	0.36	0.55	0.36	0.36	0.97	0.35	0.36	0.35	0.36	0.35	0.36	0.35	0.36	0.36	0.87
_lqhoursp_3	0.09	0.09	0.86	0.08	0.09	0.49	0.08	0.09	0.08	0.09	0.08	0.08	0.08	0.08	0.08	0.85
_lqhoursp_4	0.14	0.13	0.57	0.14	0.13	0.60	0.13	0.12	0.13	0.12	0.13	0.13	0.13	0.13	0.13	0.70
_lqhoursp_5	0.12	0.12	0.75	0.14	0.12	0.07	0.15	0.12	0.15	0.12	0.14	0.13	0.14	0.13	0.13	0.65
_lqhoursp_6	0.24	0.24	0.79	0.22	0.24	0.05	0.23	0.26	0.23	0.26	0.23	0.25	0.23	0.25	0.25	0.31
_lqyearsiv_2	0.20	0.21	0.30	0.22	0.21	0.34	0.23	0.20	0.23	0.20	0.22	0.20	0.22	0.20	0.20	0.17
_lqyearsiv_3	0.13	0.13	0.92	0.14	0.13	0.20	0.14	0.13	0.14	0.13	0.15	0.13	0.15	0.13	0.13	0.46
_lqyearsiv_4	0.09	0.09	0.86	0.08	0.09	0.18	0.09	0.10	0.09	0.10	0.09	0.10	0.09	0.10	0.10	0.60
_lqyearsiv_5	0.21	0.20	0.54	0.18	0.20	0.13	0.19	0.20	0.19	0.20	0.17	0.22	0.17	0.22	0.22	0.02
_lqyearsiv_6	0.13	0.12	0.56	0.12	0.12	0.94	0.12	0.12	0.12	0.12	0.81	0.12	0.12	0.12	0.12	0.64
_lqheduc_2 i	0.30	0.31	0.48	0.32	0.31	0.71	0.32	0.30	0.32	0.30	0.27	0.30	0.33	0.30	0.30	0.27
_lqheduc_3 i	0.26	0.26	0.97	0.26	0.26	0.90	0.25	0.27	0.25	0.27	0.30	0.28	0.24	0.28	0.28	0.06
_lqchild3_1	0.36	0.34	0.47	0.40	0.34	0.00	0.40	0.34	0.40	0.34	0.01	0.34	0.40	0.34	0.34	0.01
_lqchild3_2	0.04	0.04	0.92	0.04	0.04	0.24	0.04	0.03	0.04	0.03	0.24	0.04	0.05	0.04	0.04	0.25

Continued

Table C.1 Continued

	(1)				(2)				(3)				(4)			
	Matching with replacement		NatCen full matched sample		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs	
	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part
Number of observations	1,550	1,656	1,656	1,656	1,656	1,656	1,656	1,656	1,125	1,171	770	770	770	770	770	770
Number of statistically significant mean differences	10% level	0	0	0	24	9	24	9	22	12	22	12	22	12	22	12
_lqccosc_1 i	0.02	0.03	0.41	0.03	0.03	0.03	0.99	0.99	0.03	0.03	0.03	0.02	0.03	0.02	0.10	0.10
_lcarer_pr_1	0.02	0.02	0.91	0.02	0.02	0.02	0.55	0.55	0.02	0.02	0.03	0.03	0.03	0.03	0.87	0.87
_lqminwage_2	0.53	0.54	0.71	0.51	0.54	0.06	0.06	0.06	0.51	0.52	0.51	0.53	0.51	0.53	0.58	0.58
_lqminwage_3	0.08	0.08	0.99	0.08	0.08	0.94	0.94	0.94	0.09	0.09	0.09	0.08	0.09	0.08	0.52	0.52
_lqskill_2	0.39	0.39	0.79	0.37	0.39	0.36	0.36	0.36	0.37	0.40	0.36	0.42	0.36	0.42	0.01	0.01
_lqskill_3	0.10	0.11	0.51	0.11	0.11	0.90	0.90	0.90	0.10	0.11	0.11	0.11	0.11	0.11	1.00	1.00
_lqsappl_2 i	0.43	0.41	0.24	0.46	0.41	0.01	0.01	0.01	0.46	0.42	0.45	0.42	0.45	0.42	0.22	0.22
_lqsappl_3 i	0.40	0.41	0.48	0.39	0.41	0.11	0.11	0.11	0.40	0.42	0.41	0.42	0.41	0.42	0.72	0.72
_lqtenure_1	0.12	0.11	0.51	0.09	0.11	0.10	0.10	0.10	0.10	0.13	0.09	0.14	0.09	0.14	0.01	0.01
_lqtenure_2	0.62	0.63	0.62	0.65	0.63	0.17	0.17	0.17	0.66	0.62	0.66	0.63	0.66	0.63	0.20	0.20
_lqtenure_3	0.19	0.19	0.72	0.18	0.19	0.64	0.64	0.64	0.17	0.17	0.18	0.16	0.18	0.16	0.42	0.42
_lqtenure_4	0.06	0.05	0.22	0.06	0.05	0.20	0.20	0.20	0.06	0.05	0.06	0.04	0.06	0.04	0.23	0.23
_lprt_hour_1	0.08	0.08	0.83	0.07	0.08	0.11	0.11	0.11	0.08	0.09	0.07	0.09	0.07	0.09	0.17	0.17
_lqteleph_1	0.85	0.86	0.61	0.85	0.86	0.64	0.64	0.64	0.88	0.87	0.88	0.87	0.88	0.87	0.40	0.40
_lqscbalo_2	0.15	0.16	0.90	0.16	0.16	0.49	0.49	0.49	0.16	0.16	0.16	0.16	0.16	0.16	0.89	0.89
_lqscbalo_3	0.52	0.52	0.89	0.54	0.52	0.16	0.16	0.16	0.55	0.52	0.54	0.53	0.54	0.53	0.58	0.58
_lqservice_1	0.28	0.29	0.76	0.27	0.29	0.41	0.41	0.41	0.28	0.30	0.30	0.29	0.30	0.29	0.91	0.91

Continued

Table C.1 Continued

	(1)				(2)				(3)				(4)			
	Matching with replacement		NatCen full matched sample		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs	
	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part
Number of observations	1,550	1,656	1,656	1,656	1,656	1,656	1,656	1,656	1,125	1,171	770	770	770	770	770	770
Number of statistically significant mean differences	10% level	5% level	0	0	0	0	0	0	24	24	22	22	22	22	11	0
_lqservice_2	0.18	0.17	0.85	0.17	0.17	0.17	0.53	0.16	0.16	0.16	0.15	0.16	0.16	0.16	0.48	0.48
_lqservice_3	0.11	0.12	0.53	0.11	0.12	0.12	0.42	0.10	0.12	0.12	0.12	0.11	0.12	0.11	0.87	0.87
_lqservice_4	0.41	0.40	0.63	0.43	0.40	0.40	0.07	0.44	0.40	0.40	0.43	0.41	0.43	0.41	0.54	0.54
_lqscab_2 i	0.29	0.30	0.55	0.28	0.30	0.30	0.13	0.27	0.31	0.31	0.05	0.29	0.29	0.32	0.22	0.22
_lqscab_3 i	0.56	0.54	0.46	0.57	0.54	0.54	0.10	0.58	0.54	0.54	0.06	0.57	0.55	0.55	0.31	0.31
_lqscadlin_2	0.34	0.34	0.92	0.31	0.34	0.34	0.09	0.30	0.33	0.33	0.11	0.30	0.33	0.33	0.15	0.15
_lqscadlin_3	0.57	0.57	0.86	0.60	0.57	0.57	0.13	0.61	0.57	0.57	0.05	0.60	0.57	0.57	0.22	0.22
_lqscndadv_2	0.26	0.25	0.70	0.24	0.25	0.25	0.46	0.24	0.26	0.26	0.52	0.25	0.25	0.25	0.73	0.73
_lqscndadv_3	0.54	0.54	0.80	0.57	0.54	0.54	0.09	0.57	0.54	0.54	0.12	0.55	0.55	0.55	0.54	0.54
_lqscadv_2 i	0.34	0.35	0.74	0.32	0.35	0.35	0.10	0.31	0.35	0.35	0.07	0.32	0.34	0.34	0.45	0.45
_lqscadv_3 i	0.58	0.57	0.68	0.60	0.57	0.57	0.13	0.60	0.57	0.57	0.13	0.59	0.58	0.58	0.61	0.61
_lqljobs_2 i	0.49	0.48	0.72	0.47	0.48	0.48	0.61	0.46	0.47	0.47	0.63	0.45	0.48	0.48	0.39	0.39
_lqljobs_3 i	0.15	0.16	0.50	0.16	0.16	0.16	0.87	0.15	0.16	0.16	0.78	0.16	0.15	0.15	0.73	0.73
_lndaysben_2	0.34	0.35	0.76	0.40	0.35	0.35	0.00	0.39	0.34	0.34	0.01	0.38	0.35	0.35	0.19	0.19
_lndaysben_3	0.25	0.25	0.97	0.26	0.25	0.25	0.76	0.27	0.25	0.25	0.27	0.27	0.25	0.25	0.52	0.52
_lqlfinanc_2	0.36	0.35	0.37	0.33	0.35	0.35	0.22	0.31	0.34	0.34	0.10	0.29	0.32	0.32	0.27	0.27
_lqlfinanc_3	0.16	0.16	0.63	0.17	0.16	0.16	0.83	0.17	0.16	0.16	0.71	0.17	0.16	0.16	0.37	0.37

Continued

Table C.1 Continued

	(1)				(2)				(3)				(4)			
	Matching with replacement		NatCen full matched sample		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs	
	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part
Number of observations	1,550	1,656	1,656	1,656	1,656	1,656	1,656	1,656	1,125	1,171	770	770	770	770	770	770
Number of statistically significant mean differences	10% level	5% level	0	0	0	0	0	0	24	24	22	22	22	22	11	11
_lqleave_2	0.46	0.47	0.73	0.44	0.47	0.13	0.42	0.46	0.05	0.43	0.46	0.17	0.46	0.17	0.46	0.17
_lqleave_3	0.10	0.12	0.15	0.11	0.12	0.52	0.10	0.12	0.20	0.10	0.11	0.74	0.10	0.11	0.11	0.74
_lqhourwk_2	0.13	0.12	0.42	0.12	0.12	0.90	0.12	0.12	0.68	0.11	0.12	0.75	0.11	0.12	0.12	0.75
_lqhourwk_3	0.22	0.24	0.18	0.21	0.24	0.09	0.21	0.25	0.05	0.22	0.24	0.20	0.22	0.24	0.24	0.20
_lqhourwk_4	0.44	0.44	0.77	0.44	0.44	0.88	0.44	0.41	0.24	0.45	0.41	0.15	0.45	0.41	0.41	0.15
_lqhourwk_5	0.16	0.15	0.60	0.18	0.15	0.07	0.17	0.15	0.23	0.17	0.16	0.68	0.17	0.16	0.16	0.68
_lqsempag_2	0.50	0.48	0.24	0.52	0.48	0.03	0.51	0.49	0.26	0.51	0.48	0.24	0.51	0.48	0.48	0.24
_lqsempag_3	0.43	0.45	0.23	0.42	0.45	0.06	0.43	0.46	0.25	0.43	0.46	0.36	0.43	0.46	0.46	0.36
_lqchild11_1	0.54	0.54	0.77	0.57	0.54	0.10	0.56	0.53	0.08	0.56	0.53	0.18	0.56	0.53	0.53	0.18
_lqchild11_2	0.27	0.27	0.77	0.28	0.27	0.54	0.28	0.28	0.97	0.28	0.28	0.91	0.28	0.28	0.28	0.91
_lqchild11_3	0.08	0.08	0.50	0.08	0.08	0.81	0.08	0.08	0.57	0.08	0.08	1.00	0.08	0.08	0.08	1.00
_lqsfrel_2	0.31	0.29	0.26	0.34	0.29	0.00	0.34	0.29	0.01	0.34	0.30	0.07	0.34	0.30	0.30	0.07
_lqsfrel_3	0.35	0.36	0.55	0.33	0.36	0.06	0.34	0.37	0.13	0.33	0.35	0.36	0.33	0.35	0.35	0.36
_ldnndp_1	0.04	0.05	0.37	0.04	0.05	0.37	0.04	0.05	0.45	0.04	0.05	0.27	0.04	0.05	0.05	0.27
_ldnndp_2	0.01	0.01	0.54	0.01	0.01	0.84	0.01	0.01	0.89	0.01	0.01	1.00	0.01	0.01	0.01	1.00

Appendix D

Balancing tests for NatCen full sample and face-to-face interview respondent sample

Table D.1 Balancing tests for NatCen full sample and face-to-face
interview respondent sample

	(1)			(2)		
	NatCen full matched sample			NatCen matched sample: face-to-face interview respondents		
	Non-Part	Part	p-value of the t-test	Non-Part	Part	p-value of the t-test
Number of observations:	1,819	1,777		1,241	1,259	
Number of statistically	10% level		17			14
significant mean differences	5% level		11			12
Qattmoth	13.212	13.228	0.880	13.174	13.137	0.771
qattjob	8.753	8.790	0.630	8.783	8.732	0.589
_lqwhenwk_2	0.255	0.246	0.552	0.239	0.253	0.442
_lqwhenwk_3	0.234	0.212	0.104	0.254	0.214	0.018
_lqwhenwk_4	0.104	0.103	0.928	0.110	0.106	0.750
_lqwhenwk_5	0.067	0.106	0.000	0.075	0.113	0.001
_lqhelp_2 if	0.218	0.220	0.897	0.222	0.211	0.531
_lqhelp_3 if	0.528	0.515	0.441	0.518	0.512	0.771
_lqhelp_4 if	0.111	0.122	0.301	0.107	0.128	0.108
_Intimeres_2	0.134	0.130	0.750	0.141	0.145	0.800
_Intimeres_3	0.151	0.150	0.975	0.150	0.157	0.609
_Intimeres_4	0.123	0.127	0.715	0.132	0.133	0.971

Continued

Table D.1 Continued

	(1)			(2)		
	NatCen full matched sample			NatCen matched sample: face-to-face interview respondents		
	Non-Part	Part	p-value of the t-test	Non-Part	Part	p-value of the t-test
Number of observations:	1,819	1,777		1,241	1,259	
Number of statistically significant mean differences	10% level		17			14
	5% level		11			12
_Intimeres_5	0.052	0.086	0.000	0.051	0.077	0.007
_Intimeres_6	0.086	0.070	0.085	0.088	0.065	0.033
_Intimeres_7	0.094	0.088	0.517	0.085	0.079	0.585
_Intimeres_8	0.058	0.060	0.805	0.055	0.059	0.667
_Intimeres_9	0.051	0.056	0.494	0.043	0.056	0.136
_Intimeres_1	0.060	0.052	0.288	0.054	0.048	0.471
_Iqacqual_1	0.203	0.190	0.319	0.199	0.193	0.705
_Iqacqual_2	0.303	0.304	0.949	0.307	0.314	0.716
_Iqacqual_3	0.051	0.065	0.070	0.051	0.065	0.125
_Iqacqual_4	0.070	0.080	0.250	0.073	0.077	0.667
_Iqacqual_5	0.052	0.053	0.869	0.053	0.048	0.590
_Iqexpstud_2	0.378	0.362	0.343	0.367	0.355	0.519
_Iqexpstud_3	0.151	0.147	0.788	0.158	0.153	0.749
_Iqexpstud_4	0.229	0.242	0.368	0.224	0.245	0.207
_Iqexpstud_5	0.068	0.071	0.747	0.066	0.072	0.541
_Inc91ethn_4	0.132	0.120	0.275	0.140	0.132	0.542
_Inc91ethn_5	0.113	0.108	0.656	0.116	0.114	0.848
_Inc91ethn_6	0.091	0.095	0.650	0.081	0.091	0.376
_Inc91ethn_7	0.109	0.103	0.568	0.110	0.102	0.563
_Inc91ethn_8	0.085	0.089	0.651	0.093	0.087	0.644
_Inc91ethn_9	0.065	0.074	0.328	0.055	0.059	0.667
_Inc91ethn_1	0.062	0.068	0.510	0.049	0.058	0.327
_Iqdrivlic_1	0.461	0.479	0.274	0.454	0.479	0.220
_Iqlhealth_2	0.774	0.770	0.794	0.787	0.764	0.165
_Iqlhealth_3	0.139	0.129	0.423	0.128	0.130	0.873
_Iqhoursp_2	0.389	0.380	0.587	0.386	0.376	0.625
_Iqhoursp_3	0.079	0.084	0.566	0.074	0.088	0.199
_Iqhoursp_4	0.140	0.136	0.692	0.133	0.121	0.359
_Iqhoursp_5	0.132	0.115	0.118	0.136	0.118	0.181
_Iqhoursp_6	0.198	0.230	0.020	0.210	0.241	0.056
_Iqyearsiv_2	0.224	0.216	0.552	0.226	0.205	0.191
_Iqyearsiv_3	0.142	0.130	0.279	0.139	0.129	0.431
_Iqyearsiv_4	0.075	0.088	0.154	0.084	0.093	0.422
_Iqyearsiv_5	0.167	0.192	0.048	0.169	0.191	0.164

Continued

Table D.1 Continued

	(1)			(2)		
	NatCen full matched sample			NatCen matched sample: face-to-face interview respondents		
	Non-Part	Part		Non-Part	Part	
Number of observations:	1,819	1,777	p-value of the t-test	1,241	1,259	p-value of the t-test
Number of statistically significant mean differences	10% level		17			14
	5% level		11			12
_lqyearsiv_6	0.115	0.114	0.867	0.110	0.114	0.751
_lqheduc_2 i	0.315	0.311	0.806	0.320	0.303	0.351
_lqheduc_3 i	0.248	0.253	0.743	0.243	0.262	0.260
_lqchild3_1	0.386	0.340	0.004	0.388	0.340	0.012
_lqchild3_2	0.040	0.034	0.404	0.039	0.033	0.353
_lqccosc_1 i	0.027	0.028	0.906	0.023	0.026	0.648
_lcarer_pr_1	0.018	0.021	0.411	0.019	0.023	0.431
_lqminwage_2	0.515	0.543	0.093	0.518	0.522	0.853
_lqminwage_3	0.078	0.078	0.964	0.083	0.085	0.858
_lqlskill_2	0.374	0.393	0.256	0.373	0.403	0.119
_lqlskill_3	0.106	0.106	0.980	0.098	0.106	0.501
_lqsappl_2 i	0.434	0.405	0.071	0.442	0.413	0.138
_lqsappl_3 i	0.363	0.400	0.026	0.377	0.402	0.204
_lqtenure_1	0.092	0.107	0.130	0.098	0.124	0.035
_lqtenure_2	0.648	0.627	0.197	0.659	0.619	0.036
_lqtenure_3	0.188	0.192	0.767	0.177	0.176	0.951
_lqtenure_4	0.052	0.045	0.315	0.050	0.048	0.862
_lpert_hour_1	0.068	0.078	0.247	0.076	0.086	0.357
_lqteleph_1	0.863	0.866	0.797	0.885	0.879	0.670
_lqscbal_2	0.176	0.165	0.379	0.177	0.164	0.393
_lqscbal_3	0.526	0.506	0.252	0.533	0.507	0.182
_lqservice_1	0.273	0.289	0.285	0.281	0.295	0.432
_lqservice_2	0.175	0.178	0.847	0.168	0.170	0.917
_lqservice_3	0.128	0.132	0.711	0.125	0.134	0.487
_lqservice_4	0.404	0.380	0.146	0.407	0.380	0.163
_lqsccab_2 i	0.295	0.312	0.265	0.292	0.320	0.124
_lqsccab_3 i	0.557	0.533	0.149	0.563	0.531	0.109
_lqscadlin_2	0.325	0.345	0.215	0.324	0.345	0.271
_lqscadlin_3	0.577	0.557	0.236	0.584	0.553	0.113
_lqscndadv_2	0.300	0.294	0.728	0.294	0.297	0.872
_lqscndadv_3	0.560	0.526	0.040	0.566	0.527	0.050
_lqscadv_2 i	0.332	0.357	0.119	0.326	0.354	0.130
_lqscadv_3 i	0.570	0.551	0.261	0.573	0.552	0.292
_lqljobs_2 i	0.466	0.478	0.466	0.461	0.473	0.532

Continued

Table D.1 Continued

	(1)			(2)		
	NatCen full matched sample			NatCen matched sample: face-to-face interview respondents		
	Non-Part	Part		Non-Part	Part	
Number of observations:	1,819	1,777	p-value of the t-test	1,241	1,259	p-value of the t-test
Number of statistically significant mean differences	10% level		17			14
	5% level		11			12
_lqljobs_3 i	0.154	0.155	0.945	0.151	0.154	0.857
_Indaysben_2	0.388	0.344	0.007	0.378	0.339	0.043
_Indaysben_3	0.254	0.243	0.451	0.267	0.242	0.160
_lqlfinanc_2	0.322	0.341	0.230	0.311	0.336	0.183
_lqlfinanc_3	0.167	0.163	0.785	0.164	0.163	0.916
_lqlleave_2	0.454	0.478	0.145	0.437	0.476	0.050
_lqlleave_3	0.108	0.114	0.571	0.100	0.116	0.196
_lqhourwk_2	0.120	0.118	0.837	0.118	0.125	0.591
_lqhourwk_3	0.221	0.236	0.291	0.222	0.246	0.160
_lqhourwk_4	0.441	0.443	0.904	0.443	0.421	0.262
_lqhourwk_5	0.166	0.148	0.138	0.160	0.148	0.382
_lqsempag_2	0.510	0.482	0.088	0.504	0.487	0.403
_lqsempag_3	0.418	0.448	0.068	0.425	0.452	0.169
_lqchild11_1	0.575	0.542	0.049	0.570	0.528	0.037
_lqchild11_2	0.270	0.271	0.959	0.271	0.282	0.531
_lqchild11_3	0.071	0.074	0.746	0.073	0.078	0.614
_lqsfrrel_2	0.326	0.291	0.023	0.330	0.292	0.036
_lqsfrrel_3	0.320	0.353	0.037	0.324	0.357	0.077
_ldnnondp_1	0.042	0.049	0.301	0.044	0.050	0.440
_ldnnondp_2	0.007	0.007	0.795	0.007	0.008	0.843

Appendix E

Summary statistics (mean and standard deviation)

Table E.1 Summary statistics (mean and standard deviation)

	Non-participants	Participants	Total
qattmoth	11.99 (3.33)	13.23 (3.14)	12.05 (2.59)
qattjob	8.72 (2.26)	8.79 (2.33)	8.83 (1.77)
_Intimeres_1	0.11 (0.32)	0.18 (0.39)	0.12 (0.32)
_lqwhenwk_2	0.17 (0.38)	0.25 (0.43)	0.10 (0.31)
_lqwhenwk_3	0.24 (0.43)	0.21 (0.41)	0.14 (0.35)
_lqwhenwk_4	0.33 (0.47)	0.10 (0.30)	0.19 (0.39)
_lqwhenwk_5	0.11 (0.31)	0.09 (0.28)	0.07 (0.26)
_lqwhenwk_6	0.03 (0.18)	0.02 (0.14)	0.42 (0.49)
_lqhelp_2	0.17 (0.38)	0.22 (0.41)	0.11 (0.31)
_lqhelp_3	0.23 (0.42)	0.51 (0.50)	0.15 (0.35)
_lqhelp_4	0.42 (0.49)	0.12 (0.33)	0.64 (0.48)
			Continued

Table E.1 Continued

	Non-participants	Participants	Total
_Intimeresa2	0.09 (0.28)	0.13 (0.34)	0.09 (0.28)
_Intimeresa3	0.09 (0.29)	0.15 (0.36)	0.10 (0.29)
_Intimeresa4	0.11 (0.31)	0.13 (0.33)	0.11 (0.31)
_Intimeresa5	0.09 (0.29)	0.09 (0.28)	0.09 (0.29)
_Intimeresa6	0.09 (0.29)	0.07 (0.26)	0.09 (0.29)
_Intimeresa7	0.11 (0.32)	0.09 (0.28)	0.11 (0.31)
_Intimeresa8	0.10 (0.30)	0.06 (0.24)	0.10 (0.29)
_Intimeresa9	0.10 (0.30)	0.06 (0.23)	0.10 (0.30)
_Intimeresa10	0.11 (0.31)	0.05 (0.22)	0.10 (0.30)
_Iqacqual_1	0.16 (0.37)	0.19 (0.39)	0.10 (0.30)
_Iqacqual_2	0.21 (0.40)	0.30 (0.46)	0.13 (0.34)
_Iqacqual_3	0.05 (0.22)	0.07 (0.25)	0.03 (0.17)
_Iqacqual_4	0.04 (0.19)	0.08 (0.27)	0.02 (0.15)
_Iqacqual_5	0.08 (0.26)	0.05 (0.22)	0.04 (0.21)
_Iqexpstud_2	0.19 (0.39)	0.36 (0.48)	0.20 (0.40)
_Iqexpstud_3	0.21 (0.41)	0.15 (0.35)	0.20 (0.40)
_Iqexpstud_4	0.40 (0.49)	0.24 (0.43)	0.39 (0.49)
_Iqexpstud_5	0.08 (0.27)	0.07 (0.26)	0.08 (0.27)
_Inc91ethn_4	0.10 (0.30)	0.12 (0.32)	0.10 (0.30)
_Inc91ethn_5	0.10 (0.30)	0.11 (0.31)	0.10 (0.30)
			Continued

Table E.1 Continued

	Non-participants	Participants	Total
_Inc91ethn_6	0.10 (0.30)	0.10 (0.29)	0.10 (0.30)
_Inc91ethn_7	0.10 (0.30)	0.10 (0.30)	0.10 (0.30)
_Inc91ethn_8	0.10 (0.30)	0.09 (0.29)	0.10 (0.30)
_Inc91ethn_9	0.10 (0.30)	0.07 (0.26)	0.10 (0.30)
_Inc91ethn_10	0.10 (0.30)	0.07 (0.25)	0.10 (0.30)
_lqdrivlic_1	0.38 (0.48)	0.48 (0.50)	0.23 (0.42)
_lqlhealth_2	0.63 (0.48)	0.77 (0.42)	0.64 (0.48)
_lqlhealth_3	0.17 (0.37)	0.13 (0.34)	0.16 (0.37)
_lqhoursp_2	0.20 (0.40)	0.38 (0.49)	0.13 (0.33)
_lqhoursp_3	0.06 (0.23)	0.08 (0.28)	0.04 (0.18)
_lqhoursp_4	0.09 (0.28)	0.14 (0.34)	0.05 (0.23)
_lqhoursp_5	0.10 (0.30)	0.11 (0.32)	0.06 (0.24)
_lqhoursp_6	0.11 (0.31)	0.09 (0.28)	0.07 (0.26)
_lqhoursp_7	0.39 (0.49)	0.14 (0.35)	0.62 (0.49)
_lqyearsiv_2	0.16 (0.37)	0.22 (0.41)	0.10 (0.30)
_lqyearsiv_3	0.12 (0.33)	0.13 (0.34)	0.07 (0.26)
_lqyearsiv_4	0.12 (0.33)	0.09 (0.28)	0.07 (0.26)
_lqyearsiv_5	0.11 (0.31)	0.09 (0.28)	0.07 (0.26)
_lqyearsiv_6	0.13 (0.34)	0.11 (0.32)	0.08 (0.27)
_lqyearsiv_7	0.20 (0.40)	0.11 (0.31)	0.51 (0.50)
			Continued

Table E.1 Continued

	Non-participants	Participants	Total
_lqheduc_2	0.24 (0.43)	0.31 (0.46)	0.15 (0.36)
_lqheduc_3	0.09 (0.29)	0.13 (0.34)	0.06 (0.23)
_lqheduc_4	0.42 (0.49)	0.12 (0.33)	0.64 (0.48)
_lqchild3_1	0.34 (0.47)	0.34 (0.47)	0.20 (0.40)
_lqchild3_2	0.06 (0.23)	0.03 (0.18)	0.03 (0.18)
_lqccosc_1	0.03 (0.16)	0.03 (0.17)	0.02 (0.13)
_lcarer_pr_1	0.05 (0.22)	0.02 (0.14)	0.04 (0.21)
_lqminwage_2	0.57 (0.50)	0.54 (0.50)	0.57 (0.50)
_lqminwage_3	0.09 (0.29)	0.08 (0.27)	0.09 (0.29)
_lqlskill_2	0.37 (0.48)	0.39 (0.49)	0.37 (0.48)
_lqlskill_3	0.18 (0.39)	0.11 (0.31)	0.18 (0.38)
_lqsappl_2	0.31 (0.46)	0.40 (0.49)	0.19 (0.39)
_lqsappl_3	0.14 (0.35)	0.19 (0.39)	0.09 (0.28)
_lqsappl_4	0.37 (0.48)	0.12 (0.33)	0.61 (0.49)
_lqsappl_5	0.11 (0.31)	0.09 (0.28)	0.07 (0.26)
_lqtenure_1	0.10 (0.30)	0.11 (0.31)	0.06 (0.24)
_lqtenure_2	0.65 (0.48)	0.63 (0.48)	0.39 (0.49)
_lqtenure_3	0.16 (0.36)	0.19 (0.39)	0.10 (0.29)
_lqtenure_4	0.06 (0.24)	0.05 (0.21)	0.04 (0.19)
_lppt_hour_1	0.07 (0.25)	0.08 (0.27)	0.06 (0.23)
			Continued

Table E.1 Continued

	Non-participants	Participants	Total
_lqteleph_1	0.77 (0.42)	0.87 (0.34)	0.47 (0.50)
_lqscbalo_2	0.11 (0.31)	0.16 (0.37)	0.07 (0.25)
_lqscbalo_3	0.07 (0.26)	0.13 (0.33)	0.05 (0.22)
_lqscbalo_4	0.53 (0.50)	0.38 (0.49)	0.70 (0.46)
_lqservice_1	0.26 (0.44)	0.29 (0.45)	0.16 (0.36)
_lqservice_2	0.12 (0.33)	0.18 (0.38)	0.08 (0.27)
_lqservice_3	0.07 (0.25)	0.13 (0.34)	0.05 (0.21)
_lqservice_4	0.53 (0.50)	0.38 (0.49)	0.70 (0.46)
_lqsccab_2	0.23 (0.42)	0.31 (0.46)	0.15 (0.35)
_lqsccab_3	0.11 (0.32)	0.15 (0.36)	0.07 (0.26)
_lqsccab_4	0.53 (0.50)	0.38 (0.49)	0.70 (0.46)
_lqscadlin_2	0.26 (0.44)	0.34 (0.48)	0.16 (0.37)
_lqscadlin_3	0.14 (0.35)	0.18 (0.38)	0.09 (0.29)
_lqscadlin_4	0.53 (0.50)	0.38 (0.49)	0.70 (0.46)
_lqscndadv_2	0.27 (0.44)	0.29 (0.46)	0.16 (0.37)
_lqscndadv_3	0.14 (0.35)	0.15 (0.35)	0.08 (0.28)
_lqscndadv_4	0.53 (0.50)	0.38 (0.49)	0.70 (0.46)
_lqscadv_2	0.28 (0.45)	0.36 (0.48)	0.17 (0.38)
_lqscadv_3	0.15 (0.36)	0.17 (0.38)	0.09 (0.29)
_lqscadv_4	0.53 (0.50)	0.38 (0.49)	0.70 (0.46)

Continued

Table E.1 Continued

	Non-participants	Participants	Total
_lqljobs_2	0.44 (0.50)	0.48 (0.50)	0.45 (0.50)
_lqljobs_3	0.22 (0.42)	0.15 (0.36)	0.22 (0.41)
_Indaysben_2	0.33 (0.47)	0.34 (0.48)	0.33 (0.47)
_Indaysben_3	0.34 (0.47)	0.24 (0.43)	0.33 (0.47)
_lqlfinanc_2	0.34 (0.47)	0.34 (0.47)	0.35 (0.48)
_lqlfinanc_3	0.21 (0.41)	0.16 (0.37)	0.21 (0.41)
_lqlleave_2	0.35 (0.48)	0.48 (0.50)	0.36 (0.48)
_lqlleave_3	0.15 (0.36)	0.11 (0.32)	0.15 (0.36)
_lqhourwk_2	0.11 (0.32)	0.12 (0.32)	0.07 (0.25)
_lqhourwk_3	0.20 (0.40)	0.24 (0.42)	0.13 (0.33)
_lqhourwk_4	0.40 (0.49)	0.44 (0.50)	0.24 (0.43)
_lqhourwk_5	0.24 (0.43)	0.15 (0.36)	0.54 (0.50)
_lqsempag_2	0.34 (0.47)	0.48 (0.50)	0.21 (0.41)
_lqsempag_3	0.15 (0.36)	0.24 (0.43)	0.09 (0.29)
_lqsempag_4	0.37 (0.48)	0.12 (0.33)	0.61 (0.49)
_lqsempag_5	0.11 (0.31)	0.09 (0.28)	0.07 (0.26)
_lqchild11_1	0.46 (0.50)	0.54 (0.50)	0.28 (0.45)
_lqchild11_2	0.27 (0.45)	0.27 (0.44)	0.16 (0.37)
_lqchild11_3	0.12 (0.33)	0.07 (0.26)	0.07 (0.26)
_lqsfrrel_2	0.23 (0.42)	0.29 (0.45)	0.14 (0.35)
			Continued

Table E.1 Continued

	Non-participants	Participants	Total
_lqsfrrel_3	0.10 (0.30)	0.14 (0.35)	0.06 (0.24)
_lqsfrrel_4	0.37 (0.48)	0.12 (0.33)	0.61 (0.49)
_lqsfrrel_5	0.11 (0.31)	0.09 (0.28)	0.07 (0.26)
_ldnnondp_1	0.06 (0.23)	0.05 (0.22)	0.05 (0.23)
_ldnnondp_2	0.02 (0.13)	0.01 (0.09)	0.02 (0.13)
a2118	0.99 (0.11)	0.99 (0.08)	0.98 (0.13)
a2117	0.99 (0.11)	0.99 (0.12)	0.98 (0.13)
a2116	0.98 (0.14)	0.98 (0.14)	0.98 (0.15)
a2115	0.98 (0.15)	0.97 (0.17)	0.98 (0.15)
a2114	0.97 (0.18)	0.96 (0.20)	0.96 (0.18)
a2113	0.95 (0.21)	0.94 (0.23)	0.95 (0.21)

Appendix F

Logit coefficients using the administrative data (odds-ratio)

Table F.1 Logit coefficients using the administrative data (odds-ratio)

	(1)	(2)	(3)
sextype	0.77 (0.090)*	0.73 (0.086)**	0.72 (0.085)**
ageband==2	1.14 (0.15)	1.10 (0.14)	1.24 (0.16)
ageband==3	1.06 (0.13)	1.09 (0.15)	1.21 (0.16)
ageband==4	1.11 (0.14)	1.25 (0.17)	1.38 (0.192)*
ageband==5	1.10 (0.14)	1.27 (0.182)+	1.41 (0.203)*
ageband==6	1.11 (0.15)	1.28 (0.20)	1.43 (0.221)*
ageband==7	0.68 (0.122)*	0.81 (0.16)	0.91 (0.18)
ageband==8	0.72 (0.17)	0.93 (0.24)	1.07 (0.27)
ageband==9	0.60 (0.26)	0.79 (0.34)	0.91 (0.40)
Continued			

Table F.1 Continued

	(1)	(2)	(3)
gor==2	1.01 (0.12)	1.01 (0.12)	1.01 (0.12)
gor==3	0.95 (0.14)	1.00 (0.15)	1.02 (0.15)
gor==4	0.98 (0.14)	0.95 (0.14)	0.96 (0.14)
gor==5	1.07 (0.14)	1.03 (0.14)	1.04 (0.14)
gor==6	0.78 (0.109)+	0.78 (0.109)+	0.78 (0.109)+
gor==7	1.21 (0.16)	1.17 (0.16)	1.17 (0.16)
gor==8	0.90 (0.15)	0.87 (0.14)	0.86 (0.14)
gor==9	0.68 (0.084)**	0.66 (0.082)**	0.67 (0.083)**
gor==10	0.99 (0.13)	0.95 (0.12)	0.94 (0.12)
gor==11	1.05 (0.14)	1.08 (0.15)	1.08 (0.15)
gor==12	0.86 (0.11)	0.84 (0.11)	0.85 (0.11)
pcageyc==2		1.25 (0.122)*	1.32 (0.130)**
pcageyc==3		1.40 (0.106)**	1.48 (0.114)**
pcageyc==4		1.16 (0.11)	1.25 (0.122)*
pcageyc==5		1.15 (0.11)	1.24 (0.116)*
pcageyc==6		0.98 (0.11)	1.10 (0.13)
pcnumbch==3		0.84 (0.047)**	0.84 (0.047)**
pcnumbch==4		0.59 (0.050)**	0.59 (0.050)**
pcnumbch==5		0.39 (0.052)**	0.39 (0.053)**
disabled		0.53 (0.082)**	0.58 (0.090)**
			Continued

Table F.1 Continued

	(1)	(2)	(3)
disalen		0.85 (0.026)**	0.85 (0.026)**
prehist3==1			1.02 (0.25)
prehist3==10			1.20 (1.25)
prehist3==11			1.24 (0.30)
prehist3==101			1.10 (0.83)
prehist3==110			1.61 (1.67)
prehist3==111			1.29 (0.32)
prehist3==1001			1.92 (1.23)
prehist3==1011			1.04 (1.09)
prehist3==1100			1.65 (1.71)
prehist3==1101			1.61 (1.23)
prehist3==1111			0.94 (0.24)
prehist3==10001			0.56 (0.58)
prehist3==10011			2.31 (1.49)
prehist3==10111			0.64 (0.66)
prehist3==11001			1.43 (1.09)
prehist3==11011			0.71 (0.74)
prehist3==11111			0.97 (0.25)
prehist3==100000			0.95 (0.41)
prehist3==100001			1.00 (0.38)
			Continued

Table F.1 Continued

	(1)	(2)	(3)
prehist3==100011			0.75 (0.33)
prehist3==100101			4.66 (5.16)
prehist3==100111			1.07 (0.44)
prehist3==101111			1.88 (0.689)+
prehist3==110000			0.41 (0.30)
prehist3==110001			0.59 (0.28)
prehist3==110011			0.50 (0.22)
prehist3==110111			1.21 (0.50)
prehist3==111000			0.96 (0.50)
prehist3==111001			0.54 (0.26)
prehist3==111011			1.24 (0.46)
prehist3==111100			1.44 (0.52)
prehist3==111101			1.01 (0.40)
prehist3==111110			0.84 (0.44)
prehist3==111111			0.94 (0.22)
prehist3==222222			1.01 (0.74)
(sum) a2118			5.05 (1.911)**
(sum) a2117			0.41 (0.158)*
(sum) a2116			1.38 (0.51)
(sum) a2115			0.94 (0.29)
			Continued

Table F.1 Continued

	(1)	(2)	(3)
(sum) a2114			1.08 (0.31)
(sum) a2113			0.92 (0.26)
proprio			0.70 (0.057)**
Observations	65,887	65,887	65,731
Age	0.02	0.03	0.02
Region	0.00	0.00	0.00
YCAge		0.00	0.00
Numbch		0.00	0.00
C&S			0.43
Recent			0.00

Standard errors in parentheses.

+ significant at 10%; * significant at 5%; ** significant at 1

Appendix G

Model summary statistics by stata

Table G.1 Model summary statistics by stata

	s1	s2	s3	s4	s5	s6
chi2	106.154***	40.608	59.637**	58.849***	30.411	38.972**
r2_p	0.058	0.046	0.035	0.025	0.028	0.036
N	3,816	1,672	3,408	5,191	2,793	4,022
	s7	s8	s9	s10	s11	s12
chi2	58.522	54.095	45.625	50.375**	43.047**	101.876***
r2_p	0.073	0.05	0.07	0.06	0.058	0.051
N	1,334	1,888	1,214	1,724	1,359	4,433
	s13	s14	s15	s16	s17	s18
chi2	103.288***	91.228***	74.022***	33.606	31.467	74.724***
r2_p	0.064	0.041	0.051	0.024	0.034	0.02
N	2,704	3,683	2,701	2,883	2,133	9,301
	s19	s20	s21	s22	s23	s24
chi2	77.169**	33.088	55.719**	20.025	34.137*	127.633***
r2_p	0.096	0.094	0.147	0.058	0.104	0.059
N	1,411	531	845	721	833	5,797

Standard errors in parentheses.

+ significant at 10%; * significant at 5%; ** significant at 1

Appendix H

Technical annex

Table H.1 Trends in average numbers of new entrants to NDLP per month averaged by year: 1998 to 2003

Year	Mean	Std. dev.	The average number of months
1998	3,363	2,302	3
1999	8,380	2,350	12
2000	8,095	1,926	12
2001	8,561	2,014	12
2002	10,716	2,373	12
2003	11,551	2,509	12
Total	9,170	2,873	63

Source: Author's calculations from NDED.

Table H.2 Different sample definitions

Sample Number	Sample description (groups)	Number of NDLP participants	Number of non-participants	Total sample size
I	A, B, C, D and E	4,245	65,606	69,851
II	B and C	1,787	39,277	41,064
III	B, C1 and C2	1,777	1,819	3,596
IV	B, C1 and C2.with $ps < .3$	1,656	1,653	3,309
V	B1 and C1	1,259	1,241	2,500
VI	B1 and C1 with $ps < .3$	1,171	1,125	2,296
VII	B, C1 and C2 (except those non-participants who join NDLP after April 2001)	1,777	1,819	3,500
VIII	B, C1 and C2 (except those non-participants who join NDLP after April 2001 and their re respective matches)	1,777	1,819	3,500
IX	Sample VII with $ps < .3$	1,656	1,653	3,309
X	B1 and C1 (except those non-participants who join NDLP after April 2001 and their re respective matches)	1,259	1,241	2,500
XI	Sample X with $ps < .3$	1,171	1,125	2,296

Note: 'ps' denotes the estimated propensity score.

The structure of the data collection implies that $N^P \subseteq N^A$ and that

$$N^F \subseteq N^M \subseteq N^Q \subseteq N^A.$$

That is, the administrative data contain all the sub-samples, the postal survey respondents are a subset of the lone parents in the administrative data, the matched participants and non-participants for whom face-to-face interviews are attempted are a subset of the postal survey respondents, and the face-to-face interview completers are a subset of the matched samples.

Two other features of the data add additional complexity. Firstly, the data include additional variables merged into the administrative data from other sources based on the geographical location of each lone parent. These data which are denoted by Z , consist of various economic and social characteristics of the local area. These variables can serve as additional conditioning variables for the matching, or could potentially function as exclusion restrictions (instruments) in analyses that attempt to account for selection on unobservables as well as selection on observables.

Secondly, the two sets of outcome variables are measured in somewhat different ways. As already noted, the administrative data include a panel of outcomes, with information on IS receipt on each member of N^A in each period starting at $t = 1$ in Figure 2.1.1. In addition, self-reported outcomes are collected at one point in time from both the postal questionnaire and as part of the face-to-face interview. These self-reported outcomes are denoted by g^Q and g^F , respectively.

Table H.3 summarizes the data collection. It indicates the types of variables associated with each data source and which variables are available for each sub-sample within the data. that $i \in N^A$, $j \in N^Q$ and $k \in N^F$ are used.

Table H.3 Summary of data collection process

Sample	Time period collected	Sample denoted	Individual characteristics						
			X^A	X^Q	X^F	Z^A	Y^A	S^F	S^Q
Administrative data	1	N^A	✓			✓	✓		
Postal questionnaire data respondents (Groups A, B and C)	q	N^Q	✓	✓		✓	✓		✓
Participant in NDLP	p	N^P	✓			✓	✓		
Participant and non-participant matched sample (Groups B1 and C1)	p	N^M	✓	✓		✓	✓		✓
Face to Face Interview.	f	N^F	✓	✓	✓	✓	✓	✓	✓

Figure H.1 The number of NDLP participants – October 1998 to December 2003

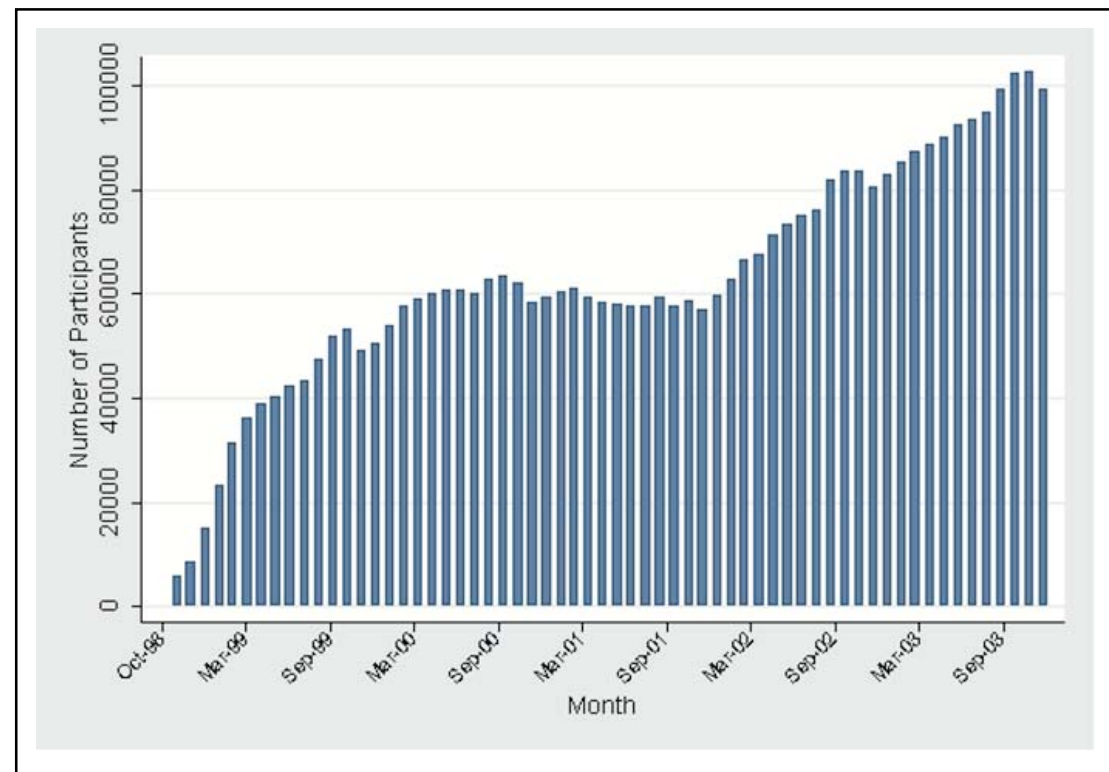
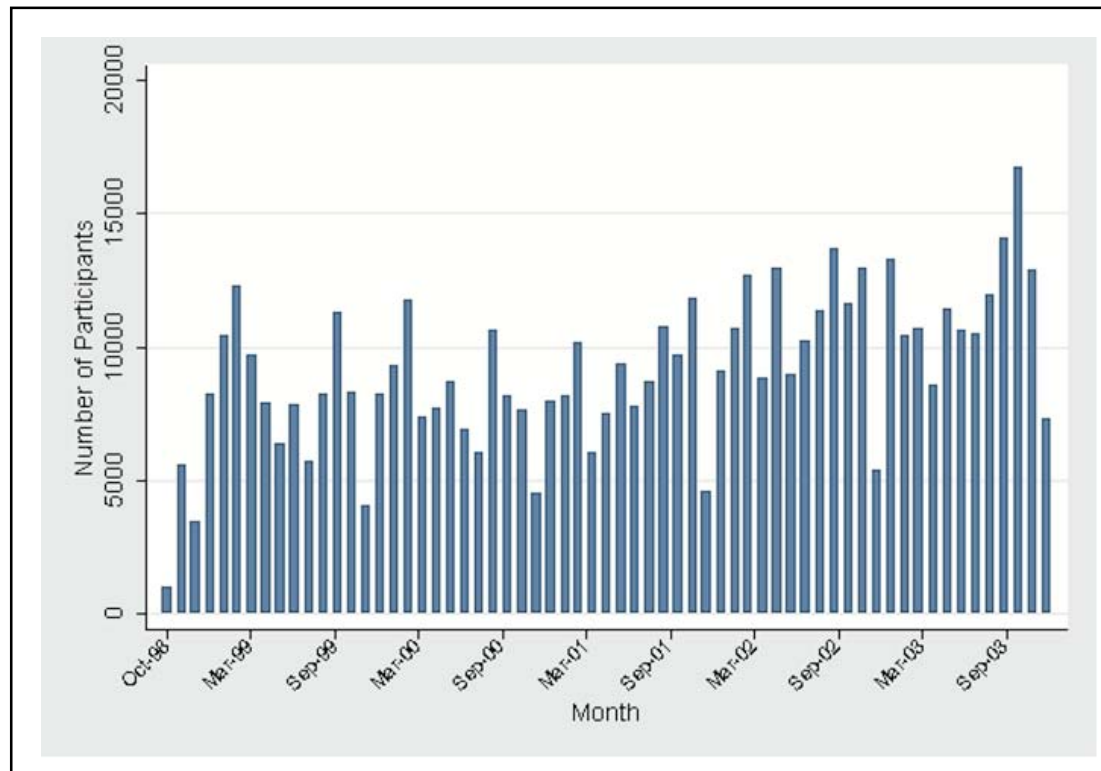


Figure H.2 The number of NDLP new entrants – October 1998 to December 2003



Appendix I

Basic notation, the matching methodology and the NatCen model

I.1 The basic notation

In the notation the following conventions are adopted. Superscripts indicate either a particular data source (e.g., the postal survey) or the point at which the matching takes place. Subscripts indicate the set of people for whom the data exist or the time period over which the data are available. In the case of coefficients in equations, the subscripts indicate the equation to which the coefficient belongs.

As discussed in Chapter 2 and illustrated in Figure 2.1, NatCen collected the data for the New Deal for Lone Parents (NDLP) evaluation at several points in time. The data collection process starts at time period $t = 1$ with an administrative data set, itself an extract from a larger administrative database. The administrative data for the NDLP evaluation contain information, which is denoted X_i^A , on the socio-economic characteristics of each lone parent eligible to enter the NDLP. The administrative data also contain information on outcome variables of interest in the evaluation, such as participation in IS. Let Y_{it} denote the vector of outcome variables for individual i at time t . Let N^A denote set of lone parents with complete records in the administrative data.

At some later date $t = q$, all of the lone parents in the administrative data receive a postal questionnaire, which was mailed to a sample of the eligible population and this sample was geographically clustered, with primary sampling unit (PSU) selection probabilities dependent on historical rates of NDLP participation. Some, but not all, of the individuals who receive the postal survey complete it and return it to the NatCen. Define a binary variable $Q \in \{0,1\}$ indicating response to the postal survey,

let X^Q denote the vector of characteristics collected from the postal survey, and let N^Q denote the set of respondents to the postal survey. During the interval shown in Figure 2.1, any eligible lone parent, regardless of whether he or she responds to either survey, can participate in NDLP. Denote the time of participation by $t = p$, the set of individuals who choose to participate by N^P and let $D \in \{0,1\}$ indicate participation in the NDLP.

Following the return of the postal surveys, a statistical matching exercise matches up each respondent who completed a postal survey to a single non-respondent with similar observable characteristics, drawn from the (much larger) population of non-participants who completed the postal survey. Denote the set of those who get matched by N^M . Later, at time $t = f$, NatCen attempts to complete a face-to-face interview with each member of N^M . Let N^F denote the subset of N^M completing a face-to-face interview, X^F denote the vector of variables (other than outcomes) collected in the face-to-face interview, and define the binary variable $F \in \{0,1\}$ indicating completion of the face-to-face interview. Finally, let $X_i = \{X_i^A, X_i^Q, X_i^F\}$ denote the complete set of individual characteristics from all three data sources and subscripts i, j and k such that $i \in N^A, j \in N^Q$ and $k \in N^F$ are used.

1.2 The complete model

In this section, a complete and very general version of the model of the NDLP participation process, the labour market outcome processes and the survey response processes are written. Writing down this general model allows the explicit assumptions that underlie the NatCen analysis to be seen. The most general possible model is not written down; in particular, linear functional forms for the survey response and NDLP participation functions is assumed. In some cases, there are several variations, each of which presumes a different conditioning set. The choice of conditioning sets depends on measurement issues and on choices regarding how best to deal with survey non-response.

The complete form of this model can be written in the following notation:

Postal survey questionnaire return:

$$(1) \quad Q_i^* = \beta_1^Q X_i^A + u_{i1}$$

where $Q_i = 1$ if $Q_i^* \geq 0$ and $Q_i = 0$ otherwise.

Participation in the NDLP:

$$(2a) \quad D_{it}^* = \beta_2^A X_i^A + \gamma_2^A Z_i^A + \delta_{\tau 2} L^\tau Y_{it}^A + u_{j2}$$

$$(2b) \quad D_{jt}^* = \beta_2^A X_j^A + \beta_2^Q X_j^Q + \gamma_2^A Z_j^A + \delta_{\tau 2} L^\tau Y_{jt}^A + \alpha S_j^Q + u_{j2}$$

where $D_{it} = 1$ if $D_{it}^* > 0$ and $D_{it} = 0$ otherwise, and similarly for D_{jt} and where L^τ is the lag operator for τ periods of time, with $\tau > 0$.

Matching:

Hence, the matching function M can be written as

$$M_j^Q = \Phi(X_j^A, X_j^Q, Z_j^A, L^T Y_{j\tau}^A, S_j^Q, R, \Omega),$$

where $R = 1$ if matching with replacement is used and $R = 0$ otherwise, and where Ω denotes the choice of matching algorithm (e.g., single nearest neighbour matching, kernel matching with a Gaussian kernel and so on).

Response to the face-to-face interview survey:

$$(3) \quad F_j^* = \beta_3^A X_k^A + \beta_3^Q X^Q + \gamma_3^A Z^A + u_{k3}$$

where $F_j = 1$ if $F_j^* > 0$ and $F_j = 0$ otherwise.

Outcomes from the administrative data:

$$(4a) \quad Y_{it}^A = \pi(X_i^A, Z_i^A, M_i^Q, D_{it}, u_{i4}),$$

$$(4b) \quad Y_{jt}^A = \pi(X_j^A, X_j^Q, Z_j^A, M_j^Q, D_{jt}, u_{j4}),$$

$$(4c) \quad Y_{kt}^A = \pi(X_k^A, X_k^Q, X_k^F, Z_k^A, M_k^Q, D_{kt}, u_{k4}).$$

Rematching:

$$M_k^F = \Phi(X^A, X^Q, X^F, Z^A, R, \Omega).$$

Self-reported outcomes from the postal and face-to-face interview surveys:

$$S_j^Q = \xi(X_j^A, X_j^Q, Z_j^A, D_j, u_{k5}),$$

$$S_k^F = \xi(X_k^A, X_k^Q, X_k^F, Z_k^A, M_k^F, D_k, u_{k6}).$$

In all cases, α , β , δ and γ are parameters to be estimated and all of the u are mean-zero stochastic unobservables.

1.3 Non-response

Non-response to both the postal survey and face-to-face interview survey represents a key issue in the NatCen evaluation, and in any re-evaluation utilising the same data. The NatCen evaluation makes the assumption that non-response depends only on observable variables. In terms of the notation in this section, they assume

independence between the unobserved determinants of postal survey response and outcomes, $u_{i1} \perp u_{i4}$, and independence between the unobserved determinants of face-to-face interview response and outcomes, conditional on response to the postal survey, $u_{k3} \perp u_{k4} \mid Q = 1$

If the assumption that non-response depends only on observables is inappropriate, then the resulting bias depends on the nature of the relationship between the unobserved determinants of postal survey response and outcomes. In formal terms, suppose that we assume a simple linear form for the conditional mean function for outcomes, so that $Y_{ik} = X_i \beta + \varepsilon_i$, where ε_i satisfies the usual assumptions. Then if an estimate of the parameters of the mean outcome function are sought conditional on response to the postal and face-to-face interview survey:

$$E(Y_{kt} \mid Q_i = 1, F_i = 1, X_i, Z_i) = (X_i \beta + E(\varepsilon_i \mid u_{i1} \geq \beta_1^Q X^A, u_{k3} \geq \theta \Xi)) = X_i \beta + h(\beta_1^Q X^A, \theta \Xi)$$

where $\theta = (\beta_3^A, \beta_3^Q, \beta_3^A)$ and $\Xi = (X^A, X^Q, Z^A)$. The exact form of the bias in the conditional expectation of Y_{kt} depends on the function h , which in turn depends on the assumed joint distribution of ε_i, u_{i1} and u_{k3} and the constituent elements of X^A and Ξ . Dealing with the bias when the assumption that non-response depends only on observables fails, requires an exclusion restriction (or instrument) – a variable that affects non-response but does not affect labour market outcomes other than through its effect on non-response.

1.4 The matching model

In this section, the reasoning behind matching estimators of program impact is briefly outlined. Matching estimators assume selection on observables. That is, they assume that, conditional on some set of observable characteristics, treatment status (program participation) is independent of outcomes in the untreated state (or in both the treated and untreated states, depending on the parameter of interest). Unlike standard regression adjustment mechanisms, matching allows the evaluator to avoid making a functional form assumption about the outcome equation. This means that non-linearities in the conditional mean function that a regression equation might miss (given the standard practice of including each variable linearly with few or no higher order or interaction terms), matching will pick up. On this point see the discussion in Black and Smith (2004).

Although it is impossible to prove definitively that the ‘natural’ rate of outcomes observed for the matched comparison sample is identical to the ‘natural’ rate of outcomes that would have been observed for the participant sample if these rates could have been observed, the research design adopted for NDLP represents a very serious attempt to do so. The analysis used to check this is described in the technical report (Phillips et al, 2003). A way in which this could be achieved would be to find, for each participant, a non-participant who was identical in all relevant respects. These ‘relevant respects’ would be those that affect employment chances – not just an identical age and gender, but an identical work history, attitudes to work and so

on. There are three fundamental difficulties with this: One is that it is not possible to know all the factors that affect employment. The second is that some of these factors, for example an individual's attitude towards balancing childcare and work, are inherently difficult to define and measure. The third is that the list of possible variables is too large to allow a match to be achieved on each one separately. The third problem is assessed by using an alternative method of matching using a score that represents the probability of participation (the propensity score). This is well described in the literature (see, for example, Rosenbaum and Rubin, 1983 and Bryson, Dorsett and Purdon, 2002).

Cell matching represents the simplest matching approach; it compares groups of participants and non-participants with exactly the same observable characteristics. When conditioning on several observable variables, cell matching quickly runs into the so-called 'curse of dimensionality', because there will be combinations of values for the variables for which the sample includes participants but no non-participants.

Rosenbaum and Rubin (1983) suggest overcoming the curse of dimensionality by matching participants and non-participants based on their so-called 'propensity score' – the estimated probability of participation given the observable matching variables. This reduces the matching from a multi-dimensional problem (where the number of dimensions depends on the number of matching variables) to a one-dimensional problem. Matching methods have recently generated a lot of interest in the applied econometrics literature, particularly in the context of evaluating employment and training programs. Key papers include Dehejia and Wahba (1999, 2001), Heckman, Ichimura and Todd (1997, 1998), Heckman, Ichimura, Smith and Todd (1998) and Smith and Todd (2005). NatCen employs propensity score matching in their evaluation of the NDLP as a method for choosing the non-participants for the face-to-face interviews.

The basic steps in propensity score matching are as follows:

The probability or propensity of participating in a programme is estimated for a sample of participants and a pool of non-participants. This is typically done using a statistical model, such as a logit model, where the predictors of participation are all the variables thought to influence both participation and the subsequent rate of outcomes.

Each participant is then matched to a non-participant, the most simple method involving the selection from the pool of the non-participant with the closest propensity score (nearest-neighbour matching).

The success of propensity score matching largely depends on whether the propensity score model includes all the relevant factors which affect propensity to participate in the programme, whether good data have been collected for all of these factors, and whether the structure of the model is appropriate. If this is the case then the estimated propensity score will be accurate and the matched comparison sample design will give an unbiased estimate of programme additionality. The less accurate the score, the more the additionality estimate is likely to be biased. It must be remembered that some factors (such as personal motivation just before participation) are hard to observe and measure.

The potential outcome model due to Roy (1951) and Rubin (1974) provides a useful notation. Let Y_{i1} represent an outcome when individual i takes part in the NDLP and let Y_{i0} represent the outcome when individual i does not participate in NDLP, where $Y_i \in \{Y_i^A, S_i^Q, S_i^F\}$ refers to any single outcome from the administrative data or from one of the two surveys. Of course, only at most one of the two potential outcomes Y_{i1} and Y_{i0} is observed for each person. Using the potential outcome notation, the realised outcome for individual i as:

$$Y_i^A = D_i Y_{i1} + (1 - D_i) Y_{i0}.$$

In this framework, the evaluation problem constitutes a missing data problem. To see this, note that the impact of participation for individual i equals $Y_{i1} - Y_{i0}$. If both potential outcomes are observed, estimating the impact becomes a simple matter of subtraction.

The main parameter of interest in this evaluation is the average effect of the NDLP treatment on the treated (ATT). This parameter can be written as:

$$ATT = E(Y_{i1} - Y_{i0} | D_i = 1) = E(Y_{i1} | D_i = 1) - E(Y_{i0} | D_i = 1),$$

where both expectations are taken over all participants. In words, the ATT gives the mean effect of the NDLP program on those we observe to participate in it.

The first term can be written $E(Y_{i1} | D_i = 1)$, using data on participants. The trick comes in estimating the second term, $E(Y_{i0} | D_i = 1)$, because by definition there is no observation of the non-participation outcome, Y_{i0} for participants, $D_i = 1$. A randomised experiment produces the second term by construction, by randomly forcing individuals who otherwise would have participated not to participate. The assumptions required for propensity score matching estimators to consistently estimate this term are now considered.

The crucial assumption underlying the propensity score method is the conditional independence assumption.²⁰

²⁰ In the statistics literature, this assumption is sometimes called the 'unconfoundedness' or 'ignorable treatment assignment' assumption. We follow the usage in the econometric literature and refer to it as the conditional independence assumption.

Definition 1 Conditional independence (Rosenbaum and Rubin, 1983)

Assignment to treatment (NDLP participation in our context) is conditionally independent of outcomes in the untreated state, Y_{i0}^A , given pretreatment variables W_i , if:

$$D_i \perp Y_{i0}^A | W_i,$$

for some set of observed variables $W_i \subset (X_i, Z_i, L^\tau Y_{i\tau}^A)$, where L is the lag operator and $\tau > 0$, so that $L^\tau Y_{i\tau}^A$ denotes pre-treatment values of the outcome variables in the administrative data.

Conditional independence becomes an assumption once we assume that some in the data make the conditional independence condition hold. The Conditional Independence Assumption (CIA) is a strong and restrictive assumption. In our context, it requires that all of the variables that affect participation in the NDLP as well as labour market outcomes in the absence of participation are observed. Nothing guarantees that any set of variables W in our data cause the CIA to hold. In the absence of experimental data, a case is made based on economic theory and prior knowledge about the determinants of program participation and labour market outcomes that all the important factors are observed.

The power and attraction of the CIA assumption is that it validates comparisons between participants and non-participants with the same (or, in practice, closely similar) values of the covariates, W . As noted earlier, matching on all of the elements of W raises a curse of dimensionality, leading Rosenbaum and Rubin (1983) to suggest matching on the propensity score. The propensity score is now defined.

Definition 2 Propensity score

The propensity score is the conditional probability of treatment:

$$P(W_i) \equiv \Pr(D_i = 1 | W_i).$$

The propensity score has two important properties:

Lemma 1 The balancing property.

The balancing property holds that

$$D_i \perp W_i | P(W_i).$$

In other words, the conditioning variables are independent of participation conditional on the propensity score. The balancing property holds whether or not W_i cause the conditional independence assumption to hold. Thus, testing this condition does not test the CIA. However, testing this condition – the so-called balancing test – does allow a test of the propensity score specification. An empirical failure of the balancing property indicates an insufficiently flexible propensity score specification. The balancing test is discussed in more detail in Chapter 0.

Combined with the CIA, the balancing property suggests the key property of the propensity score:

Lemma 2 Conditional Independence Given the Propensity Score

If assignment to treatment is conditionally independent given W , then assignment to treatment is conditionally independent given the propensity score estimated as a function of W . In notation,

$$D_i \perp Y_{i0} | P(W_i).$$

Put differently, if the CIA holds for W , then it also holds for $P(W_i)$. Rosenbaum and Rubin (1983) show that you can identify the ATT by conditioning on the propensity score. Hence the ATT can be written as

$$T = E(Y_{i1} - Y_{i0} | D_i = 1) = E[E[Y_{i1} | D_i = 1, P(W_i)] - E[Y_{i0} | D_i = 0, P(W_i)]],$$

where the outside expectation is over the distribution of among the participants. This implies that matching on the estimated propensity score can be used to construct the expected counterfactual outcome that participants would have experienced, had they not participated.

The discussion of matching here relates to the discussion of non-response in the preceding section. The NatCen evaluation matches conditional on response to the postal survey. The resulting estimates can be considered in one of several ways. If a common effect world, in which the NDLP has the same impact on all lone parents, is assumed, then assuming that the CIA holds conditional on response to the postal survey suffices to identify the ATT for all participants, not just those responding to the postal survey.

In contrast, if a heterogeneous effect world is assumed, where the impact of the NDLP varies among individuals (who then act on that variation), then there are (at least) two alternatives. Firstly, an assumption can be made that the CIA holds conditional on response to the postal survey, and produces consistent estimates of the ATT for participants who respond to the postal survey. Secondly, it can be assumed that the CIA also takes care of differential non-response to the postal survey. That is, it can be assumed that non-response to the postal survey is independent of Y_{0i} conditional on W_i . In that case, with appropriate weighting to take account of the non-response, matching will consistently estimate the ATT for all participants. The issues of weighting and non-response is returned to in Chapter 7.

1.5 The NatCen model

NatCen estimated a restricted version of the general model laid out in Section 1.3. Using the same notation as in Section 1.3, the model actually estimated in the NatCen report can be written as:

Participation in the NDLP:

$$D_j^* = \beta_2^A X_j^A + \beta_2^Q X_j^Q + u_{j2}$$

where $D_j = 1$ if $D_j^* > 0$ and $D_j = 0$ otherwise.

Matching:

$$M_j^Q = \Phi(X_j^A, X_j^Q, D_j, R = 0, \Omega),$$

where Ω denotes single nearest neighbour matching without replacement based on propensity scores that include variables from the postal survey and the administrative data. The NatCen model rejects possible matches where the matched non-participant is not on Income Support (IS) in the month that the participant being matched starts to participate in NDLP. Put differently, they force an exact match on IS receipt in the month of the implicit decision to participate in the NDLP or not. In addition, nearest neighbour or not, NatCen included all non-participant observations with estimated propensity scores above 0.3 in the matched comparison group.

Outcomes in the administrative data:

$$Y_{kt}^A = \xi(X_j^Q, D_j, M_j^Q).$$

In addition, as noted in Section I.4, the NatCen report assumes that response to the postal survey is unrelated to outcomes conditional on X^A and that response to the face-to-face interview survey is unrelated to outcomes conditional on X^A and X^Q . The NatCen report does not do rematching, other than as a sensitivity analysis, nor do they report estimate impacts on self-reported outcomes from the postal and face-to-face interview surveys.

Getting from the general model in Section I.3 to this model involves a variety of simplifying assumptions. Some are required for identification of the ATT while others serve only to make the model more econometrically tractable. This notation makes it easy to write down these assumptions explicitly and consider the importance and effects of each one.

Key assumptions underlying the NatCen model:

A1: *The Conditional Independence Assumption (CIA)*

The CIA invoked here implies that

$$D_i \perp Y_{i0}^A | W_i,$$

where NatCen assumes that $W_i = (X^Q, L^T Y_{j\tau}^A)$. In words, outcomes in the absence of NDLP participation are independent of the NDLP participation choice conditional on variables from the postal survey and a (very limited) set of past labour market outcomes drawn from the administrative data. This implies that conditioning on the exogenous characteristics Z , or on the various attitudinal characteristics collected in

X^F , is not required to induce conditional independence or, put differently, that these omitted factors do not affect both participation in NDLP and outcomes in the absence of NDLP.

A2. ***Covariate balance***

The NatCen assume that their propensity score specification, which includes no higher order or interaction terms, suffices to balance the distributions of covariates between the participants and matched non-participants. That is, they assume it satisfies the balancing property in Lemma 1 of Section I.5.

A3: ***Ignorable non-response conditional on observables***

Non-response to both the postal questionnaire and the face-to-face questionnaire is independent of outcomes conditional on observable variables from X^A in the former case and from (X^A, X^Q) in the latter case. As discussed in Section I.4, this means that the stochastic error terms (u_{j1}, u_{k3}) in survey response equations are not correlated with the unobserved component of Y_{it} , so that a simple weighting procedure corrects for the bias associated with non-response to both surveys. This assumption implies that the NatCen procedure of using all participants and non-participants responding to the face-to-face interview survey, rather than just those matched pairs wherein both the participant and non-participant respond, will yield consistent estimates, given appropriate weighting on the observable variables affecting non-response.

A4. ***Validity of over-weighting non-participants with high estimated propensity scores***

The NatCen assumes that including extra non-participants with estimated propensity scores over 0.3 does not render the impact estimates inconsistent. If, for example, Y_0 varies systematically with $P(W)$, this assumption will fail to hold in the data.

A5: ***No effect of participation window or outcome measurement window***

As discussed in Section 2.2, individuals become Group B participants, and thereby get included in the sample, only if they participate in NDLP after returning their postal survey. If the unobserved determinants of the timing of postal survey response are correlated with labour market outcomes, then this restriction will affect the external validity of the NatCen impact estimates. Similarly, the window over which individual outcomes get measured depends on the timing of response to the postal survey and to the face-to-face interview. If the unobserved determinants of these timing decisions also affect labour market outcomes, then again the external validity of the estimates will suffer, although this latter effect depends in part on the specific outcome measure in question. Measures of outcomes that occur prior to the end of the shortest window will have no external validity problems from this source.

I.6 Testing the NatCen assumptions

All of the assumptions underlying the NatCen analysis listed in Section I.6 are amenable to at least some testing using the available data. For example, their conditional independence assumption (Assumption A1) can be tested by computing matching estimates based on propensity scores that include the exogenous Z variables, the attitudinal variables from the face-to-face interview survey, or a richer set of past labour market outcome variables drawn from X^A . If substantially different estimates are obtained, this suggests (but does not prove) that their version of the CIA did not hold in the data.

A2, regarding covariate balance, can be tested by examining the balancing issue more thoroughly than does NatCen in their report. More flexible alternative propensity score specifications that utilise the same underlying set of variables can be examined to see if they do a better job of balancing the covariates and, more importantly, to see if they affect the obtained impact estimates.

A3, regarding non-response, can be tested using standard methods for detecting unobservables affecting non-response. These include trying to find an instrument for non-response in the data or by comparing the labour market outcomes of those respondents most like the non-respondents – e.g., those who took the longest to respond or had to be contacted the most times – to those least like the non-respondents. If the labour market outcomes differ between the two groups conditional on the observables used to correct for the non-response, this suggests (but does not prove) that A3 fails to hold in the data. An indirect test consists of constructing estimates using only the participant and non-participant pairs in which both respond to the face-to-face interview survey. Under A3, the resulting estimates should not differ statistically from those reported by NatCen.

Another test of A3 consists of constructing impact estimates using only variables from the administrative data. Such matching estimates can be constructed for both the full sample and for the sample of respondents. If the estimates differ between the two, this suggests (but does not prove) that A3 fails to hold in the data. A similar exercise can be performed using the full postal survey sample and the sub-sample of respondents to the face-to-face interview survey. That latter test does not pick up non-response problems related to the postal survey, but has the advantage that it can rely on the same propensity score specification utilised in the NatCen report.

A4 can be tested by re-matching the participants and non-participants who respond to the face-to-face interview. This procedure eliminates the over-weighting. If it yields estimates that differ statistically from those reported by the NatCen, then this suggests that A4 fails to hold in the data.

A5 can be tested by examining how the impact estimates vary with different definitions of the participation window and outcome windows and, in the case of the outcome window, by considering alternative outcome measures that are not affected by variations in the window among sample members.

I.7 Summary

The main objective of this chapter is to clarify the econometric model implicit in the NatCen analysis in order to make clear the precise assumptions required for their model to generate consistent estimates.

As shown in this chapter, these assumptions include:

- the relatively limited set of regressors included in the NDLP participation probability model suffices to control for non-random selection into the program. In particular, they assume eliminating selection bias does not require matching on local labour markets or matching on detailed IS and employment histories;
- non-response to the postal survey and non-response to the face-to-face interview depend only on the small set of observables used to construct the sampling weights. In particular, they must assume that non-response is unrelated to outcomes conditional on this small set of observables;
- participation in NDLP is what causes a lone parent to find employment rather than being offered a part-time job and possibly having a Work Focused Interview (WFI) and registering for NDLP to simply get the 'better off calculation' in order to explore whether the job is worth taking.

As a final remark it is important to note that the 'full' characterisation of the underlying econometric model which would capture the NatCen data is in fact quite complicated but the version of the impact assessment model used by NatCen is relatively simple and relies on a number of strong and potentially controversial assumptions.

Appendix J

Evaluating the NatCen report

J.1 Introduction

To rigorously appraise the statistical analysis in the NatCen report it is first necessary to replicate the propensity score estimation performed by NatCen. This replication is described in Section J.1. Section J.1 also assesses whether there is common support in the distribution of propensity score estimates in the participants and non-participants to the New Deal for Lone Parents (NDLP). Section J.2 further examines issues in the matching and covariate balancing in the face to face interviewees and explores the issue of matching with and without replacement. Section J.3 assesses how much these issues of exactly what sample is used in the matching and exactly what matching technique is employed matter using a variety of different possible outcomes measures.

J.2 Propensity scores and common support

J.2.1 Propensity score replication

The replication of NatCen's propensity scores began by doing exactly what they did. NatCen used 37 variables to estimate its logit propensity score model. A complete list of their regressors, along with definitions and acronyms, appears in Appendix A. Their SPSS code, the statistical package they used (SPSS), their data set and their data selection and filter commands, which served to reduce the sample to Group B participants who responded to the postal survey and non-participants who responded to the postal survey were used.

Next, the data was converted from SPSS format to Stata format and the NatCen specification was estimated using Stata. The estimated coefficients in the propensity score logit were almost exactly the same as those obtained from SPSS.

Finally, the estimated propensity score provided to us by NatCen was compared with those obtained from doing what they did and estimating them in SPSS. Table 0.1 presents a random sample of the propensity score provided by the NatCen and the respective match estimated by us. The two sets of scores closely resembled one another, but were not quite the same. The correlation between the two sets of scores is 0.9989, and in almost all cases small differences between the two sets of scores were found in the third or fourth decimal place. Despite these differences, the close similarity between the two sets of scores, and those generated in the same manner using Stata, indicate that the estimation of the scores did not cause problems for the NatCen evaluation.

Table J.1 Random sample of the propensity scores estimated by NatCen and the authors

Observation serial number	Participation status	Propensity score	
		Estimated by NatCen and included in their dataset	Estimated by the authors using the NatCen syntax and their dataset
89	Participant	0.10499	0.101340
94	Participant	0.05791	0.057314
118	Non-Participant	0.09241	0.092089
128	Participant	0.07953	0.075889
135	Participant	0.26407	0.258298
140	Non-Participant	0.07876	0.081605
144	Non-Participant	0.01214	0.011988
164	Non-Participant	0.02901	0.027705
168	Non-Participant	0.03558	0.035114
169	Non-Participant	0.01276	0.012476
190	Non-Participant	0.01464	0.014519
192	Participant	0.09219	0.092361
199	Non-Participant	0.11734	0.112601
211	Non-Participant	0.32484	0.315494
219	Participant	0.03092	0.029640
230	Non-Participant	0.12675	0.127181
266	Participant	0.12968	0.127878
296	Non-Participant	0.03086	0.029851

J.2.2 Propensity score common support condition

As previously mentioned, propensity score matching requires that $0 < \Pr(D = 1|W) < 1$. The treatment on the treated parameter estimated in the NatCen evaluation of the NDLP requires only the second inequality hold in the sample. In words, the second inequality requires that for each value of $P(W)$ among participants in the sample, some non-participants in the sample have similar values of $P(W)$. The definition of similar depends on the particular method employed to impose the support

condition; methods in the literature include the use of callipers combined with nearest neighbour matching, as well as methods based on the estimated densities of $f(\hat{P}(W)|D=1)$ and $f(\hat{P}(W)|D=0)$. See Heckman, Ichimura and Todd (1997) and Smith and Todd (2005) for discussions.

In this section, things are kept simple and the sample densities of the estimated propensity scores are graphically examined. Figure J.1, Figure J.2, and Figure J.3, present histograms of the estimated propensity scores for the face-to-face interview sample, the matched sample from the NatCen matching algorithm, and the full sample, respectively.

Figure J.3 reveals a pattern typical of evaluations of voluntary active labour market policies, with many more observations with low participation probabilities than with high ones, and with non-participants over-represented at the low end of the estimated propensity score distribution.

Comparing Figure J.1 and Figure J.2, shows that the matching did what it should do by making the two densities look much more alike. The over-sampling of non-participants with estimated propensity scores over 0.3 does not even show up visibly in Figure J.1, which suggests, as shown later, that it will make little difference to the impact estimates. Comparing Figure J.1 and Figure J.2 shows that non-response to the face-to-face interview survey did little to affect either the symmetry of the two distributions or the support condition. Finally, although there are not many observations with high estimated propensity scores for either participants or non-participants, they both appear in roughly equal proportions in Figure J.1, indicating no support problems at this level of precision.

Figure J.1 Histogram: Propensity score distribution for the face-to-face interviewed sample (Groups B1 and C1)

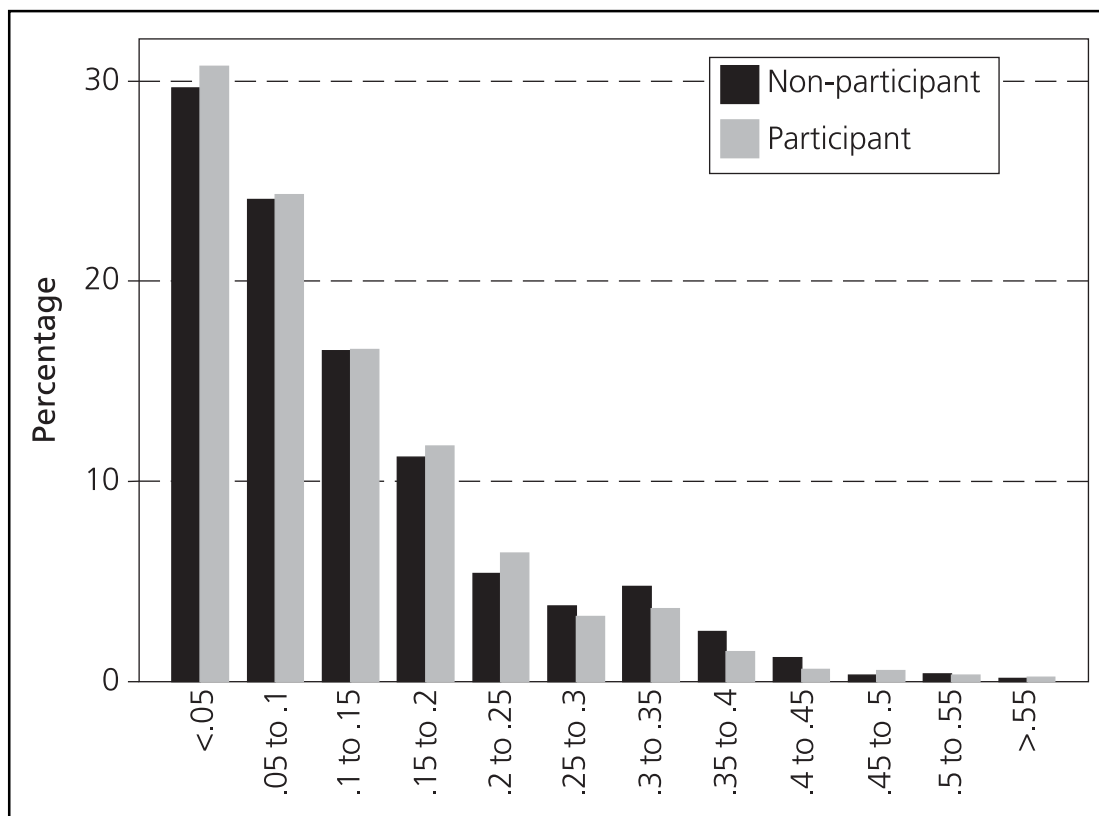


Figure J.2 Histogram: Propensity score distribution for the matched sample (Groups B1, B2, C1 and C2)

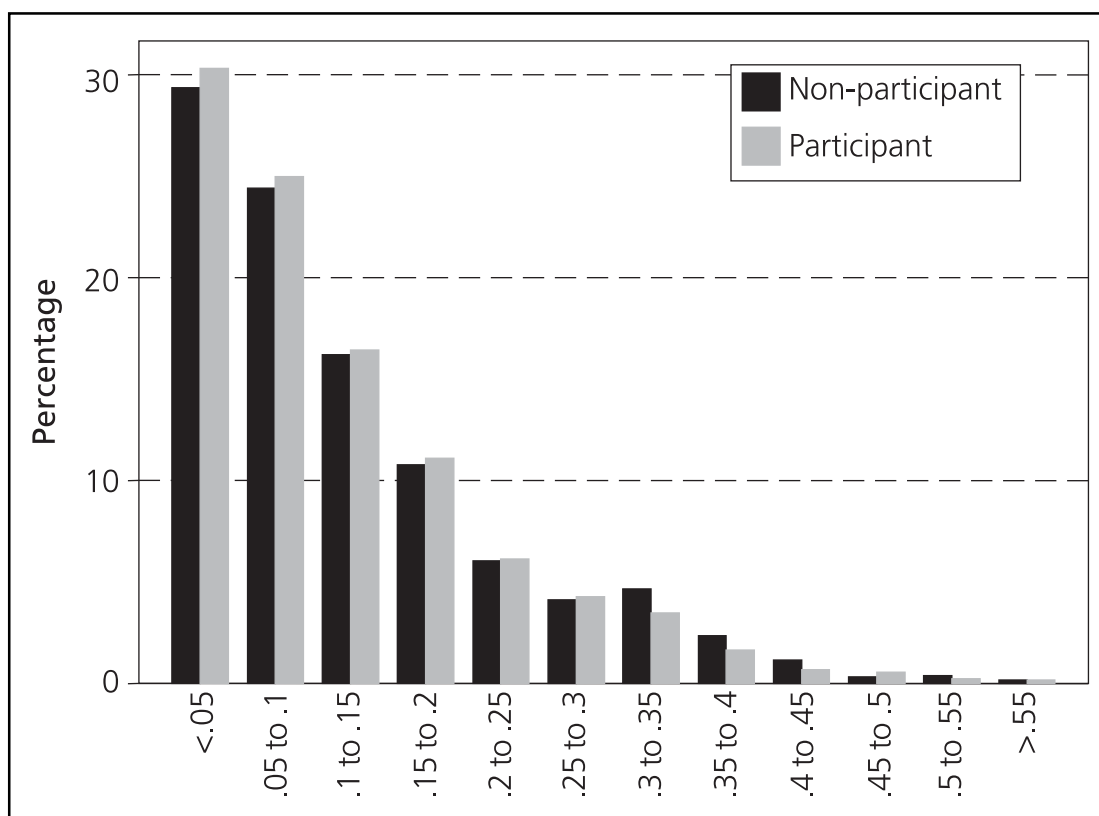
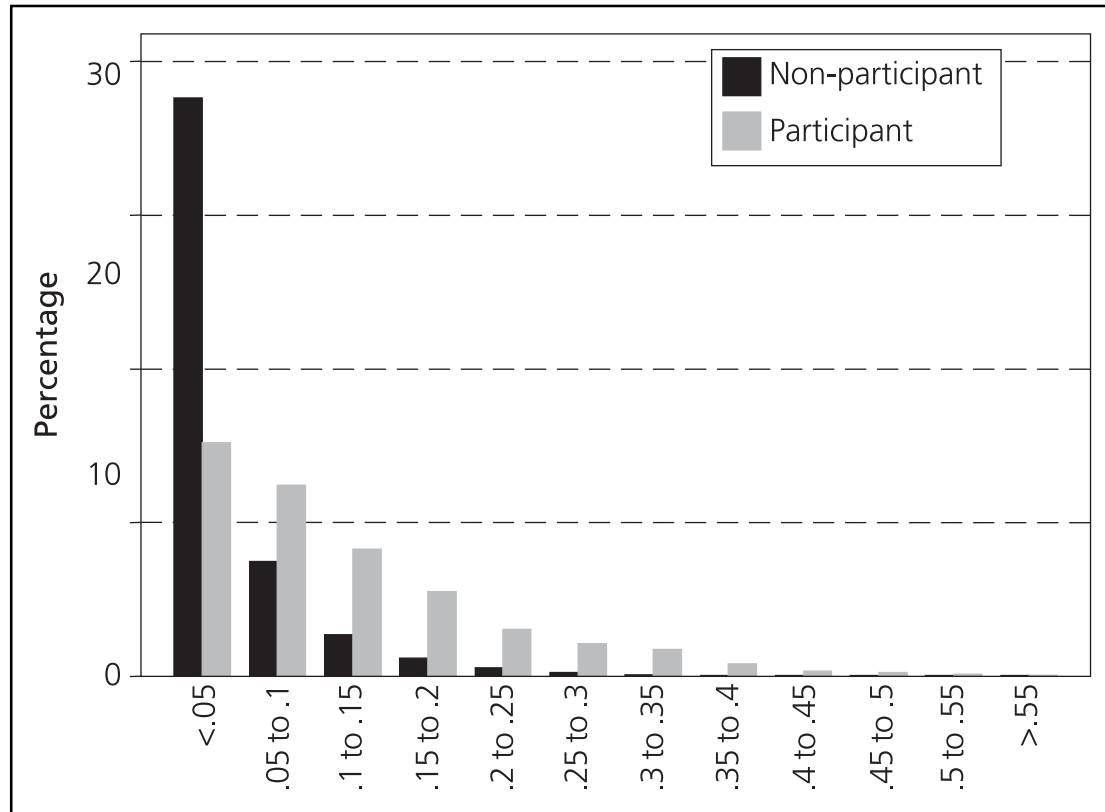


Figure J.3 Histogram: Propensity score distribution for the full sample (Groups B and C)



To further compare the distributions of estimated propensity scores in the full sample, particularly in the upper tail of the distribution, Box-Cox plots are presented in Figures J.4 and J.5. The first Box-Cox Plot in Figure J.4 covers the entire sample and shows that the upper tail of the distribution of the estimated propensity scores is quite thin for both the participants and the non-participants. For the plots in Figure J.5, the sample is divided into 'hex-tiles' based on the combined distribution of scores. Table J.2 displays the boundaries of the six-tiles, along with the number of participant and non-participant observations in each and other descriptive statistics. Taken together, Figure J.4 and Figure J.5 demonstrate that the two distributions look very much the same, other than in the upper tail. Depending on the exact method used to enforce the common support condition, Figure J.5 indicates the potential for common support problems for a handful of participant observations without non-participant observations close at hand, particularly in the presence of non-response.

Figure J.4 Box-Cox Plot of propensity scores for full sample, by participation status

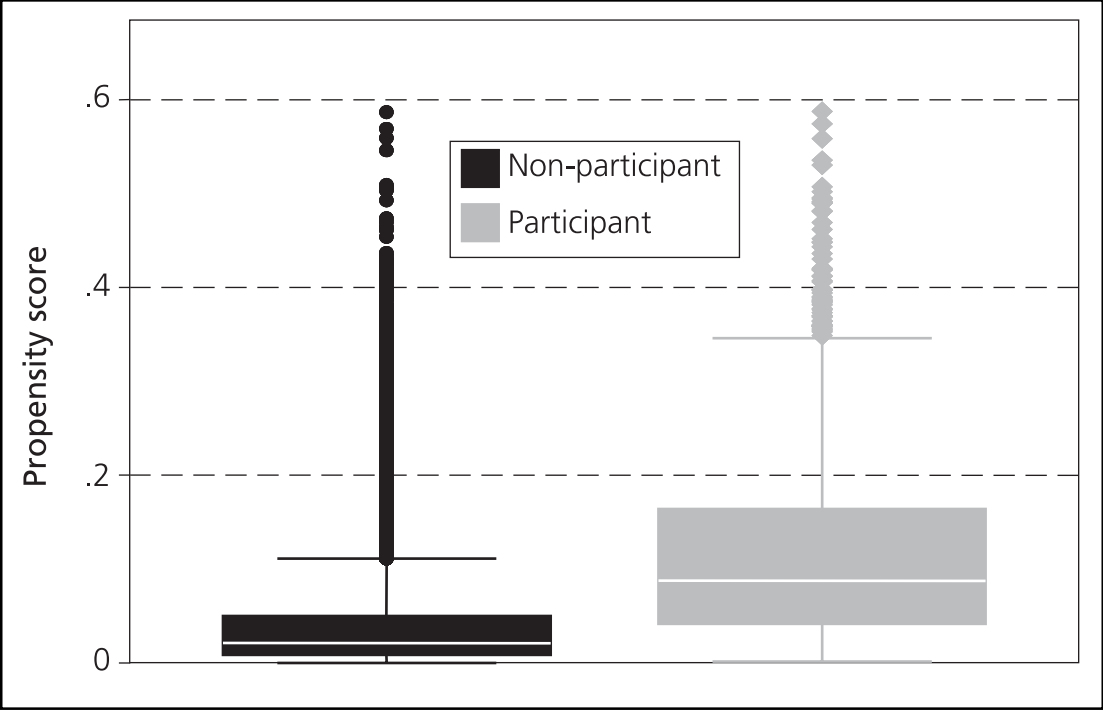


Figure J.5 Box-Cox plots of propensity scores for hex-tiles, by participation status

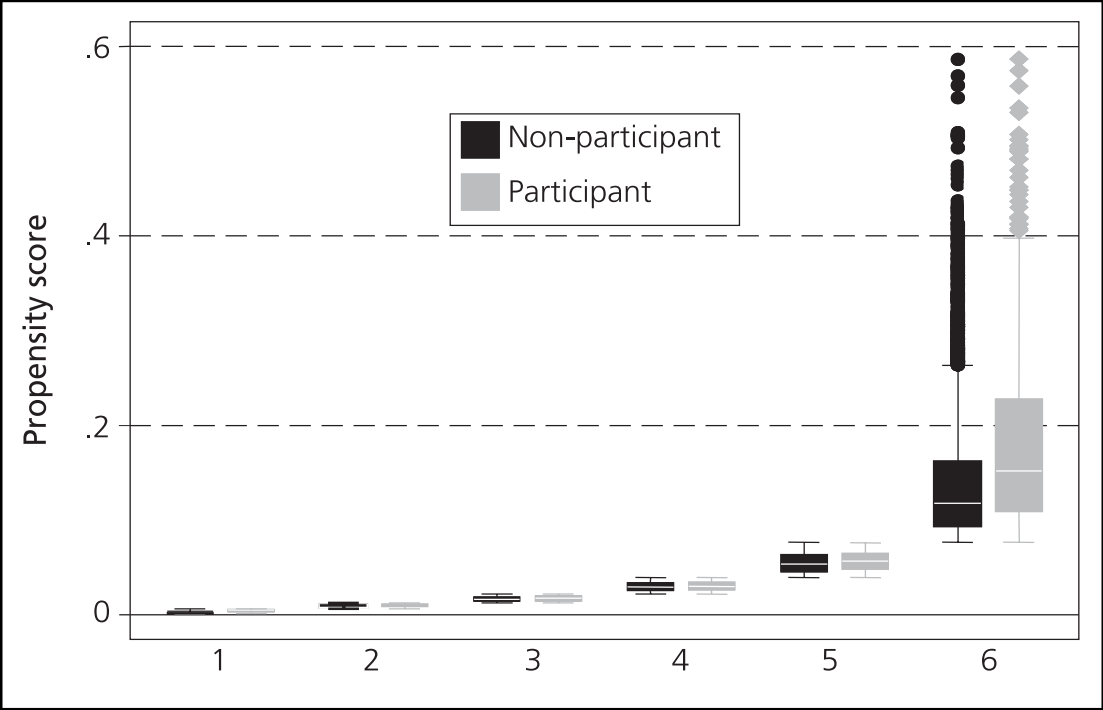


Table J.2 Six-tiles of the estimated propensity scores

Tiles	Group	Obs	Mean score	Std. dev.	Min score	Max score
1	Non-Participant	6,816	0.0041718	0.001420	0.000437	0.006552
	Participant	25	0.0046644	0.001345	0.001399	0.006417
2	Non-Participant	6,781	0.0092493	0.001664	0.006552	0.012304
	Participant	59	0.0096777	0.001607	0.006728	0.012295
3	Non-Participant	6,725	0.0167171	0.002793	0.012305	0.021962
	Participant	115	0.0171449	0.002746	0.012522	0.021949
4	Non-Participant	6,622	0.0297243	0.005003	0.021964	0.039388
	Participant	219	0.0300241	0.004869	0.022285	0.039359
5	Non-Participant	6,458	0.0549170	0.010536	0.039394	0.076278
	Participant	382	0.0566337	0.010270	0.039499	0.076245
6	Non-Participant	5,863	0.1383585	0.065411	0.076299	0.635578
	Participant	977	0.1841051	0.097445	0.076365	0.624328

J.3 Matching and covariate balancing

In this section, a number of issues are examined related to the creation of the matched sample of non-participants in the NatCen report. In sub section 0, response rates are examined in the matched sample. In sub section 0, the NatCen procedure of single nearest neighbour matching without replacement is repeated, but a matching routine is employed written in Stata rather than the matching routine written by NatCen in the Blaise statistical language. In sub section 0, the matching algorithm is changed from single nearest neighbour matching without replacement to single nearest neighbour matching with replacement. A comparison of matching with and without replacement provides evidence on average match quality in the comparison group sample drawn by NatCen without replacement. It also foreshadows the analysis using ex post re-matching in Section 0. In sub section 0, the success of the NatCen matching is evaluated at producing participant and matched non-participant sample with balanced covariates.

J.3.1 Response to the face-to-face interviews

In this section, response rates to the face-to-face interviews are considered. These response rates have implications for the ability of the impact analysis to rely on comparisons of matched pairs wherein both the matched participant and the matched non-participant responded to the face-to-face interview survey.

NatCen completed face-to-face interviews with 2,500 individuals out of the 3,596 initially identified by their matching algorithm (including the over-sampling of the non-participants with estimated propensity scores over 0.3). This figure represents a response rate of 69.5 per cent.

This response rate alone does not indicate whether the non-response process tended to preserve or break up matched pairs. In theory, matched pairs could always respond or not respond together, or matched pairs could disagree whenever possible, so that if one matched observation failed to respond, its partner always responded. Empirically, only 1,540 out of the 2,500 individuals in the sample with complete face-to-face interviews belong to a pair in which both members responded. In Section 0, how the impact estimates vary when we employ only observations from pairs wherein both members responded is shown, and also where re-matching to take account of the non-response (and resultant pair separation) is documented.

Table J.3 Complete pairs of matched face-to-face interview respondents

	Response in respect of participation		
	Non-participant	Participant	Total
Responded and matched partner did respond	770 (42.33)	770 (43.33)	1,540 (42.83)
Responded but matched partner did not respond	471 (25.89)	489 (27.52)	960 (26.7)
Did not respond to the face-to-face interview	578 (31.78)	518 (29.15)	1,096 (30.48)
Total	1,819 (100)	1,777 (100)	3,596 (100)

Note: Percentages in parentheses.

J.3.2 Matching without replacement

NatCen's program that performs single nearest neighbour matching without replacement was written in the programming language of the Blaise statistical software package. The NatCen routine identified a total 3,596 potential candidates for the face-to-face interview survey. As shown in Table J.3, the 3,596 includes 1,777 participants and 1,819 non-participants. In addition to the 1,777 matched nearest neighbours, the non-participant group includes all unmatched non-participants with estimated propensity scores above 0.3.

In order to test the NatCen matching program, and to understand the matching process more generally, our own code was written for single nearest neighbour matching without replacements using the Stata programming language. The matching routine was applied to the full sample of 40,979 postal survey respondents provided by the NatCen and described in Table J.4. For comparability, matching on the estimated propensity scores provided by NatCen which proved impossible to precisely replicate are indicated in Section 0.

Table J.4 Postal survey respondent sample

	Frequency	Per cent	Cumulative
Non-Participant	39,192	95.64	95.64
Participant	1,787	4.36	100
Total	40,979	100	

Following the NatCen analysis, the 287 observations with estimated propensity scores in excess of 0.3 were put aside, as they were all selected for face-to-face interviews regardless of match. The Stata routine for nearest neighbour matching without replacement was then run. The results from this exercise appear in Table J.5. The routine matched the 1,656 participants with estimated propensity scores less than 0.3 to the same number of non-participants. Out of the 1,656 matched non-participants identified by the routine, only 60 per cent were also selected by the NatCen matching algorithm.

Table J.5 Matching without replacement in the full sample

	Non-participant	Participant	Total
Selected by our algorithm only	644	0	644
Selected by NatCen algorithm and our algorithm	1,009	1,656	2,665
Total	1,653	1,656	3,309

The single nearest neighbour matching without replacement algorithm was applied to the data on the 2,500 individuals with face-to-face interview data. Of these, 2,296 had estimated propensity scores below 0.3. In this case, the 1,171 participants were matched to the same number of non-participants; of the latter, only 689 (58 per cent) were also selected by the NatCen Blaise matching algorithm. Table J.6 summarises this analysis.

Table J.6 Matching without replacement in the face-to-face interview sample

	Non-participant	Participant	Total
Selected by our algorithm only	436	0	436
Selected by NatCen algorithm and our algorithm	689	1,171	1,860
Total	1,125	1,171	2,296

J.3.3 Matching with replacement

The 'psmatch.do' routine written by Sianesi (2001b) in the Stata programming language was used to perform nearest neighbour matching with replacement. The

sample for this exercise consists of the full postal survey respondent sample. As in the analyses in Section 0, the estimated propensity scores provided by NatCen were used for comparability. It was found that of the 1,819 matched non-participants in the NatCen matched sample, the Sianesi (2001b) routine selected 963, or 54 per cent, as matches. Of the 1,241 non-participants completing the face-to-face interview survey, matching with replacement selected 647, or 51 per cent, as matches.

Checking the extent to which many non-participant observations get matched to multiple participants represents an important check on matching with replacement. If many non-participant observations get re-used, this suggests that a substantial reduction in bias may result, relative to matching without replacement; at the same time, because fewer actual observations play a role in constructing the counterfactual mean outcome, standard errors will increase. A finding that many non-participant observations get re-used (and, concomitantly, that many do not get used at all) also suggests the value of examining matching strategies such as multiple nearest neighbour matching and kernel matching, that utilise more than one observation in constructing the expected counterfactual for each participant.

Table J.7 displays the results from matching on the full sample of postal survey respondents. It was found that the vast majority of the non-participant observations that get matched, get matched to only one participant. Just 120 non-participants get matched to more than one participant and, of those, only 19 get matched more than twice. Table J.8 provides the corresponding numbers when the observations with estimated propensity scores in excess of 0.3 were dropped from the postal survey respondent sample, leaving a total of 40,698 observations.

The key finding here is that there is surprisingly little re-use of non-participant observations in the matching process. At first, it may seem that an error has been made, as the number of observations used more than once by the matched with replacement, falls well below the number of original NatCen matched non-participants not matched under matching with replacement. In fact, the results here combine two phenomena: First, some non-participant observations matched by NatCen without replacement get replaced by nearer non-participants matched multiple times under matching with replacement. Second, as in Section 0, some of the non-participants matched to only one participant by NatCen using their algorithm, get replaced by other non-participants matched to only one participant in the Sianesi (2001b) routine. The overall lesson from this section is that there is modest scope for differences between the original NatCen estimates and the estimates obtained from matching with replacement using the Sianesi (2001b) routine, both because it makes some different single matches and because some single matches get replaced with multiple matches.

Table J.7 Number of times non-participant observations used in matching with replacement: full postal survey respondent sample

Number of times used	Non-participant	Participant	Total
1	1,515	1,777	3,292
2	101	0	101
3	17	0	17
4	1	0	1
5	1	0	1
Total	1,635	1,777	3,412

Table J.8 Number of times non-participant observations used in matching with replacement: postal survey respondent sample with estimated propensity scores < 0.3

Number of times used	Non-participant	Participant	Total
1	1,456	1,656	3,112
2	82	0	82
3	12	0	12
Total	1,550	1,656	3,206

J.3.4 Covariate balancing

As discussed in Section 0, the suitability of a propensity score specification can be tested by testing whether or not, conditional on the estimated propensity scores, it ‘balances’ the distributions of the covariates used to estimate the score. Balancing in this context means that distribution of covariates for participants conditional on $\hat{P}(W)$ does not differ statistically from that of non-participants. This includes, in principle, all moments of the distribution, though in practice most balancing tests look only at conditional means. Balancing tests do **not** indicate anything about whether or not the variables included in the score suffice to make the CIA hold. Balancing tests do indicate whether or not the parametric specification employed to estimate the score is sufficiently flexible. The literature often employs balancing tests in an iterative fashion to refine the propensity score specification, adding higher order and interaction terms until the specification passes the test at a certain level. For good examples of the use of balancing tests in the literature, see Dehejia and Wahba (1999, 2002) and Lechner (1999, 2002).

In this section, covariate balance in the data from the NDLP evaluation is examined. Table J.9 presents covariate means, and p-values from t-tests of differences in means, for three different samples based on the NatCen matching and for matching with replacement using the Sianesi (2001b) matching routine in Stata. The first set of columns, marked (1), in the table gives the results for the nearest neighbour matching with replacement. Column (2) gives the results for the individuals selected for face-to-face interviews by the NatCen matching algorithm, Column (3) gives the

results for the individuals who actually completed a face-to-face interview and Column (4) gives the results for the complete matched pairs of face-to-face interview completers. In each case, observations with estimated propensity scores in excess of 0.3 are excluded in order to focus, for the present, on differences not resulting from the over-sampling of non-participants with high estimate scores. At the top of each set of columns, the number of t-tests for each sample that reject the null of equal means at the five and ten per cent significance levels are indicated.

Three interesting findings emerge from Table J.9: First, comparing columns (1) and (2) we see that the nearest neighbour matching with replacement does better at satisfying the balancing condition than does nearest neighbour matching without replacement, using the same set of estimated propensity scores. This is the result one would expect, given that matching with replacement yields closer matches. What is somewhat surprising is the strength of the improvement. Second, comparing sets (2) and (3) it is apparent that non-response to the face-to-face interviews actually improves the balance, as measured by the number of rejected nulls. This finding reflects the net of two opposing forces. Non-random non-response should make things worse, while the reduced sample size makes rejecting the null more difficult. Finally, comparing Columns (3) and (4), it can be seen that restricting attention to matched pairs in which both members responded to the face-to-face interview survey improves the balance. In this case, match quality and sample size work in the same direction; both should reduce the number of rejections in set (4) relative to set (3).

These results suggest that the NatCen could have amended their propensity score specification at the margin, to improve its performance in terms of balancing the covariates. They also suggest assigning a somewhat higher weight to estimates from nearest neighbour matching with replacement than to the estimates based on the NatCen matched samples.

Table J.10 repeats the analysis of Table J.9 for sets (2) and (3), but includes observations with estimated propensity scores greater than 0.3. Thus, it compares the full NatCen matched sample and the sample of face-to-face interview respondents. Comparing the results in Table J.10 with those in Table J.9 indicates the effect on the balancing results of the NatCen's over-sampling of non-participant observations with high estimated propensity scores. The second set of results in Table J.10 corresponds to the sample employed by NatCen to construct their primary impact estimates.

The table reveals a lack of balance for both the full NatCen matches sample and the sample of face-to-face interview respondents actually used to construct the NatCen impact estimates. Comparing Tables J.9 and J.10 makes it clear that the failure of the balancing test documented in Table J.9 does not result from the over-sampling of non-participants with high propensity scores. Overall, the results in this section indicate that NatCen should have made their propensity score specification more flexible before proceeding to a final analysis.

Table J.9 Extract of Balancing Tests for Samples with estimated propensity scores < 0.3

	(1)		(2)		(3)		(4)	
	Matching with replacement		NatCen full matched sample		NatCen matched sample: face-to-face interview respondents		NatCen matched sample: complete respondent pairs	
	Non-Part	Part	Non-Part	Part	Non-Part	Part	Non-Part	Part
Number of observations	1,550	1,656	1,656	1,656	1,125	1,171	770	770
Number of statistically significant mean differences	10% level	0	p-value of the t-test	p-value of the t-test	p-value of the t-test	p-value of the t-test	p-value of the t-test	p-value of the t-test
			0	24	22	11	11	0
			0	9	12	0	0	0
QATTMOTH - Attitudes towards mother working	13.04	13.11	0.51	13.06	13.11	0.64	13.02	13.01
QATTJOB - Attitudes towards retaining a job	8.67	8.74	0.40	8.69	8.74	0.55	8.72	8.69
_IQWHENWK_2 - when hoping to start work (next year)	0.26	0.26	0.84	0.27	0.26	0.37	0.26	0.26
_IQWHENWK_3 - when hoping to start work (next 3 years)	0.25	0.23	0.15	0.26	0.23	0.04	0.28	0.23
_IQWHENWK_4 - when hoping to start work (not in next 3 years)	0.10	0.11	0.45	0.11	0.11	0.73	0.12	0.11
_IQWHENWK_5 - when hoping to start work (currently in work/ don't know/missing)	0.11	0.11	0.94	0.07	0.11	0.00	0.08	0.13

Note: Full table in Appendix C.

Table J.10 Extract of balancing tests for NatCen full sample and face-to-face interview respondent sample

	(1)			(2)		
	NatCen full matched sample		p-value of the t-test	NatCen matched sample: face-to-face interview respondents		p-value of the t-test
	Non-Part	Part		Non-Part	Part	
	1,819	1,777		1,241	1,259	
Number of observations:						
Number of statistically significant mean differences	10% level		17			14
	5% level		11			12
QATTMOTH - Attitudes towards mother working	13.212	13.228	0.880	13.174	13.137	0.771
QATTJOB - Attitudes towards retaining a job	8.753	8.790	0.630	8.783	8.732	0.589
_IQWHENWK_2 - when hoping to start work (next year)	0.255	0.246	0.552	0.239	0.253	0.442
_IQWHENWK_3 - when hoping to start work (next 3 years)	0.234	0.212	0.104	0.254	0.214	0.018
_IQWHENWK_4 - when hoping to start work (not in next 3 years)	0.104	0.103	0.928	0.110	0.106	0.750
_IQWHENWK_5 - when hoping to start work (currently in work/ don't know/missing)	0.067	0.106	0.000	0.075	0.113	0.001

Note: Full table in Appendix D.

J.4 Sensitivity of the impact estimates to the matching method

In this section, the sensitivity of the impact estimates to various aspects of the matching method employed to obtain them is explored. For simplicity, only a subset of the outcome variables examined in the NatCen report is considered. Table J.11 lists the ten outcome variables we examine. The training and attitudes to work outcomes are omitted in part because they have a much higher rate of missing values. All of the outcome variables in the set tell essentially the same story, which suggests that not much is lost by focusing on only a subset. The list in Table J.11 includes outcome variables from both the administrative data and from the face-to-face interview survey. Some of the variations on the matching estimator which are examined require outcome data on individuals not responding to the face-to-face interview (either because they were not matched by the NatCen or because they were matched but did not respond); for those variations, only the outcome variables from the administrative data are utilised. Variables measured relative to the time of NDLP participation require a pseudo-participation date for non-participants. The participation date for their matched participant plays this role.

Table J.11 Outcome definitions

Variable Name	Variable label	Value code	Value label
Variables from the administrative data			
WLEFTISB	whether left IS at any time since participation	0	No
		1	Yes
WLEFTISC	whether left IS at any time since sample date	0	No
		1	Yes
PART180B	whether on IS 6 months after participation	0	No
		1	Yes
PART180C	whether on IS 6 months after sample date	0	No
		1	Yes
TONISA	duration of time on IS since participation measured in days	0	No
		1	Yes
TONISC	duration of time on IS since sample date measured in days	0	No
		1	Yes
Variables from the Face-to-face interview:			
JOBNOW	in full-time or part-time work as of the face-to-face interview	0	No
		1	Yes
JOBFTNOW	currently in full-time work as of the face-to-face interview	0	No
		1	Yes
ANYJBAD	had some job during or after participation date	0	No
		1	Yes
WORKAT6	working 6 months after participation	0	No
		1	Yes

Before the estimates are presented, it is important to make a few remarks regarding the three main potential sources of differences among the impact estimates presented in this section.

The first potential source of differences is the role played by those observations with a propensity score higher than 0.3 (for more details see Section I.5). Table J.11 presents the various matching estimates when observations with estimated propensity scores in excess of 0.3 are included, while Table J.11 presents estimates that exclude these observations. Comparing the two sets of estimates enables the determination of the extent to which the over-sampling of these individuals affects the estimates.

The second potential source of differences between impact estimates is differences in the samples of participants and non-participants utilised in constructing the estimates. The samples differ for several reasons, which are described in detail in the

discussion below. The first three columns of Table J.11 report estimates for outcomes drawn from the administrative data and from the face-to-face interview survey. All of the columns in Table J.11 report estimates for outcomes drawn from the administrative data only, as the matching estimators examined in these tables include non-participants who did not complete the face-to-face interview.

The third and last source of potential differences among the impact estimates consists of differences in the matching algorithms. A variety of different matching methods are utilised including single nearest neighbour matching with and without replacement and kernel matching with a variety of different bandwidths. The variety of methods employed allows the examination of the sensitivity of the results to changes in the matching algorithm within a set of reasonable alternatives.

The impact estimates of Table J.11 are now considered. The first five sets of estimates of Table J.11 examine the effects of changing the sample utilised in the matching. In so doing, they illuminate the effects of differential non-response to the postal and face-to-face interview surveys. They also show the value of ex post re-matching (with replacement) within the sample of face-to-face interview respondents. Column (1) presents estimates using the algorithm from the NatCen report. These estimates consist of the simple mean difference between participant respondents and non-participant respondents to the face-to-face interview survey. While NatCen did single nearest neighbour matching to determine who they attempted to interview, due to non-response, this sample includes many individuals whose match did not respond to the survey. Column (2) presents mean-difference estimates using only the 770 complete matched pairs among the respondents to the face-to-face interview survey. Column (3) presents estimates from re-matching the face-to-face interview respondents using single nearest neighbour matching **with** replacement. Column (4) presents estimates based on single nearest neighbour matching with replacement using the full NatCen matched samples, both those responding to the face-to-face interview survey and those who did not do so. Because this estimate includes non-respondents to the face-to-face interview survey, only present impacts on outcome variables drawn from the administrative data can be presented. Similarly, Column (5) presents estimates from single nearest neighbour matching with replacement on the entire postal survey respondent sample. Again, only impacts on outcomes from the administrative data are reported.

Several things are learned from the estimates in the first five columns. Restricting the sample of face-to-face interview respondents changes the estimates remarkably little. Re-matching with replacement in the face-to-face interview sample changes the estimates remarkably little. Matching with replacement using the entire NatCen matched sample, rather than just the face-to-face interview sample, changes the estimates remarkably little, which suggests that non-response to the face-to-face interview is more or less random.

Finally, Column (5) in Table J.11 shows that matching with replacement using the entire sample of postal survey respondents does make a difference to the estimates.

Most of the estimated impacts get larger in absolute value, with three of them flipping signs. Large changes are seen for whether or not the respondent is on Income Support (IS) six months after participation (part180b), which increases (in absolute value) from -0.267 to 0.084, for the duration of the IS spell starting at participation (Tonisa), which changes from -59.6 to 40.0 and for the IS spell starting at the sampling date (Tonisc), which changes from -59.3 to -32.3. These results are surprising and there is no good explanation for them; speculation about possible causes is discussed below.

The next six columns in Table J.11 present matching estimates obtained using kernel matching. Because kernel matching uses, on average, more distant non-participant observations than nearest neighbour matching (with distance measured in propensity score terms), it has a larger bias. At the same time, because kernel matching uses more information – weighting every non-participant observation by its distance in constructing the estimated counterfactual mean – kernel matching estimates will have a lower variance. The six sets of kernel matching estimates differ only in their bandwidths. The larger the bandwidth, the greater the weight attached to more distant observations. All of the kernel matching estimates use the standard normal density as the kernel. Columns (6) to (9) present estimates for bandwidths of 0.05, 0.01, 0.005 and 0.001. These were selected as reasonable bandwidths following visual inspection of the distributions of estimated propensity scores. Columns (10) and (11) present kernel matching estimates obtained using various ‘rule of thumb’ bandwidths from Silverman (1986). These bandwidths, which fall in the interior of the range selected through visual inspection, minimise the mean squared error under certain assumptions (that do not apply in this context).

The kernel matching estimates closely resemble the estimates from the other matching algorithms (Column (5)) for the outcomes measured relative to the sample date, but differ from them quite strongly for the outcomes measured relative to participation. At the present time, there is no clear explanation for this puzzle. Estimates rarely differ this much between kernel and nearest neighbour matching in the published literature that compares the two on common data. Within the set of kernel matching algorithms, bandwidth choice matters slightly in some cases, but never very much, which is consistent with the published literature.

The next two columns in Table J.11 present estimates from nearest neighbour matching with replacement with a calliper. The calliper is a range around the estimated propensity score value for each participant observation. Participants with no non-participants within their calliper get dropped from the estimation for being outside the ‘common support’ discussed in Section I.5. This has two effects. Firstly, it changes the definition of the parameter being estimated from the impact of treatment on the treated to the impact of treatment on treated individuals in the common support. Secondly, it means that no really poor matches get included in the estimation. The first effect is the cost of the second. Column (12) presents estimates with a calliper of 0.01 and Column (13) presents estimates with a calliper of 0.005. The first row tells the story in this case. The calliper of 0.01 results in the loss of only

two observations from the sample, while the calliper of 0.005 results in the loss of seven observations. Not surprisingly, this has essentially no effect on the estimates, when compared to the estimates from single nearest neighbour with replacement (and no calliper) on the full sample of postal respondents in Column (5).

The final set of estimates in Table J.11 relies on single nearest neighbour matching without replacement in the full sample of postal survey respondents. Comparing the estimates in Column (14) to those in Column (5) illustrates the effect of matching without replacement in a sample with fewer non-participants than participants with high propensity scores. In such a sample, nearest neighbour matching without replacement (unless disciplined by a calliper, which raises its own issues) forces some fairly distant matches. The estimates here show a modest (but far from trivial) effect for some outcome variables. For example, the estimates for having left IS at any time since participation (*wleftisb*) differ by only 0.01 – 0.373 with replacement versus 0.363 without it. In contrast, for duration of the IS spell starting at participation, the two estimates differ by over 50 per cent – 40.0 days with replacement and 25.5 days without replacement. These results highlight the dangers of matching without replacements in situations with unbalanced distributions of propensity scores.

The estimates excluding observations with propensity scores above 0.3 are presented in Table J.11, so they are not described in detail here. Excluding individuals with an estimated propensity score over 0.3 does not affect the impact estimates, which implies that the over-sampling of these individuals is not a major problem in this evaluation.

Overall, much the same conclusion is reached as that reached in the recent academic literature, such as Heckman, Ichimura, Smith and Todd (1998), Dehejia and Wahba (1999, 2001) and Smith and Todd (2005). Other than the choice of whether to match with replacement or not, the details of the matching algorithm, within the set of reasonable choices, do not have major substantive effects on the resulting estimates. By reasonable choices, we mean putting to one side the choice of large or tiny bandwidths, or tiny callipers, whose inappropriateness even a casual inspection of the data would reveal.

The primary exception to this general conclusion concerns the differences in the estimates for the participation relative to the sample and participation data variables (*PART180B* and *PART180C*) and the length of time on IS since participation variable (*TONISA*) between Columns (1) to (4) and Columns (5) to (13). There is no good explanation for these differences, but there are some concerns regarding data anomalies based on a preliminary examination of these variables in the administrative data.

This chapter explored the robustness of the NatCen results to changes in the matching algorithm employed and in the specification of the propensity score model, conditional on using only variables from the postal survey.

- repeated the NatCen matching using routines written in Stata rather than in BLAISE;
- performed a balancing test using the NatCen propensity scores;
- examined matching using only matched pairs in which both persons respond;
- rematched using the face-to-face interview sample and a variety of matching methods including nearest neighbour with and without replacement and kernel matching with various bandwidths; and
- rematched using the postal interview sample and alternative propensity score specifications that employ only variables available on the postal survey.

Table J.12 **Distribution of NDLP participants and total eligible population by claim duration and child age by Wave 1/ Wave 2 or booster sample**

Stata	NDLP participants in eligible population			Total eligible population		
	Wave 1/2	Booster	Total	Wave 1/2	Booster	Total
1 - y0-3, less than 3 months	104	516	620	1,853	10,139	11,992
2 - y0-3, 3 months - less than 6months	354	0	354	6,198	0	6,198
3 - y0-3, 6 months- less than 1 year	615	0	615	11,405	0	11,405
4 - y0-3, 12 months to 2 yrs	866	0	866	17,883	0	17,883
5 - y0-3, 2 yrs to less than 3 yrs	490	0	490	11,347	0	11,347
6 - y0-3, 3 yrs or more	495	0	495	21,122	0	21,122
7 - y3-5, less than 3 months	61	276	276	782	3,899	4,681

Continued

Table J.12 Continued

Stata	NDLP participants in eligible population			Total eligible population		
	Wave 1/2	Booster	Total	Wave 1/2	Booster	Total
8 - y3-5, 3 months - less than 6months	130	0	130	2,071	0	2,071
9 - y3-5, 6 months- less than 1 year	213	0	213	3,264	0	3,264
10 - y3-5, 12 months to 2 yrs	341	0	341	5,838	0	5,838
11 - y3-5, 2 yrs to less than 3 yrs	309	0	309	4,899	0	4,899
12 - y3-5, 3 yrs or more	1,048	0	1,048	22,425	0	22,425
13 - y5-11, less than 3 months	113	470	583	1,435	7,401	8,836
14 - y5-11, 3 months -less than 6months	295	0	295	3,932	0	3,932
15 - y5-11, 6 months-less than 1 year	341	0	341	5,687	0	5,687
16 - y5-11, 12 months to 2 yrs	561	0	561	9,819	0	9,819
17 - y5-11, 2 yrs to less than 3 yrs	380	0	380	7,337	0	7,337
18 - y5-11, 3 yrs or more	2,138	0	2,138	48,290	0	48,290
19 - y11-16, less than 3 months	39	252	291	815	4,384	5,199
20 - y11-16, 3 months -less than 6months	150	0	150	2,229	0	2,229
21 - y11-16, 6 months-less than 1 year	184	0	184	3,189	0	3,189
22 - y11-16, 12 months to 2 yrs	228	0	228	5,207	0	5,207
23 - y11-16, 2 yrs to less than 3 yrs	170	0	170	3,849	0	3,849
24 - y11-16, 3 yrs or more	1,190	0	1,190	29,990	0	29,990
Total	10,815	1,514	12,329	230,866	25,823	256,689

Table J.13 Distribution of NDLP participants and total sample by duration of claim and age of youngest child by Wave 1/ Wave 2 or booster sample

Stata	NDLP participants in sample			Total sample		
	Wave 1/2	Booster	Total	Wave 1/2	Booster	Total
1 - y0-3, less than 3 months	119	130	249	1,834	2107	3,941
2 - y0-3, 3 months - less than 6months	124	0	124	1,814	0	1,814
3 - y0-3, 6 months- less than 1 year	238	0	238	3,503	0	3,503
4 - y0-3, 12 months to 2 yrs	320	0	320	5,354	0	5,354
5 - y0-3, 2 yrs to less than 3 yrs	139	0	139	2,869	0	2,869
6 - y0-3, 3 yrs or more	122	0	122	4,174	0	4,174
7 - y3-5, less than 3 months	69	53	122	779	688	1,467
8 - y3-5, 3 months - less than 6months	161	0	161	2,055	0	2,055
9 - y3-5, 6 months- less than 1 year	93	0	93	1,303	0	1,303
10 - y3-5, 12 months to 2 yrs	115	0	115	1,810	0	1,810
11 - y3-5, 2 yrs to less than 3 yrs	106	0	106	1,428	0	1,428
12 - y3-5, 3 yrs or more	267	0	267	4,568	0	4,568
13 - y5-11, less than 3 months	134	106	240	1,419	1046	2,465
14 - y5-11, 3 months - less than 6 months	334	0	334	3,885	0	3,885
15 - y5-11, 6 months- less than 1 year	205	0	205	2,825	0	2,825
16 - y5-11, 12 months to 2 yrs	193	0	193	2,981	0	2,981
17 - y5-11, 2 yrs to less than 3 yrs	124	0	124	2,213	0	2,213
18 - y5-11, 3 yrs or more	497	0	497	9,660	0	9,660
19 - y11-16, less than 3 months	48	69	117	807	827	1,634
20 - y11-16, 3 months -less than 6months	55	0	55	646	0	646

Continued

Table J.13 Continued

Stata	NDLP participants in sample			Total sample		
	Wave 1/2	Booster	Total	Wave 1/2	Booster	Total
21 - y11-16, 6 months-less than 1 year	51	0	51	911	0	911
22 - y11-16, 12 months to 2 yrs	47	0	47	1,027	0	1,027
23 - y11-16, 2 yrs to less than 3 yrs	41	0	41	909	0	909
24 - y11-16, 3 yrs or more	271	0	271	6,027	0	6,027
Total	3,873	358	4,231	64,801	5,028	69,829

Table J.14 Non-response to postal questionnaire by participation in NDLP

Whether participated in NDLP	Response to the postal questionnaire		Total
	Non-respondent	Respondent	
Non-participants in NDLP			
Number	25,793	38,683	64,476
%	40.0	60.0	100
Participants in NDLP			
Number	1,247	2,980	4,227
%	29.5	70.5	100
Total			
Number	27,040	41,663	68,703
%	39.36	60.64	100

Pearson's chi-squared for the hypothesis that the rows and columns in a two-way table are independent: Pearson chi2(1) = 183.3549 Pr = 0.000

Table J.15 Non-response to face-to-face interview survey by participation in NDLP

Non-participant in NDLP	Face-to-face interviewed		Total
	Not interviewed	Interviewed	
Non-participant in NDLP			
Number	569	1,236	1,805
%	31.52	68.48	100
Participant in NDLP			
Number	518	1,256	1,774
%	29.2	70.8	100
Total			
Number	1,087	2,492	3,579
%	30.37	69.63	100

Pearson's chi-squared for the hypothesis that the rows and columns in a two-way table are independent: Pearson chi2(1) = 2.2850 Pr = 0.131

The final number of observations on each stata in the sample used in these calculations is tabulated below:

Table J.16 Sample category – age of youngest child – benefit duration

	Frequency	Per cent	Cumulative
1 - y0-3, lt 3months	3,942	5.64	5.64
2 - y0-3, 3-lt6months	1,814	2.6	8.24
3 - y0-3, 6months-lt12months	3,503	5.01	13.26
4 - y0-3, 12months to lt 2 yrs	5,357	7.67	20.92
5 - y0-3, 2 yrs to lt 3 yrs	2,869	4.11	25.03
6 - y0-3, 3 yrs or more	4,179	5.98	31.01
7 - y3-5, lt 3months	1,468	2.1	33.12
8 - y3-5, 3-lt6months	2,055	2.94	36.06
9 - y3-5, 6months-lt12months	1,303	1.87	37.92
10 - y3-5, 12months to lt 2 yrs	1,810	2.59	40.51
11 - y3-5, 2 yrs to lt 3 yrs	1,428	2.04	42.56
12 - y3-5, 3 yrs or more	4,571	6.54	49.1
13 - y5-11, lt 3months	2,826	4.05	53.15
14 - y5-11, 3-lt6months	3,885	5.56	58.71
15 - y5-11, 6months-lt12months	2,826	4.05	62.76
16 - y5-11, 12months to lt 2 yrs	2,983	4.27	67.03
17 - y5-11, 2 yrs to lt 3 yrs	2,213	3.17	70.2
'8 - y5-11, 3 yrs or more	9,663	13.83	84.03
Continued			

Table J.16 Continued

	Frequency	Per cent	Cumulative
19 - y11-16, lt 3months	1,634	2.34	86.37
20 - y11-16, 3-lt6months	646	0.92	87.29
21 - y11-16, 6months-lt12months	911	1.3	88.6
22 - y11-16, 12months to lt 2 yrs	1,028	1.47	90.07
23 - y11-16, 2 yrs to lt 3 yrs	909	1.3	91.37
24 - y11-16, 3 yrs or more	6,028	8.63	100
Total	69,851	100	

Table J.17 Variables used and respective values

Variable Name	Variable label	Value code	Value label
QUESRESP:	respondent of the postal questionnaire	0	Non-respondent
		1	Respondent
DTA3596:	candidate for the face-to-face interview (selected by the matching)	0	Non candidate
		1	Candidate
DTA2500:	responded to the face-to-face interview survey	0	Non-respondent
		1	Respondent
CXPSEX	Sex	1	F
		2	M
CXPTRFG	respondent has a partner	1	N
		2	Y
CNPCHILD	number of children from 0-12 years		
PARTIP	whether participated in NDLP	0	Not participated
		1	Participated
RURALFLG	rural ward flag	0	Non-rural
		1	Rural
JCREG2	JobCentreplus region	1	East of England
		2	East Midlands
		3	London
		4	North East
		5	North West
		6	Scotland
		7	South East
		8	South West
		9	Wales
		10	West Midlands
		11	Yorkshire and the Humber

Continued

Table J.17 Continued

Variable Name	Variable label	Value code	Value label
WLEFTISB	whether left IS at any time since participation	0	No
		1	Yes
WLEFTISC	whether left IS at any time since participation	0	No
		1	Yes
WSTRISB	whether returned to IS since participation	0	No
		1	Yes
WSTRISC	whether returned to IS since sample date	0	No
		1	Yes
PART180C	whether on IS 6 months after sample data	0	No
		1	Yes
PART180B	whether on IS 6 months after participation	0	No
		1	Yes

Model specifications:

Model 1:

$$Y_i = \alpha + CXPSEX_i \beta_1 + CXPPTFRG_i \beta_2 + CNPCHILD_i \beta_3 + RURALFLG_i \beta_4 + PARTIP_i \beta_5 + \varepsilon_i$$

Model 2:

$$Y_i = \alpha + CXPSEX_i \beta_1 + CXPPTFRG_i \beta_2 + CNPCHILD_i \beta_3 + RURALFLG_i \beta_4 + PARTIP_i \beta_5 + JCREG2_i \beta_6 + \varepsilon_i$$

Model 3:

$$Y_i = \alpha + CXPSEX_i \beta_1 + CXPPTFRG_i \beta_2 + CNPCHILD_i \beta_3 + RURALFLG_i \beta_4 + OUTCOME_i^k \beta_5 + \text{int} : PARTIP_i * OUTCOME_i^k \beta_6 + \varepsilon_i$$

Model 4:

$$Y_i = \alpha + CXPSEX_i \beta_1 + CXPPTFRG_i \beta_2 + CNPCHILD_i \beta_3 + RURALFLG_i \beta_4 + PARTIP_i \beta_5 + JCREG2_i \beta_6 * OUTCOME_i^k \beta_7 + \varepsilon_i$$

Model 5:

$$Y_i = \alpha + CXPSEX_i \beta_1 + CXPPTFRG_i \beta_2 + CNPCHILD_i \beta_3 + RURALFLG_i \beta_4 + \text{int} : PARTIP_i * OUTCOME_i^k \beta_5 + JCREG2_i \beta_6 + \varepsilon_i$$

Where,

Y_i is either QUESRESP_i or DTA2500_i

:JCREG2: mult categorical variable representing 11 different regions in the UK

$OUTCOME_i^k$: vector of 6 different ex-post outcome measures from the Administrative data (wleftisb, wleftisc, wstrisb, wstrisc, part180b, part180c)

Table J.18 Multivariate probit models of non-response to both surveys

	Response to postal questionnaire (dep:quesresp)				Response to face-to-face interviewsurvey (dep: Dta2500)				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)
Sex==Male	-0.296 (0.021)**	-0.296 (0.021)**	-0.282 (0.021)**	-0.282 (0.021)**	-0.281 (0.021)**	-0.268 (0.107)*	-0.28 (0.107)**	-0.274 (0.108)*	-0.287 (0.108)**
Respondent has a partner ==Yes	-0.391 (0.074)**	-0.388 (0.074)**	-0.317 (0.076)**	-0.316 (0.076)**	-0.314 (0.077)**	-0.242 (0.400)	-0.188 (0.404)	-0.239 (0.399)	-0.277 (0.403)
Number of children from 0-12 years	-0.019 (0.005)**	-0.021 (0.005)**	-0.02 (0.005)**	-0.021 (0.005)**	-0.021 (0.005)**	0.074 (0.026)**	0.08 (0.027)**	0.07 (0.027)**	0.078 (0.027)**
Rural ward flag	0.117 (0.015)**	0.085 (0.016)**	0.131 (0.015)**	0.096 (0.017)**	0.097 (0.017)**	0.204 (0.064)**	0.143 (0.069)*	0.221 (0.064)**	0.152 (0.069)*
Whether participated in NDLP (PARTIP)	0.273 (0.021)**	0.273 (0.021)**	0.144 (0.173)	-0.846 (0.088)**	0.14 (0.173)	0.059 (0.044)	0.063 (0.045)	0.422 (0.248)+	0.152 (0.048)**
Whether left IS at any time since participation (WLEFTISB)			-0.315 (0.061)**	0.83 (0.050)**	-0.314 (0.061)**			0.2 (0.230)	-0.533 (0.170)
Interaction: WLEFTISB*PARTIP			1.007 (0.000)**		1.014 (0.000)**			-0.23 (0.353)	
Whether left IS at any time since sample date (WLEFTIS)			-0.225 (0.172)	-0.083 (0.015)**	-0.236 (0.172)			-0.508 (0.242)*	0.22 (0.177)**
Interaction: WLEFTISC*PARTIP			-0.035 (0.175)		-0.033 (0.175)			0.26 (0.367)	
Whether returned to IS since participation (WSTRISB)			-0.076 (0.015)**	0.443 (0.095)**	-0.079 (0.015)**				
									Continued

Continued

Table J.18 Continued

	Response to postal questionnaire (dep:quesresp)					Response to face-to-face interviewsurvey (dep: Dta2500)			
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)
Location: East Midlands		-0.037 (0.032)		-0.033 (0.032)	-0.032 (0.032)		0.205 (0.137)		0.202 (0.137)
Location: London		-0.065 (0.028)*		-0.077 (0.029)**	-0.076 (0.029)**		-0.071 (0.126)		-0.137 (0.126)
Location: North East		-0.001 (0.034)		0.004 (0.035)	0.003 (0.035)		0.385 (0.147)**		0.381 (0.148)**
Location: North West		-0.128 (0.028)**		-0.127 (0.028)**	-0.127 (0.028)**		0.082 (0.121)		0.062 (0.121)
Location: Scotland		-0.108 (0.030)**		-0.109 (0.030)**	-0.108 (0.030)**		0.106 (0.129)		0.113 (0.129)
Location: South East		-0.014 (0.030)		-0.015 (0.031)	-0.014 (0.031)		0.14 (0.131)		0.134 (0.131)
Location: South West		-0.03 (0.032)		-0.033 (0.033)	-0.033 (0.033)		0.341 (0.138)*		0.328 (0.139)*
Location: Wales		-0.048 (0.032)		-0.047 (0.032)	-0.047 (0.032)		0.484 (0.141)**		0.482 (0.142)**
Location: West Midlands		-0.053 (0.031)+		-0.048 (0.032)	-0.046 (0.032)		0.033 (0.139)		0.008 (0.140)
Location: Yorkshire and the		-0.029 (0.034)		-0.024 (0.034)	-0.024 (0.034)		0.021 (0.147)		-0.002 (0.148)
Constant	0.299 (0.011)**	0.37 (0.028)**	0.03 (0.022)	0.135 (0.034)**	0.104 (0.034)**	0.344 (0.055)**	0.209 (0.122)+	0.418 (0.202)*	0.506 (0.170)**
Observations	67,894	67,817	65,336	65,261	65,261	3,544	3,540	3,544	3,540

Standard errors in brackets. + significant at 10%; * significant at 5%; ** significant at 1%.

Table J.19 Relative frequencies of the population, sample and NatCen weights

Category	Relative frequency	Relative frequency	Relative frequency population/	NATCen weights	
	Sample	Population	Rel fre sample	Stata weights	Postal questionnaire
1 - y0-3, lt 3 months	0.031	0.010	0.313	0.32	0.33
2 - y0-3, 3-lt 6 months	0.032	0.033	1.022	0.98	1.03
3 - y0-3, 6 months-lt 12 months	0.061	0.057	0.925	0.96	1.00
4 - y0-3, 12 months to 24 months	0.083	0.080	0.969	0.98	1.01
5 - y0-3, 2 yrs to lt 3 yrs	0.036	0.045	1.262	1.15	1.12
6 - y0-3, 3 yrs or more	0.032	0.046	1.453	1.49	1.43
7 - y3-5, lt 3 months	0.018	0.006	0.317	0.33	0.35
8 - y3-5, 3-lt 6months	0.042	0.012	0.289	0.30	0.30
9 - y3-5, 6 months-lt 12 months	0.024	0.020	0.820	0.74	0.76
10 - y3-5, 12 months to 24 months	0.030	0.032	1.062	0.94	0.95
11 - y3-5, 2 yrs to 3 yrs	0.027	0.029	1.044	0.99	0.96
12 - y3-5, 3 yrs or more	0.069	0.097	1.406	1.46	1.39
13 - y5-11, lt 3 months	0.035	0.010	0.302	0.32	0.34
14 - y5-11, 3-lt 6months	0.086	0.027	0.316	0.30	0.31
15 - y5-11, 6 months-12 months	0.053	0.032	0.596	0.59	0.61
16 - y5-11, 12 months to 24 months	0.050	0.052	1.041	0.96	0.98
17 - y5-11, 2 yrs to lt 3 yrs	0.032	0.035	1.097	0.98	0.95
18 - y5-11, 3 yrs or more	0.128	0.198	1.541	1.49	1.44
19 - y11-16, lt 3 months	0.012	0.004	0.291	0.31	0.33
20 - y11-16, 3-lt 6months	0.014	0.014	0.977	1.01	1.05
21 - y11-16, 6 months-12 months	0.013	0.017	1.292	1.03	1.07
22 - y11-16, 12 months - 24 months	0.012	0.021	1.737	1.48	1.53
23 - y11-16, 2 yrs to 3yrs	0.011	0.016	1.485	1.24	1.21
24 - y11-16, 3 yrs or more	0.070	0.110	1.573	1.47	1.40

Table J.20 Impact estimates from matching adjusted for non-response

		Weighted (1)	(2) Single nearest neighbour matching with replacement (sample of 40,698 postal survey respondents with estimated propensity scores < 0.3)	(3) Single nearest neighbour matching with replacement (sample of 40,964 postal survey respondents)	Non-Weighted (4) Single nearest neighbour matching with replacement (sample of 40,698 postal survey respondents with estimated propensity scores < 0.3)
Number of Participants		1,777	1,656	1,777	1,656
WLEFTISB	Whether left IS at any time since participation	0.371 (0.015)	0.354 (0.014)	0.373 (0.015)	0.382 (0.015)
WLEFTISC	Whether left IS at any time since sample date	0.232 (0.016)	0.219 (0.015)	0.233 (0.017)	0.236 (0.017)
PART180B	Whether on IS 6 months after participation	0.09 (0.017)	0.106 (0.016)	0.084 (0.017)	0.106 (0.018)
PART180C	Whether on IS 6 months after sample date	0.022 (0.01)	0.022 (0.009)	0.023 (0.010)	0.024 (0.01)
TONISA	Duration of time on IS since participation measured in days	41.699 (4.32)	44.427 (4.136)	40.008 (4.691)	45.7 (4.772)
TONISC	Duration of time on IS since sample date measured in days	-31.897 (4.109)	-29.522 (3.924)	-32.340 (4.257)	-32.344 (4.329)

Table J.21 Impact estimates for different response windows: full postal survey respondent sample

Number of days		Single nearest neighbour matching with replacement					
		Full window	-6 to 20	21 to 49	50 to 83	84 to 119	120 to 197
Number of participants		1,777	365	352	354	358	348
WLEFTISB	Whether left IS at any time since participation						
	Effect	0.373	0.318	0.372	0.452	0.388	0.336
	Std Error	(0.015)	(0.034)	(0.034)	(0.033)	(0.033)	(0.033)
WLEFTISC	Whether left IS at any time since sample date						
	Effect	0.233	0.205	0.250	0.280	0.229	0.201
	Std Error	(0.017)	(0.037)	(0.037)	(0.036)	(0.036)	(0.037)
PART180B	Whether on IS 6 months after participation						
	Effect	0.084	0.036	0.040	0.099	0.140	0.106
	Std Error	(0.017)	(0.037)	(0.038)	(0.038)	(0.037)	(0.038)
PART180C	Whether on IS 6 months after sample date						
	Effect	0.023	-0.088	-0.020	0.045	0.081	0.098
	Std Error	(0.010)	(0.028)	(0.023)	(0.022)	(0.019)	(0.017)
TONISA	Duration of time on IS since participation measured in days						
	Effect	40.008	34.890	31.193	42.607	49.589	41.793
	Std Error	(4.691)	(11.417)	(10.669)	(9.426)	(9.193)	(8.245)
TONISC	Duration of time on IS since sample date measured in days						
	Effect	-32.340	-42.674	-47.349	-43.144	-24.455	-3.443
	Std Error	4.257	10.127	9.495	8.781	8.744	8.460

Table J.22 Impact estimates for different response windows: full face-to-face interview respondent sample

Number of days		Single nearest neighbour matching with replacement					
		Full window	-6 to 20	21 to 49	50 to 83	84 to 119	120 to 197
Number of participants		1,259	254	255	250	258	242
WLEFTISB	Whether left IS at any time since participation						
	Effect	0.301	0.268	0.336	0.326	0.292	0.273
	Std Error	(0.021)	(0.045)	(0.043)	(0.044)	(0.044)	(0.043)
WLEFTISC	Whether left IS at any time since sample date						
	Effect	0.298	0.276	0.32	0.317	0.284	0.273
	Std Error	(0.021)	(0.045)	(0.043)	(0.045)	(0.045)	(0.044)
PART180B	Whether on IS 6 months after participation						
	Effect	-0.252	-0.247	-0.299	-0.264	-0.233	-0.216
	Std Error	(0.019)	(0.041)	(0.038)	(0.041)	(0.041)	(0.039)
PART180C	Whether on IS 6 months after sample date						
	Effect	-0.063	-0.18	-0.116	-0.031	0.0	0.013
	Std Error	(0.01)	(0.032)	(0.025)	(0.018)	(0.017)	(0.011)
TONISA	Duration of time on IS since participation measured in days						
	Effect	-63.856	-61.423	-69.241	-70.317	-57.814	-62.97
	Std Error	(4.732)	(11.716)	(10.063)	(9.112)	(9.049)	(8.109)
TONISC	Duration of time on IS since sample date measured in days						
	Effect	-62.255	-80.845	-80.929	-67.062	-46.284	-33.273
	Std Error	(4.889)	(11.796)	(10.5)	(9.746)	(9.204)	(8.472)
JOBNOW	In full-time or part-time work as of the face-to-face interview						
	Effect	0.252	0.226	0.278	0.282	0.225	0.264
	Std Error	(0.021)	(0.044)	(0.044)	(0.045)	(0.044)	(0.044)
JOBFTNOW	Currently in full-time work as of the face-to-face interview						
	Effect	0.246	0.255	0.278	0.251	0.263	0.195
	Std Error	(0.02)	(0.042)	(0.04)	(0.043)	(0.04)	(0.041)
ANYJBAD	Had some job during or after participation date						
	Effect	0.232	0.23	0.303	0.273	0.186	0.195
	Std Error	(0.02)	(0.043)	(0.042)	(0.043)	(0.041)	(0.042)
WORKAT6	Working 6 months after participation						
	Effect	0.246	0.238	0.247	0.27	0.228	0.254
	Std Error	(0.02)	(0.041)	(0.043)	(0.043)	(0.042)	(0.042)

**Table J.23 Impact estimates for different response windows:
postal survey respondent sample with estimated
propensity scores < 0.3**

		Single nearest neighbour matching with replacement					
Number of days		Full window	-6 to 20	21 to 49	50 to 83	84 to 119	120 to 197
Number of participants		1,656	338	328	334	331	325
WLEFTISB	Whether left IS at any time since participation						
	Effect	0.382	0.349	0.378	0.464	0.381	0.335
	Std Error	(0.015)	(0.035)	(0.034)	(0.033)	(0.033)	(0.034)
WLEFTISC	Whether left IS at any time since sample date						
	Effect	0.236	0.222	0.262	0.275	0.221	0.2
	Std Error	(0.017)	(0.038)	(0.038)	(0.037)	(0.038)	(0.038)
PART180B	Whether on IS 6 months after participation						
	Effect	0.106	0.056	0.076	0.108	0.154	0.138
	Std Error	(0.018)	(0.039)	(0.039)	(0.039)	(0.039)	(0.039)
PART180C	Whether on IS 6 months after sample date						
	Effect	0.024	-0.086	-0.015	0.048	0.082	0.092
	Std Error	(0.01)	(0.028)	(0.023)	(0.022)	(0.019)	(0.018)
TONISA	Duration of time on IS since participation measured in days						
	Effect	45.7	39.784	41.524	44.18	55.103	48.052
	Std Error	(4.772)	(11.758)	(11)	(9.762)	(9.552)	(8.484)
TONISC	Duration of time on IS since sample date measured in days						
	Effect	-32.344	-46.938	-42.122	-45.964	-22.755	-3.065
	Std Error	(4.329)	(10.348)	(9.66)	(9.163)	(8.968)	(8.642)

Table J.24 Impact estimates for different response windows: face-to-face interview respondent sample with estimated propensity scores < 0.3

		Single nearest neighbour matching with replacement					
Number of days		Full window	-6 to 20	21 to 49	50 to 83	84 to 119	120 to 197
Number of participants		1,171	238	231	242	230	230
WLEFTISB	Whether left IS at any time since participation						
	Effect	0.297	0.23	0.29	0.267	0.202	0.225
	Std Error	(0.022)	(0.048)	(0.045)	(0.046)	(0.047)	(0.045)
WLEFTISC	Whether left IS at any time since sample date						
	Effect	0.302	0.235	0.281	0.267	0.206	0.234
	Std Error	(0.022)	(0.049)	(0.045)	(0.046)	(0.048)	(0.046)
PART180B	Whether on IS 6 months after participation						
	Effect	-0.256	-0.23	-0.253	-0.195	-0.197	-0.165
	Std Error	(0.019)	(0.045)	(0.041)	(0.043)	(0.044)	(0.04)
PART180C	Whether on IS 6 months after sample date						
	Effect	-0.068	-0.174	-0.104	-0.045	0.018	0.014
	Std Error	(0.01)	(0.033)	(0.025)	(0.018)	(0.017)	(0.011)
TONISA	Duration of time on IS since participation measured in days						
	Effect	-64.055	-51.446	-55.869	-54.778	-46.725	-55.147
	Std Error	(4.756)	(12.245)	(10.651)	(9.319)	(9.647)	(8.176)
TONISC	Duration of time on IS since sample date measured in days						
	Effect	-63.413	-75.662	-69.24	-55.086	-35.372	-26.596
	Std Error	(4.903)	(12.281)	(10.882)	(9.673)	(9.742)	(8.582)
JOBNOW	In full-time or part-time work as of the face-to-face interview						
	Effect	0.248	0.188	0.244	0.226	0.183	0.229
	Std Error	(0.022)	(0.047)	(0.046)	(0.046)	(0.046)	(0.045)
JOBFTNOW	Currently in full-time work as of the face-to-face interview						
	Effect	0.243	0.216	0.231	0.213	0.216	0.151
	Std Error	(0.02)	(0.045)	(0.043)	(0.044)	(0.043)	(0.042)
ANYJBAD	Had some job during or after participation date						
	Effect	0.233	0.202	0.285	0.222	0.119	0.142
	Std Error	(0.021)	(0.046)	(0.044)	(0.045)	(0.044)	(0.044)
WORKAT6	Working 6 months after participation						
	Effect	0.244	0.196	0.208	0.203	0.219	0.207
	Std Error	(0.021)	(0.045)	(0.045)	(0.045)	(0.044)	(0.044)

Table J.25 Outcome definitions

Variable name	Variable label	Value code	Value label
Variables from the administrative data:			
WLEFTISB	whether left IS at any time since participation	0	No
		1	Yes
WLEFTISC	whether left IS at any time since sample date	0	No
		1	Yes
PART180B	whether on IS 6 months after participation	0	No
		1	Yes
PART180C	whether on IS 6 months after sample date	0	No
		1	Yes
TONISA	duration of time on IS since participation measured in days	0	No
		1	Yes
TONISC	duration of time on IS since sample date measured in days	0	No
		1	Yes

Table J.26 Impacts from alternative single nearest neighbour matching with replacement using LAD variables

Original	Unemployment ilo	Deprivation 2001					
		Full sample (39,907 dataset)	Face-to-face interview respondents (2,500 dataset)	NatCen matched sample (3,596 dataset)	Full sample (39,907 dataset)	Face-to-face interview respondents (2,500 dataset)	NatCen matched sample (3,596 dataset)
WLEFTISB	Whether left IS at any time since participation	0.51 (0.01)	0.28 (0.02)	0.27 (0.03)	0.50 (0.01)	0.29 (0.02)	0.25 (0.03)
WLEFTISC	Whether left IS at any time since sample date	0.25 (0.02)	0.28 (0.02)	0.27 (0.03)	0.26 (0.02)	0.29 (0.02)	0.26 (0.03)
PART180B	Whether on IS 6 months after participation	0.50 (0.01)	-0.25 (0.02)	-0.23 (0.02)	0.49 (0.01)	-0.25 (0.02)	-0.22 (0.02)
PART180C	Whether on IS 6 months after sample date	0.03 (0.01)	-0.05 (0.01)	-0.06 (0.01)	0.01 (0.01)	-0.06 (0.01)	-0.06 (0.01)
TONISA	Duration of time on IS since participation measured in days	168.54 (3.65)	-52.47 (4.69)	-52.74 (5.46)	164.40 (3.74)	-52.92 (4.69)	-52.63 (5.75)
TONISC	Duration of time on IS since sample date measured in days	-35.29 (4.35)	-51.24 (4.91)	-56.02 (5.51)	-42.24 (4.25)	-51.96 (4.97)	-51.53 (5.99)

Notes:

- (1) The approximate variance calculated by psmatch for nearest neighbour matching assumes independent observations, fixed weights, homoskedasticity of the outcome variable within the treated and within the control groups and that the variance of the outcome does not depend on the propensity score.
- (2) Standard error in parentheses

Table J.27 Total average treatment effect on the treated according to different samples

Sample	Description	Cross-section estimate	Pre-programme difference	Difference estimate
I	A, B, C, D and E	24.06 (0.15)	10.15 (0.09)	13.91 (0.52)
II	B and C	26.62 (0.17)	10.78 (0.08)	15.84 (0.58)
III	B, C1 and C2	20.52 (0.24)	4.40 (0.16)	16.12 (1.41)
IV	III with $ps < .3$	21.43 (0.24)	4.61 (0.19)	16.81 (1.65)
V	B1 and C1	20.84 (0.23)	5.38 (0.20)	15.45 (1.58)
VI	V with $ps < .3$	21.50 (0.22)	6.04 (0.24)	15.45 (1.81)
VII	III (no switcher)	19.25 (0.17)	5.27 (0.16)	13.98 (0.96)
VIII	III (no switcher and matched counterpart)	18.61 (0.16)	5.38 (0.14)	13.23 (0.78)
IX	III (no switcher and matched counterpart) with $ps < .3$	20.26 (0.17)	5.71 (0.16)	14.55 (0.97)
X	V (no switcher and matched counterpart)	18.58 (0.13)	6.73 (0.20)	11.84 (0.99)
XI	V (no switcher and matched counterpart) with $ps < .3$	19.97 (0.14)	7.36 (0.21)	12.62 (1.11)

Note: Std. Errors in parentheses.

Table J.28 Card and Sullivan cell match (A, B, C, D and E)

PREHIST	Comparison		Treatment		TT	WEIGHT	CONTR
	N	RATE (%)	N	RATE (%)			
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	0.10	-1.03
000000	2,502	66.2	225	47.4	-18.8	0.05	-1.00
000011	3,658	65.3	349	55.8	-9.6	0.08	-0.78
011111	2,598	72.5	196	55.6	-16.9	0.05	-0.78
000111	2,651	69.1	206	57.1	-12.0	0.05	-0.58
001111	2,198	70.0	165	57.2	-12.8	0.04	-0.50
100001	367	61.2	41	49.0	-12.2	0.01	-0.12
101111	330	76.1	28	58.5	-17.6	0.01	-0.12
110111	329	79.2	26	64.6	-14.6	0.01	-0.09
111011	382	73.8	28	60.5	-13.4	0.01	-0.09
111100	210	71.8	20	54.4	-17.4	0.00	-0.08
111101	354	72.6	21	56.3	-16.3	0.00	-0.08
110001	328	60.4	21	45.6	-14.9	0.00	-0.07
110011	479	68.7	36	61.1	-7.6	0.01	-0.06
Others	2,646		193				-0.32
Total	65,594		4,235				-17.61

Table J.29 Card and Sullivan cell match (B and C)

PREHIST	Comparison		Treatment		TT	WEIGHT	CONTR
	N	RATE (%)	N	RATE (%)			
111111	25,858	83.2	989	60.2	-23.0	0.56	-12.82
000001	2,434	62.2	143	48.4	-13.8	0.08	-1.11
000000	1,536	67.3	86	46.4	-20.8	0.05	-1.01
000011	2,147	67.8	151	58.4	-9.5	0.08	-0.80
011111	1,516	73.9	76	56.3	-17.5	0.04	-0.75
000111	1,553	70.9	110	59.6	-11.3	0.06	-0.70
001111	1,266	72.5	66	63.6	-8.9	0.04	-0.33
110111	181	81.9	11	54.5	-27.4	0.01	-0.17
111101	191	71.8	9	40.4	-31.4	0.01	-0.16
101111	187	76.3	16	59.7	-16.6	0.01	-0.15
100001	198	61.2	12	41.7	-19.6	0.01	-0.13
010011	26	72.8	4	29.7	-43.1	0.00	-0.10
111001	186	69.0	6	46.5	-22.5	0.00	-0.08
100111	182	75.9	11	64.6	-11.4	0.01	-0.07
001101	43	73.4	3	41.5	-31.9	0.00	-0.05
Others	1,761		84				-0.23
Total	39,265		1,777				-18.66

Table J.30 Card and Sullivan cell match except 111111 (A, B, C, D and E)

PREHIST	Comparison		Treatment		TT	WEIGHT	CONTR
	N	RATE (%)	N	RATE (%)			
000001	4,154	62.2	404	51.3	-10.9	0.21	-2.24
000000	2,502	66.2	225	47.4	-18.8	0.11	-2.16
000011	3,658	65.3	349	55.8	-9.6	0.18	-1.70
011111	2,598	72.5	196	55.6	-16.9	0.10	-1.69
000111	2,651	69.1	206	57.1	-12.0	0.11	-1.26
001111	2,198	70.0	165	57.2	-12.8	0.08	-1.08
100001	367	61.2	41	49.0	-12.2	0.02	-0.26
101111	330	76.1	28	58.5	-17.6	0.01	-0.25
110111	329	79.2	26	64.6	-14.6	0.01	-0.19
111011	382	73.8	28	60.5	-13.4	0.01	-0.19
111100	210	71.8	20	54.4	-17.4	0.01	-0.18
111101	354	72.6	21	56.3	-16.3	0.01	-0.17
110001	328	60.4	21	45.6	-14.9	0.01	-0.16
110011	479	68.7	36	61.1	-7.6	0.02	-0.14
111001	397	68.4	23	57.9	-10.5	0.01	-0.12
Others	2,249		170				-0.54
Total	23,186		1,959				-12.33

Table J.31 Card and Sullivan cell match except 111111 (B and C)

PREHIST	Comparison		Treatment		TT	WEIGHT	CONTR
	N	RATE (%)	N	RATE (%)			
000001	2,434	62.2	143	48.4	-13.8	0.18	-2.50
000000	1,536	67.3	86	46.4	-20.82	0.11	-2.27
000011	2,147	67.8	151	58.4	-9.45	0.19	-1.81
011111	1,516	73.9	76	56.3	-17.53	0.10	-1.69
000111	1,553	70.9	110	59.6	-11.28	0.14	-1.58
001111	1,266	72.5	66	63.6	-8.85	0.08	-0.74
110111	181	81.9	11	54.5	-27.41	0.01	-0.38
111101	191	71.8	9	40.4	-31.37	0.01	-0.36
101111	187	76.3	16	59.7	-16.63	0.02	-0.34
100001	198	61.2	12	41.7	-19.55	0.02	-0.30
010011	26	72.8	4	29.7	-43.09	0.01	-0.22
111001	186	69.0	6	46.5	-22.49	0.01	-0.17
100111	182	75.9	11	64.6	-11.38	0.01	-0.16
001101	43	73.4	3	41.5	-31.89	0.00	-0.12
001001	37	64.8	3	39.8	-25.01	0.00	-0.10
Other	1,724		81				-0.38
Total	13,407		788				-13.12

Table J.32 P-value of the joint significance of selected categories on the propensity score estimation

	(1)	(2)	(3)
Observations	65,887	65,887	65,731
Age	0.02	0.03	0.02
Region	0.00	0.00	0.00
Age of the youngest child		0.00	0.00
Number of children		0.00	0.00
Card and Sullivan regime dummies			0.43
Recent			0.00

A table with all coefficients can be found in Appendix ??.

Table J.33 Total average treatment effect

Sample	Description	Cross-section estimate	Pre-programme difference	Difference estimate
I	A, B, C, D and E	23.05 (0.15)	5.05 (0.08)	18.00 (0.47)

Table J.34 Proportion of lone parents by age of the youngest child and length of time on IS (%)

Banded age of the youngest child	Banded duration of claim						Total
	Under 3 months	3 months	6 months	12 months	2yrs to u	3yrs and	
Under 3 years	0.80	2.68	4.94	7.75	4.91	9.15	30.24
3 to under 5 years	0.34	0.9	1.41	2.53	2.12	9.71	17.01
5 to under 11 years	0.62	1.7	2.46	4.25	3.18	20.92	33.14
11 to under 16	0.35	0.97	1.38	2.26	1.67	12.99	19.61
Total	2.12	6.25	10.2	16.78	11.88	52.77	100

Source: DWP administrative data (August 2000).

Table J.35 Subgroups total average treatment effect

Sample	Hardmatching by Stata	Description	Treatment Hetero		Diff
I – Flow	Yes	Children at the age of 0 and 3	16.73 (0.31)	-3.05 (0.18)	19.78 (2.21)
I – Flow	Yes	Children at the age of 3 and 5	17.71 (0.26)	-1.40 (0.14)	19.11 (1.56)
I – Flow	Yes	Children at the age of 5 and 11	14.31 (0.18)	-1.66 (0.10)	15.97 (0.77)
I – Flow	Yes	Children at the age of 11 and 16	9.92 (0.33)	-2.22 (0.27)	12.14 (3.13)
I – Stock	Yes	Children at the age of 0 and 3	17.03 (0.28)	-0.37 (0.06)	17.39 (1.47)
I – Stock	Yes	Children at the age of 3 and 5	18.75 (0.16)	0.32 (0.09)	18.42 (0.57)
I – Stock	Yes	Children at the age of 5 and 11	21.26 (0.18)	0.20 (0.07)	21.06 (0.62)
I – Stock	Yes	Children at the age of 11 and 16	23.67 (0.30)	-0.89 (0.14)	24.56 (1.87)
I	Yes	Children at the age of 0 and 3	16.96 (0.25)	-0.98 (0.06)	17.95 (1.15)
I	Yes	Children at the age of 3 and 5	18.58 (0.16)	0.11 (0.08)	18.47 (0.54)
I	Yes	Children at the age of 5 and 11	20.33 (0.17)	-0.04 (0.06)	20.37 (0.59)
I	Yes	Children at the age of 11 and 16	21.23 (0.25)	-1.13 (0.14)	22.37 (1.43)
I	Yes	On IS for less than 3 months	16.07 (0.19)	-0.54 (0.08)	16.62 (0.72)
I	Yes	On IS from 3 to 6 months	13.17 (0.30)	-3.11 (0.21)	16.28 (2.30)
I	Yes	On IS from 6 to 12 months	21.05 (0.21)	-1.13 (0.20)	22.18 (1.42)
I	Yes	On IS from 12 to 24 months	15.87 (0.45)	0.26 (0.16)	15.61 (3.93)
I	Yes	On IS from 24 to 36 months	18.12 (0.13)	-0.45 (0.08)	18.57 (0.42)
I	Yes	On IS for more than 36 months	25.79 (0.21)	0.03 (0.01)	25.76 (0.80)

Table J.36 NDLP claims according study window and participant status

GroupsA	C & E	A, B &D	Total	C1 & C2	B	Total
Start and end before study window	0	76	76	0	0	0
Start and end during study window	0	1,084	1,084	0	401	401
Start during and end after study window	0	1,836	1,836	0	988	988
Start before and end after study window	0	235	235	0	0	0
Start and end after study window	9,378	1,580	10,958	487	748	1,235
Total	9,378	4,811	14,189	487	2,137	2,624

Table J.37 NDLP recidivism after 1 April 2001, by participation status

Number of NDLP spells	Non-participants	Sample I		Non-participants	Sample III	
		Participants	Total		Participants	Total
1	8,558 (85.75)	1,144 (78.09)	9,702 (84.77)	435 (86.83)	535 (78.68)	970 (82.13)
2	1,251 (12.54)	269 (18.36)	1,520 (13.28)	57 (11.38)	122 (17.94)	179 (15.16)
3	147 (1.47)	47 (3.21)	194 (1.7)	7 (1.4)	20 (2.94)	27 (2.29)
4	20 (0.2)	5 (0.34)	25 (0.22)	2 (0.4)	3 (0.44)	5 (0.42)
5	4 (0.04)	0 (0)	4 (0.03)			
Total	9,980 (100)	1,465 (100)	11,445 (100)	501 (100)	680 (100)	1,181 (100)

Percentages in parentheses.

Table J.38 NDLP recidivism after 23 October 1998, by participation status

Number of NDLP spells	Sample I			Sample III		
	Non-participants	Participants	Total	Non-participants	Participants	Total
1	8,558 (85.75)	2,925 (72.13)	11,483 (81.82)	435 (86.83)	1,258 (73.91)	1,693 (76.85)
2	1,251 (12.54)	904 (22.29)	2,155 (15.35)	57 (11.38)	363 (21.33)	420 (19.06)
3	147 (1.47)	190 (4.69)	337 (2.4)	7 (1.4)	67 (3.94)	74 (3.36)
4	20 (0.2)	35 (0.86)	55 (0.39)	2 (0.4)	14 (0.82)	16 (0.73)
5	4 (0.04)	1 (0.02)	5 (0.04)			
Total	9,980 (100)	4,055 (100)	14,035 (100)	501 (100)	1,702 (100)	2,203 (100)

Percentages in parentheses.

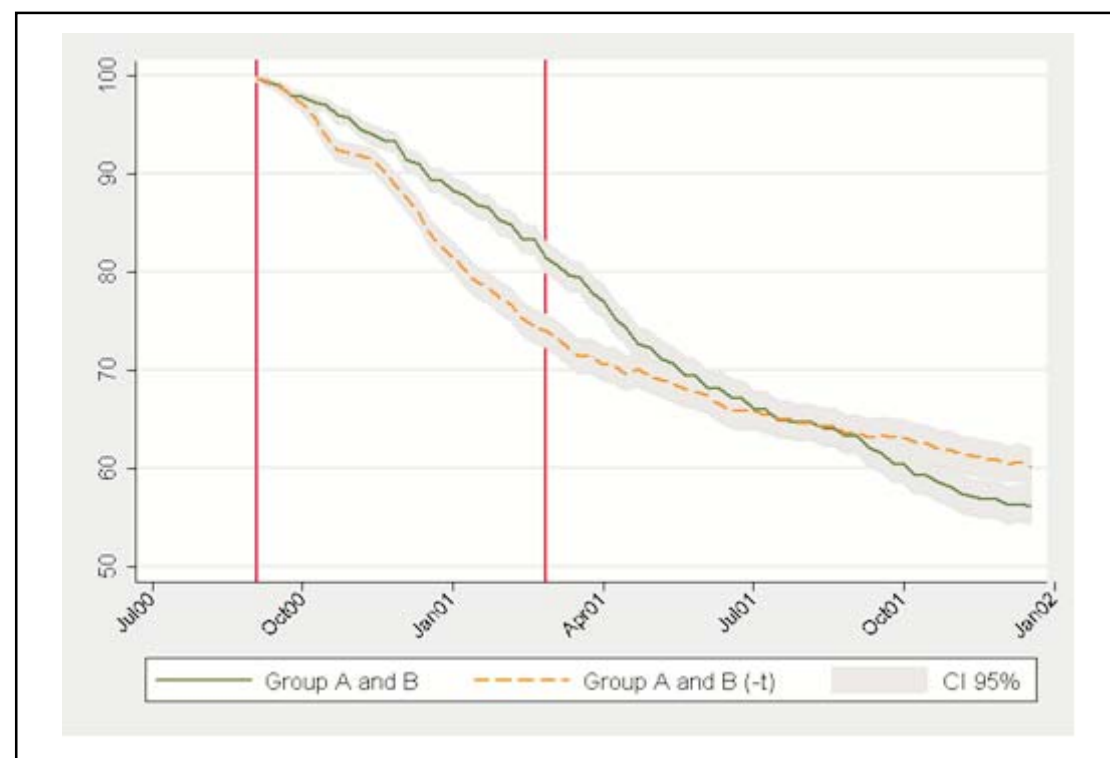
Figure J.6 Symmetric estimator of the weekly proportion of participants on benefit 1999-2002 (Sample II – postal)

Figure J.7 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample II – postal)

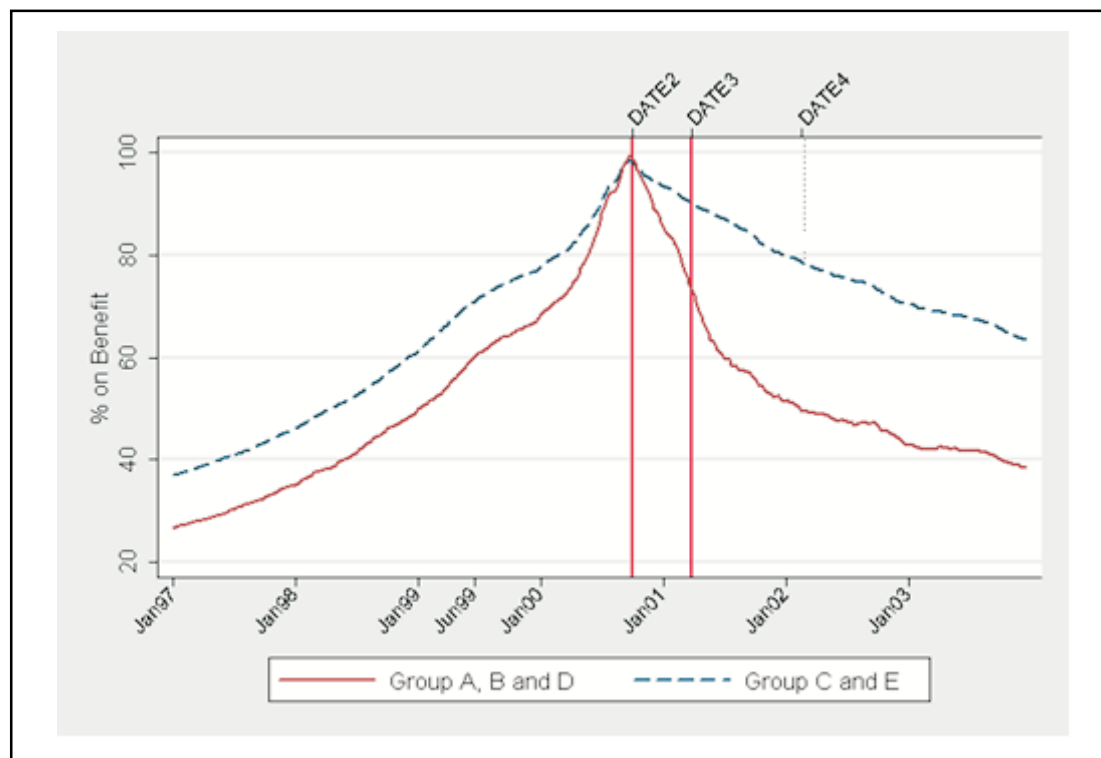


Figure J.8 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – full – weighted by Stata)

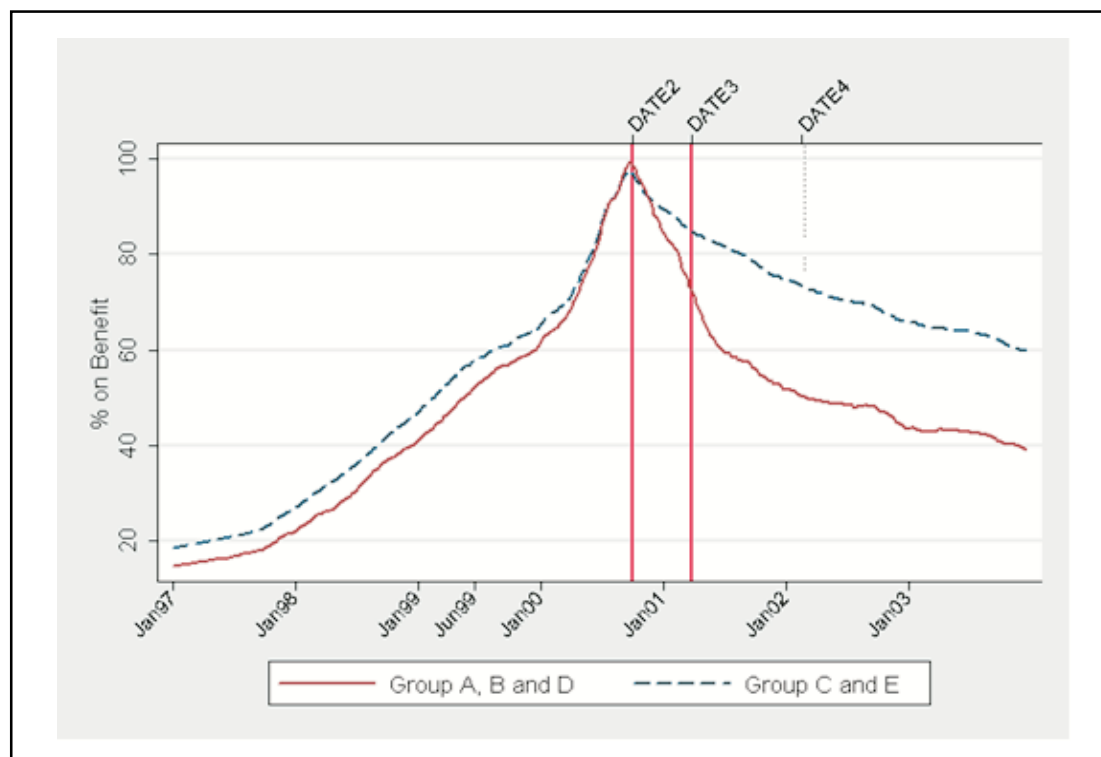


Figure J.9 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 1)

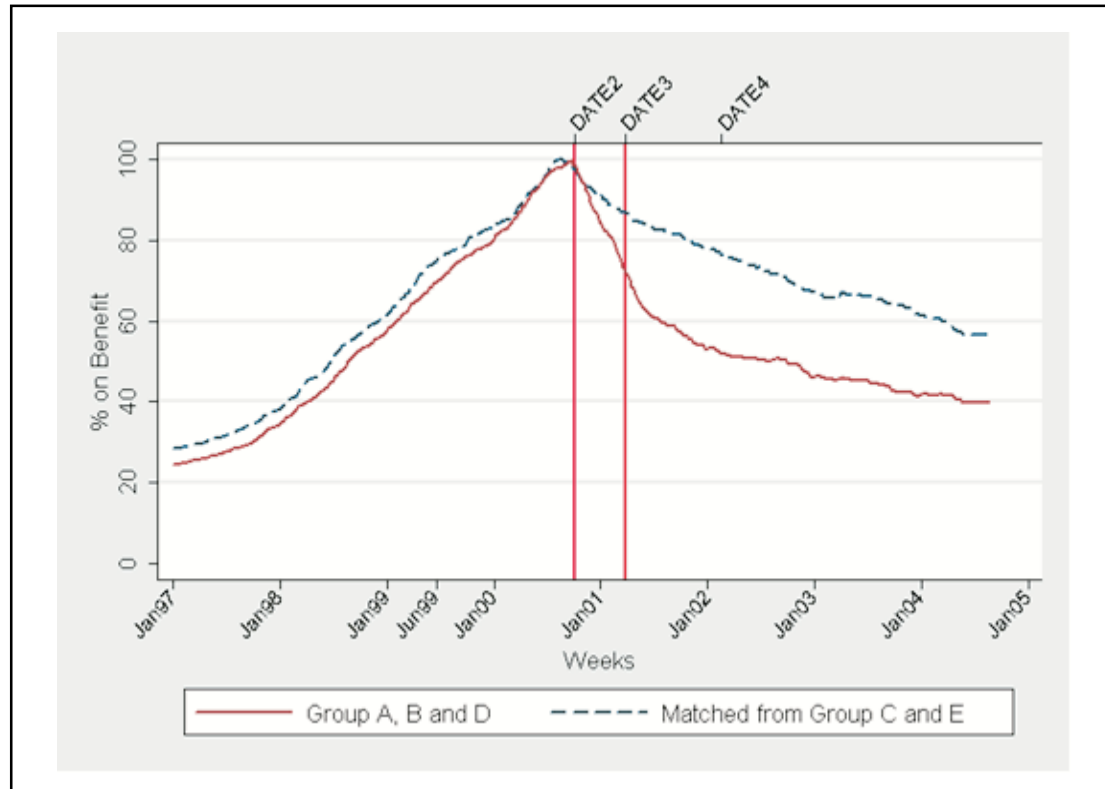


Figure J.10 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 2)

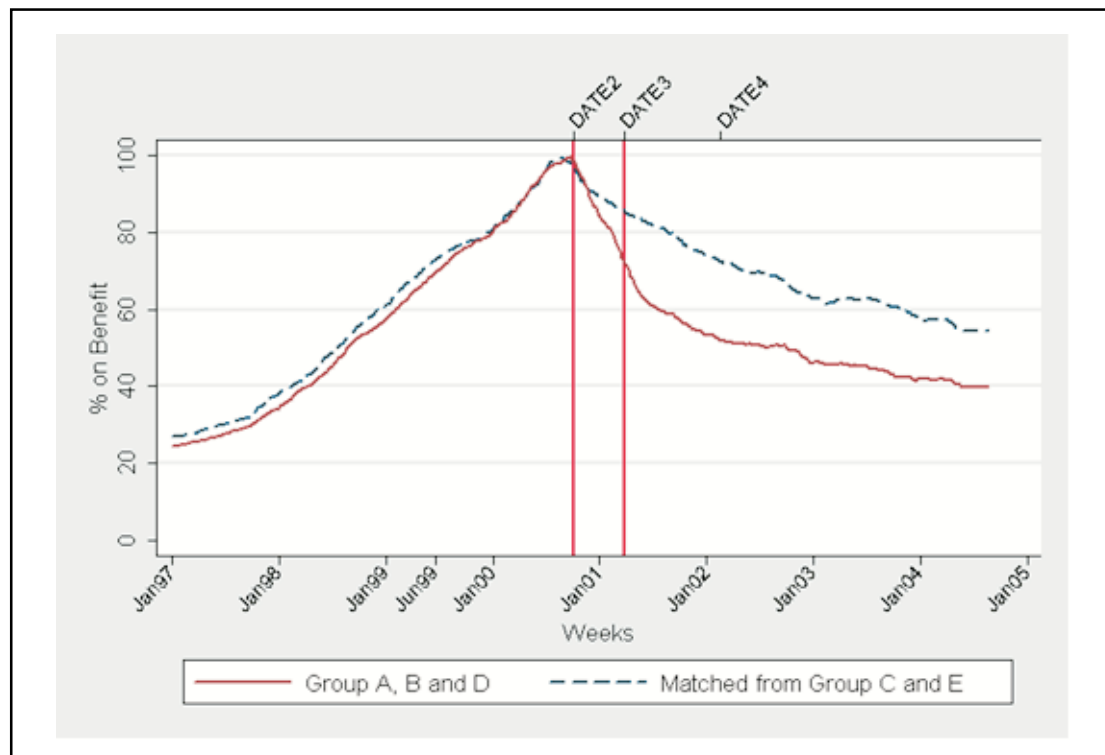


Figure J.11 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3 – Group B only)

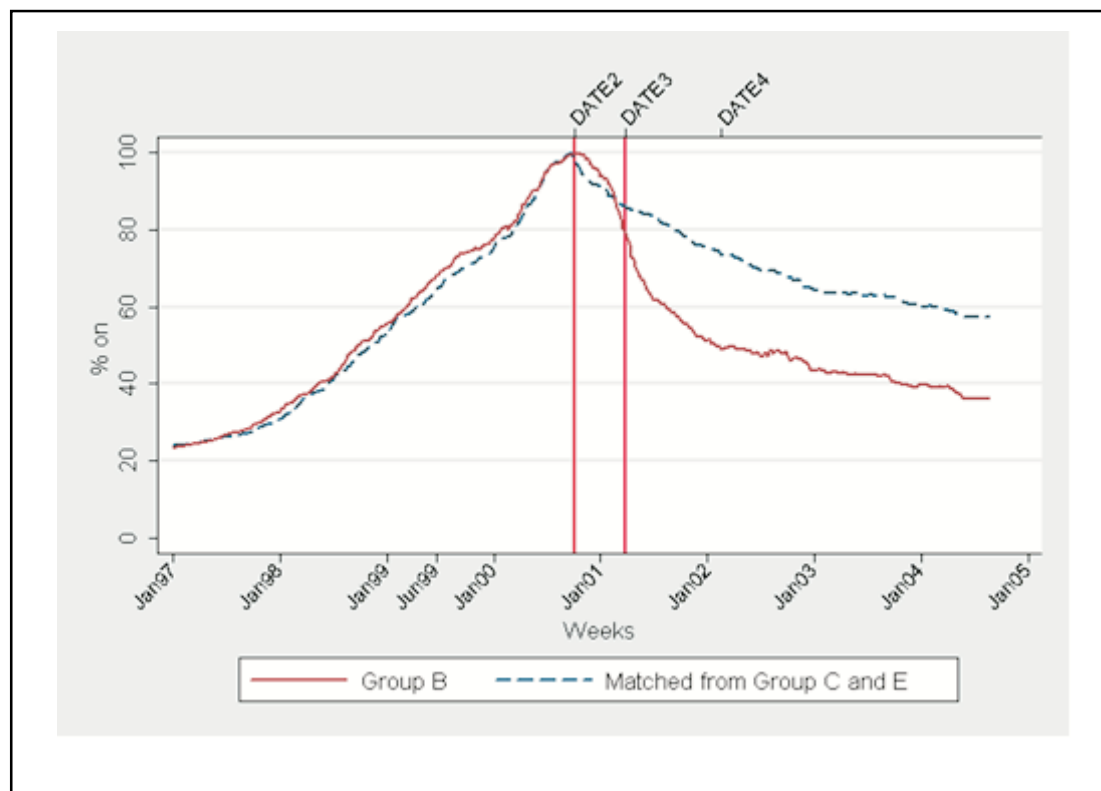


Figure J.12 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample II – NatCen specification + ps w/ replacement)

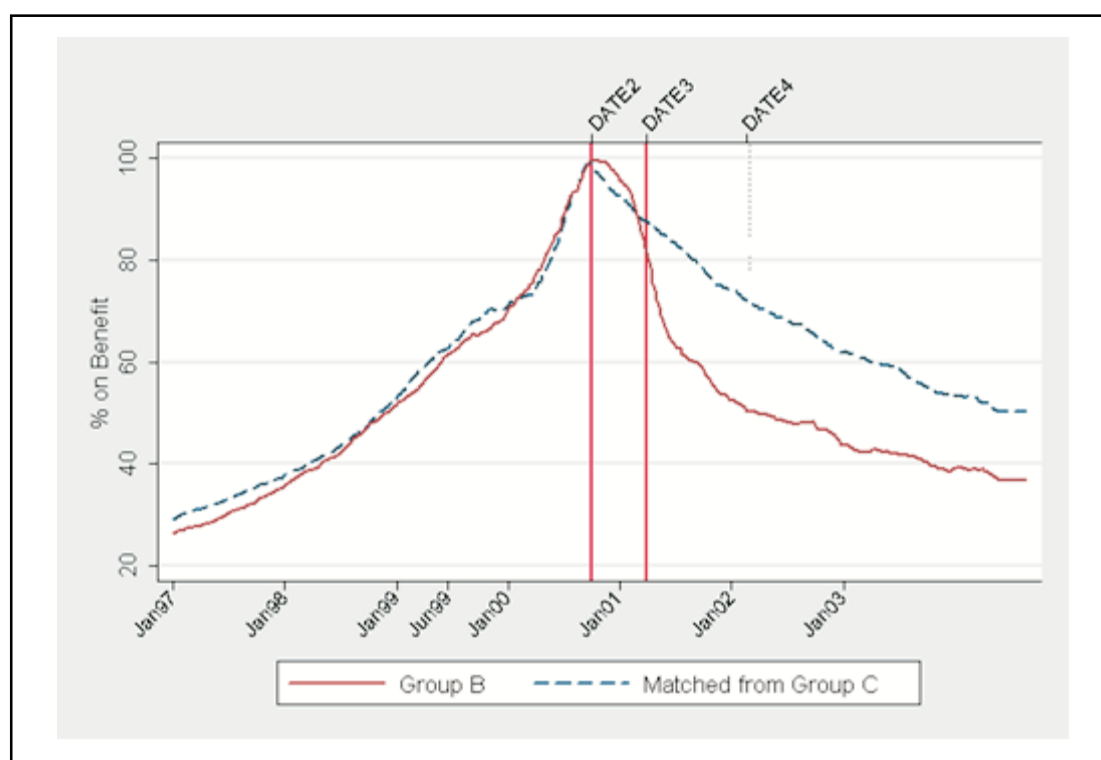


Figure J.13 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample II – NatCen specification + pre-programme + ps w/ replacement)

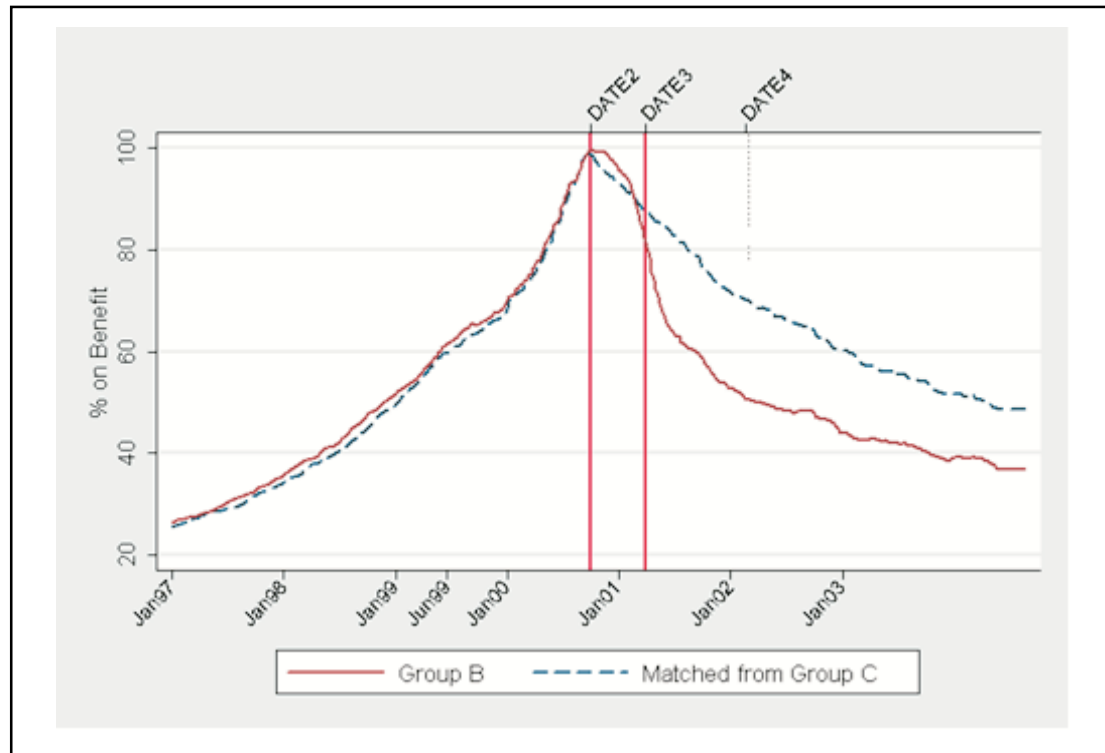


Figure J.14 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample V – face-to-face)

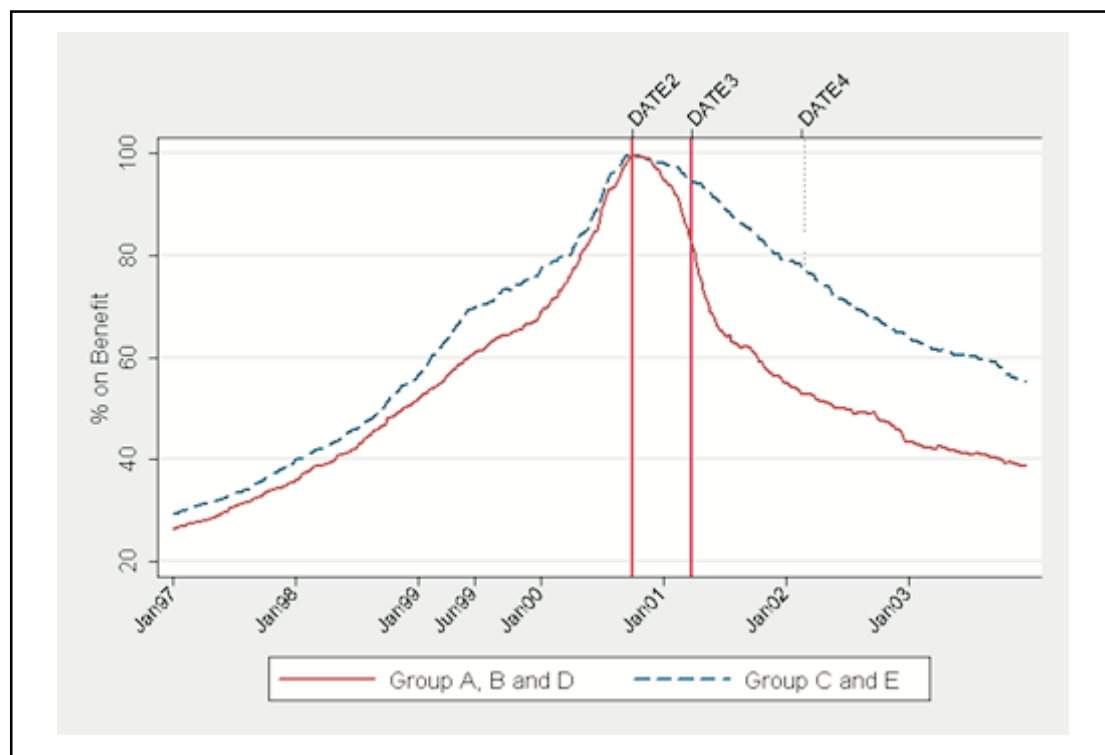


Figure J.15 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) child at between 3 and 5 years old

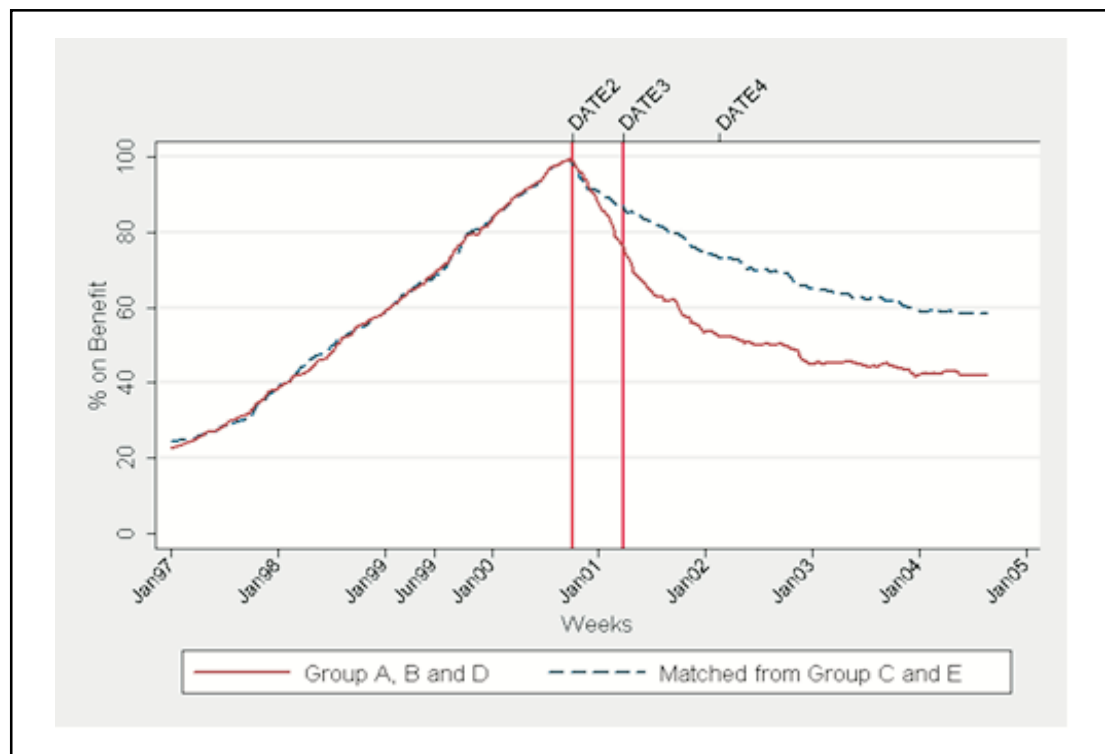


Figure J.16 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) child at between 5 and 11 years old

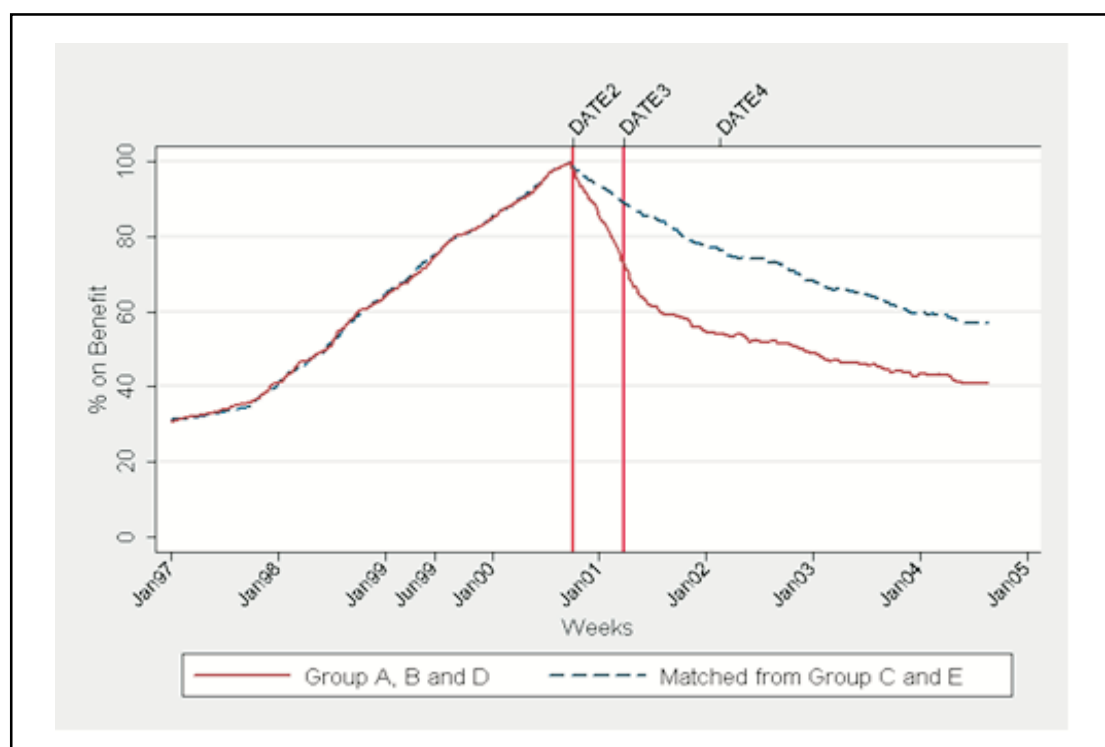


Figure J.17 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) less than 3 months on IS

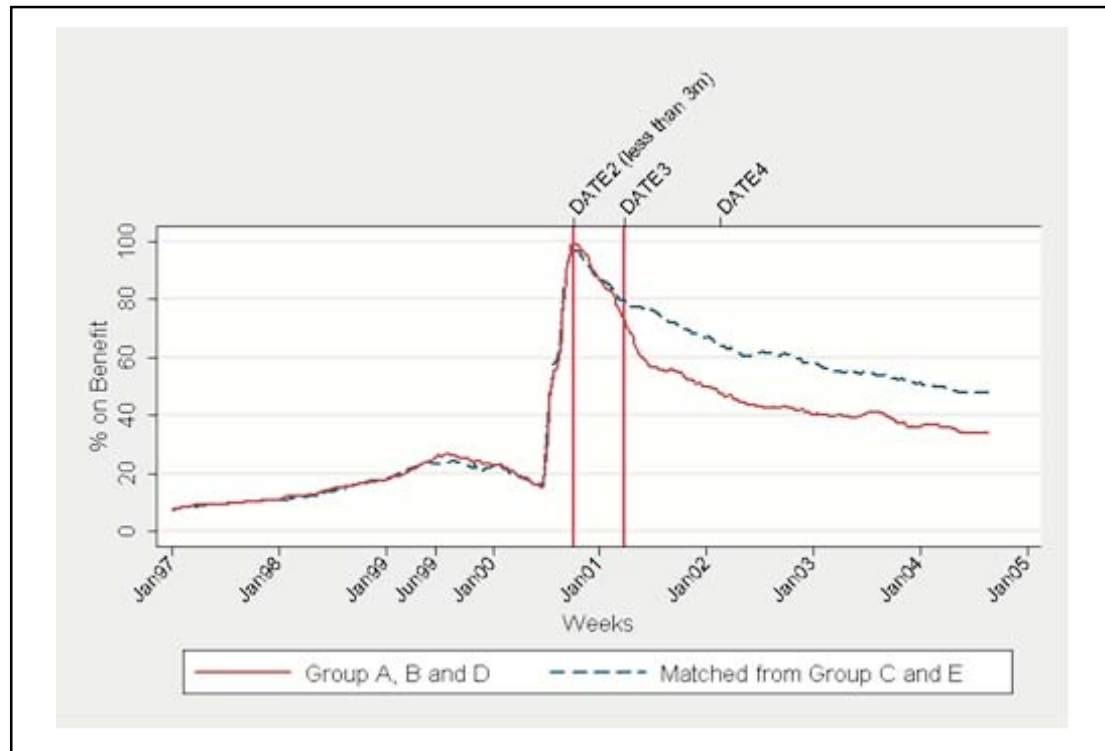


Figure J.18 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) from 3 to 6 months on IS

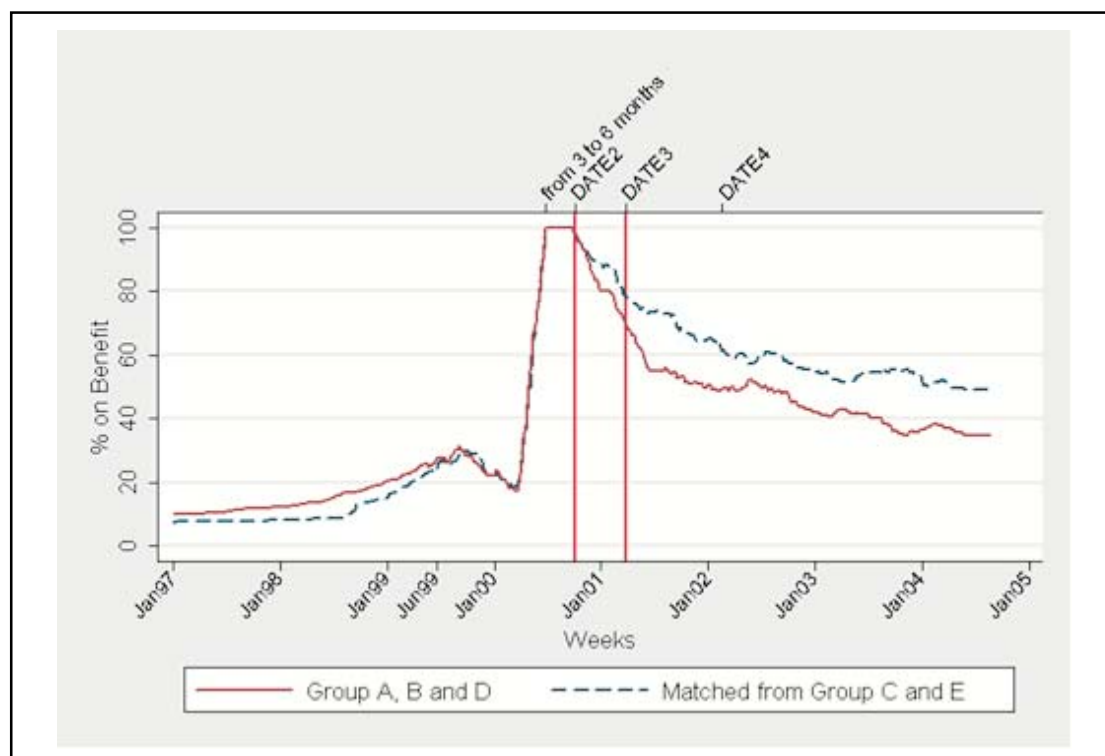


Figure J.19 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) from 6 to 12 months on IS

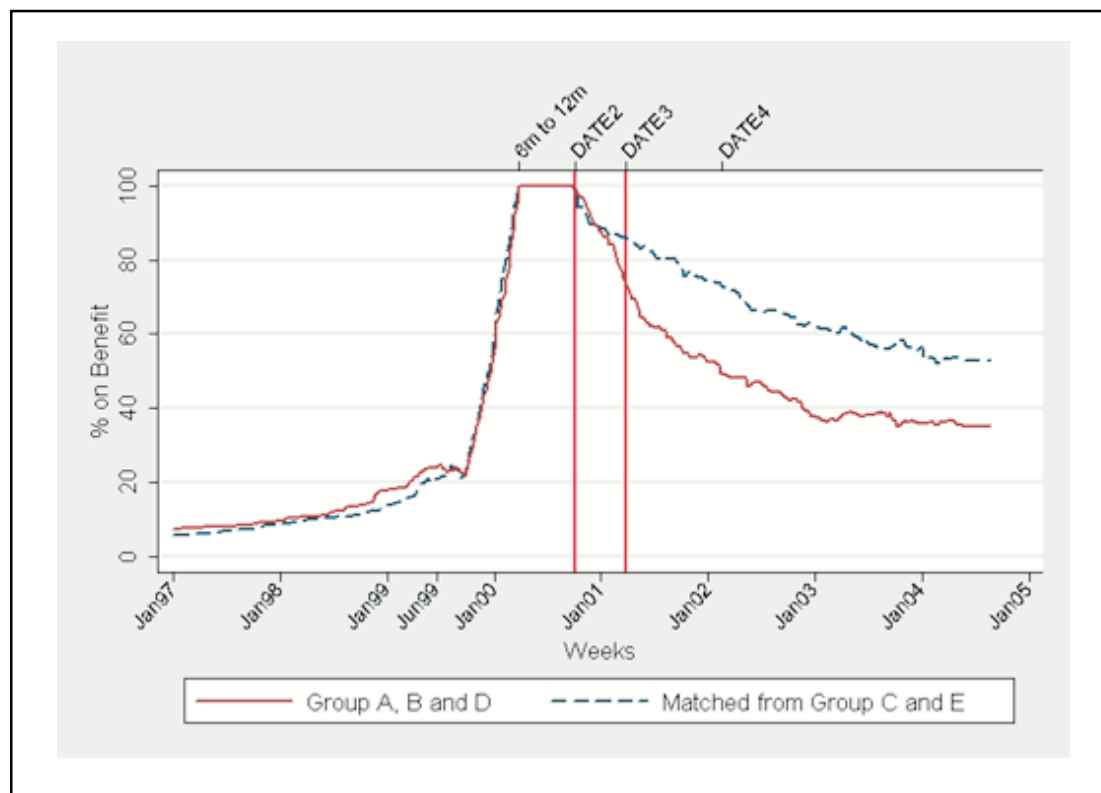


Figure J.20 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) from 12 to 24 months on IS

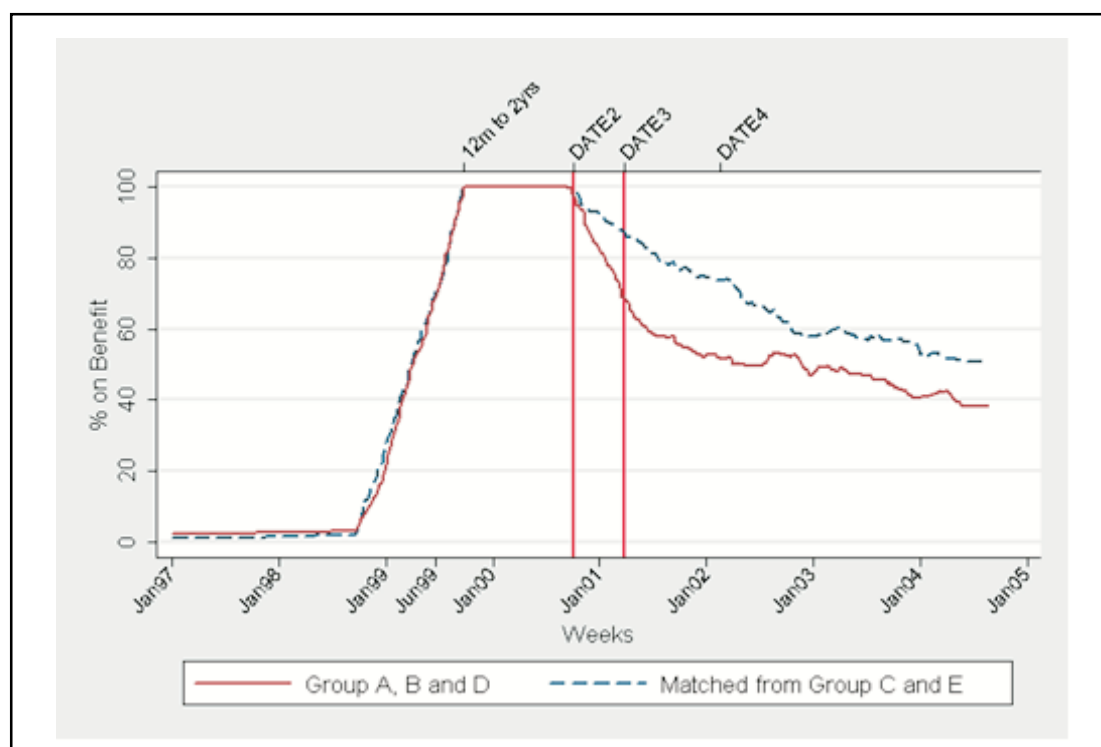


Figure J.21 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) from 24 to 36 months on IS

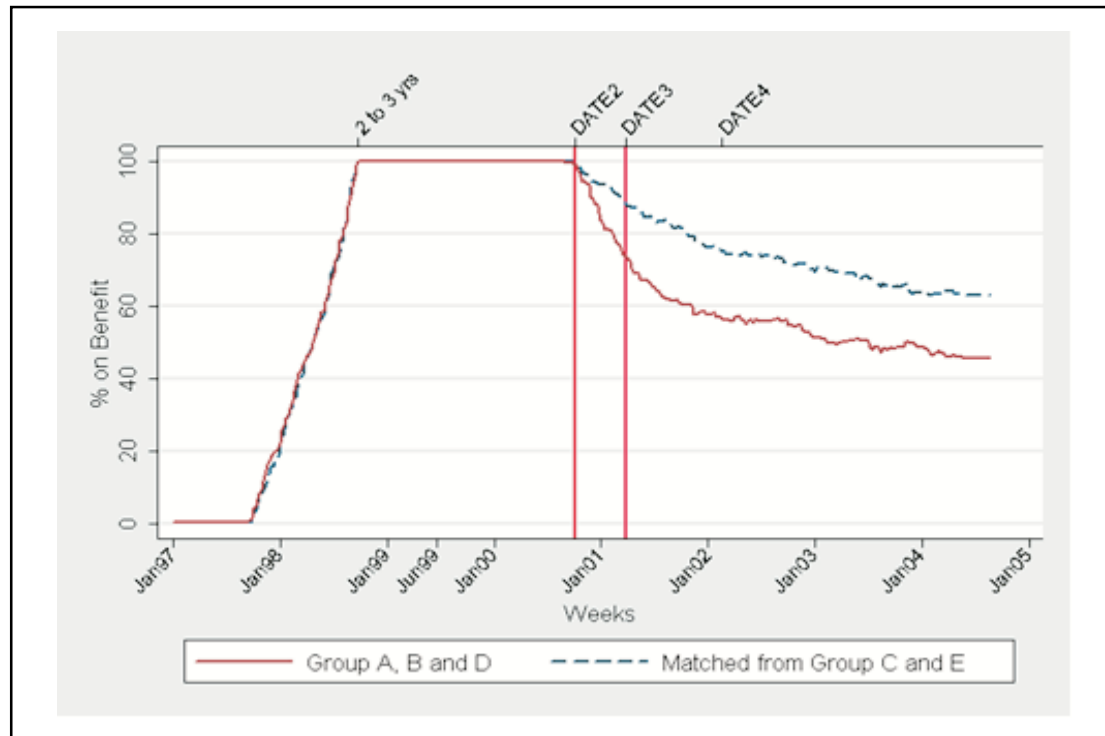
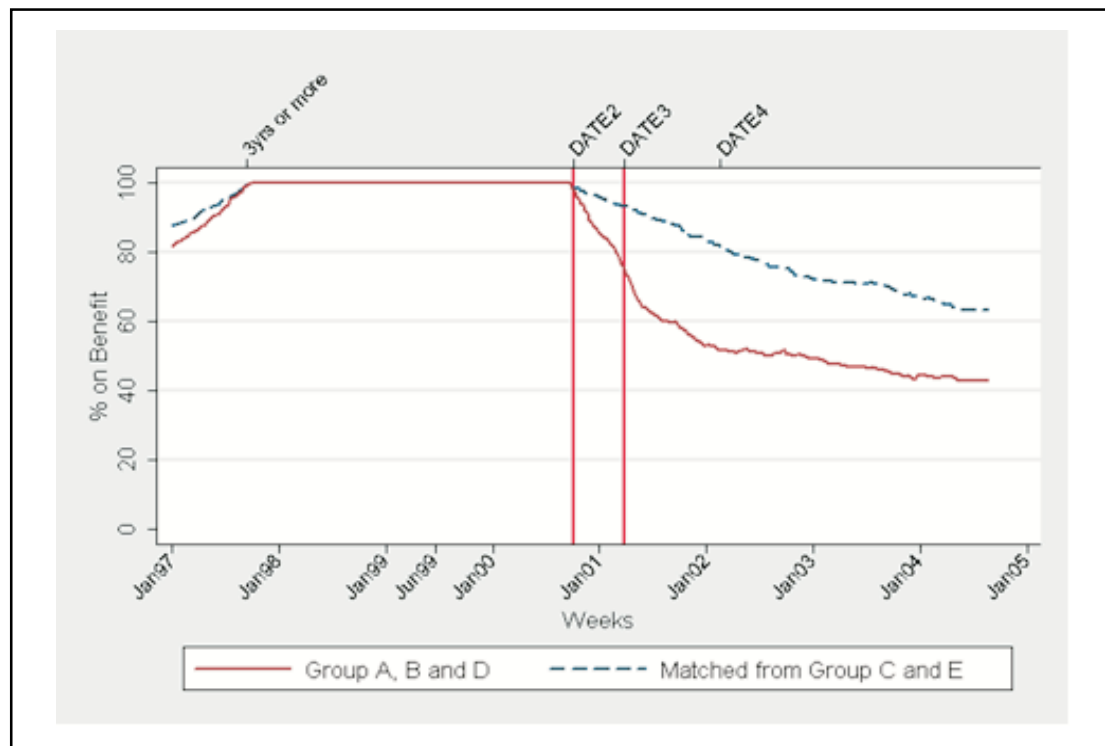


Figure J.22 Weekly proportion of participants and non-participants on benefit 1997-2003 (Sample I – Specification 3) for over 36 months on IS



References

- Abbring, J.H. and van den Berg, G.J. (2003a), The Non-parametric Identification of Treatment Effects in Duration Models. *Econometrica* 71(5): 1491-1517.
- Abbring, J.H. and van den Berg, G.J. (2003b) The identifiability of the mixed proportional hazards competing risks model, *Journal of the Royal Statistical Society Series B* vol 65, s 701-710
- Baj, J., and Trott, C.E. (1993). An Analysis of Repeating in JTPA in Illinois. Paper submitted to the Illinois Department of Commerce and Community Affairs.
- Barnow, B. (1987). The Impact of CETA Programmes on Earnings: A Review of the Literature. *Journal of Human Resources*. 22(2). 157-193.
- Becker, S. O. and Ichino, A. (2002). Estimation of Average Treatment Effects Based on Propensity Scores. *The Stata Journal* 22(4): 358-377.
- Black, D. and Smith, J. (2004). How Robust is the Evidence on the Effects of College Quality? Evidence from Matching. *Journal of Econometrics* 121(1): 99-124.
- Black, D., Smith, J., Berger, M. and Noel, B. (2003). Is the Threat of Reemployment Services More Effective than the Services Themselves? Evidence from Random Assignment in the UI System. *American Economic Review* 93(4): 1313-1327.
- Bloom, H., Orr, L., Bell, S., Cave, G., Doolittle, F. Lin, W. and Bos, J. (1997). The Benefits and Costs of JTPA Title II-A Programmes: Findings from the National Job Training Partnership Act Study. *Journal of Human Resources* 32(3): 549-576.
- Bryson, A., Dorsett, R. and Purdon, S. (2002). The Use of Propensity Score Matching in the Evaluation of Active Labour Market Policies. Department for Work and Pensions, Working Paper No.4.
- Calmfors, L. (1994). Active Labour Market Policy Unemployment - A Framework for the Analysis of Crucial Design Features. *OECD Economic Studies* 22(1): 7-47.

- Card, D. and Sullivan, D. G. (1988). Measuring the Effect of Subsidized Training Programmes on Movements In and Out of Employment. *Econometrica* 56(3): 497-530.
- Cochran, W. and Rubin, D.B. (1973). Controlling Bias in Observational Studies. *Sankhya* 35: 417-446.
- Coleman, N., Rousseau, N. et al. (2003). National Evaluation of Lone Parents Adviser Meetings: Findings from a Longitudinal Survey of Clients, BMRB Social Research.
- Davidson, C. and Woodbury, S. (1993). The Displacement Effect of Reemployment Bonus Programmes. *Journal of Labour Economics* 11(4): 575-605.
- Dehejia, R. and Wahba, S. (1999). Causal Effects in Non-experimental Studies: Re-evaluating the Evaluation of Training Programmes. *Journal of the American Statistical Association*. 94(448): 1053-62.
- Dehejia, R. and Wahba, S. (2002). Propensity Score Matching Methods for Non-experimental Causal Studies. *Review of Economics and Statistics* 84(1): 151-61.
- Department of the Environment, Transport and the Regions (2000). Indices of Deprivation.
- Dolton, P. and O'Neill, D. (2002). The Long-Run Effects of Unemployment Monitoring and Work-Search Programmes: Experimental Evidence from the United Kingdom. *Journal of Labour Economics* 20(2): 381-403.
- Dolton, P, Makepeace, G.W. and Treble, J. (1994). The Youth Training Scheme and the School to Work Transition. *Oxford Economic Papers* 46(4): 629-657.
- Dolton, P., Smith, J., Azevedo, J.P. (2003). Progress Report: Further Evaluation of the NDLP.
- DWP (2002). Lone parent brief. unpublished Sheffield.
- Eberwein, C. (2002). Correcting Estimates of expected durations from discrete-time hazard models, *Labour Economics*, 9(2), 303-339.
- Eberwein, C., Ham, J. and LaLonde, R. (2002). Alternative Methods of estimating programme effects in event history models, *Labour Economics*, 9(2), 249-278.
- Evans, M., Eyre, J. et al. (2003). New Deal for Lone Parents: Second Synthesis Report of the National Evaluation, Centre for Analysis of Social Policy, University of Bath, DWP Research Report W163, June 2003.
- Fredriksson, P. and Johansson, P. (2003a) Programme evaluation and random programme starts. Working Paper No. 2003:1. Institute for Labour Market Policy Evaluation, Uppsala, Sweden.

- Fredriksson, P. and Johansson, P. (2003b), Employment mobility, and active labour market programmes. Working Paper No. 2003:3. Institute for Labour Market Policy Evaluation, Uppsala, Sweden.
- Friedlander, D. and Burtless, G. (1995). Five Years After: The Long-Term Effects of Welfare-to-Work Programmes. New York: Russell Sage Foundation.
- Gueron, J. and Pauly, E. (1991). From Welfare to Work. New York: Russell Sage Foundation.
- Hales, J, Roth, W., Barnes, M., Millar, J., Lessof, C., Gloyer, M. and Shaw, A. Evaluation of the New Deal for Lone Parents: early lessons from the phase one prototype, DWP Research Report 109, 2000
- Hamilton, G. (2002). Moving People from Welfare to Work: Lessons from the National Evaluation of Welfare-to-Work Strategies. New York: Manpower Demonstration Research Corporation.
- Heckman, J.J., Hotz, V.J. (1989). Choosing Among Alternative Methods of Estimating the Impact of Social Programmes: The Case of Manpower Training, *Journal of the American Statistical Association*, 84, 862-874.
- Heckman, J.J., Ichimura, H. and Todd, P.E. (1997). Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme, *Review of Economic Studies*, 64, 605-654.
- Heckman, J.J., Ichimura, H. and Todd, P.E. (1998). Matching as an Econometric Evaluation Estimator, *Review of Economic Studies*, 65(2), 261-294.
- Heckman, J.J., Ichimura, H., Smith, J. and Todd, P.E. (1998), Characterising Selection Bias Using Experimental Data, *Econometrica*, 66(5), 1017-1098.
- Heckman, J.J., Lochner, L. and Taber, C. (1998), Explaining Rising Wage Inequality: Explorations with a Dynamic General Equilibrium Model of Labour Earnings with Heterogeneous Agents, *Review of Economic Dynamics*, 1(1), 1-58.
- Heckman, J.J., LaLonde, R.J., Smith, J. (1999), The Economics and Econometrics of Active Labour Market Programmes, in Ashenfelter, O. and Card, D. (eds.), *The Handbook of Labour Economics*, Volume III.
- Heckman, J. J. and Robb, R. (1986). Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes, *Economics Research Centre/NORC Discussion Paper*: 86-9: 77.
- Heckman, J. J. and Robb, R. Jr. (1985). Alternative Methods for Evaluating the Impact of Interventions. *Longitudinal Analysis of Labour Market Data*. J. J. Heckman and B. Singer.
- Heckman, J.J. and Robb, R. (1985). Alternative Methods for Evaluating the Impact of Interventions: An Overview, *Journal of Econometrics*, 30(1-2), 239-267.

- Heckman, J.J and Smith, J. (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.
- Hellerstein, J.K. and Imbens, G.W. (1999). Imposing Moment Restrictions from Auxiliary Data by Weighting, *Review of Economics and Statistics*, 81(1), 1-14.
- Horowitz J.L. and Manski, C.F. (1998). Censoring of outcomes and regressors due to survey non-response. Identification and estimation using weights and imputations, *Journal of Econometrics*, 84(1), 37-58.
- Ichimura, H. and Todd, P.E. (1999). Alignment Problem in Matching Estimators for Longitudinal Data. Unpublished manuscript, University of Pennsylvania.
- Imbens, G. (2000), The Role of Propensity Score in Estimating Dose-Response Functions, *Biometrika*, 87(3), 706-710.
- Lechner, M. (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.
- Lechner, M. (2001). Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption, in: Lechner, M., Pfeiffer, F. (eds), *Econometric Evaluation of Labour Market Policies*, Heidelberg: Physica/Springer, 43-58.
- Lechner, M. (2002). Programme Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labour Market Policies. *Review of Economics and Statistics*. 84(2), 205-220.
- Lessof, C., Miller, M., Phillips, M., Pickering, K., Purdon, S. and Hales, J., *New Deal For Lone Parents Evaluation: Findings from the Quantitative Survey*, (2003a). National Centre for Social Research, DWP Research Report WAE147, March 2003.
- Lessof, C., Phillips, M., Pickering, K. and Purdon, S. (2003). Estimating the impact of NDLP. Lone parents and employment: International comparisons of what works. J. Millar and M. E. (eds.), *Centre for the Analysis of Social Policy - University of Bath*, DWP Research Report 181, December 2003.
- Leuven, and Sianesi, B (2003) PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing, *Statistical Software Components S432001*, Boston College Department of Economics, revised 12 Sep 2005.
- Lise, J., Seitz, S. and Smith, J. (2002). Equilibrium Policy Experiments and the Evaluation of Social Programmes. Unpublished manuscript, University of Maryland.
- Little, R. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, New York, Wiley.

- Phillips, M., Pickering, K., Lessof, C., Purdon, S. and Hales, J. (2003). National Centre for Social Research. Evaluation of the New Deal for Lone Parents: Technical report for the quantitative survey, DWP Research Report 146
- Roed, K and Raaum, O. (2003) Administrative Registers – Unexplored Reservoirs of Scientific Knowledge? *Economic Journal*, vol.113, pp.F258-F281.
- Rosenbaum, P.R. and Rubin, D.B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, 70(1), 41-55.
- Rosenbaum, P.R. and Rubin, D.B. (1984). Reducing Bias in Observational Studies Using Sub-classification on the Propensity Score, *Journal of the American Statistical Association*, 79(387), 516-524.
- Rosenbaum, P.R. and Rubin, D.B. (1985), Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score, *The American Statistician*, 39(1), 33-38.
- Roy, A. 1951. Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers*. 3. 135-146.
- Rubin, D.B. (1974) Estimating Causal Effects of Treatments in Randomised and Non-Randomised Studies, *Journal of Educational Psychology*, 66, 688-701.
- Rubin, D.B. (1979) Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies, *Journal of the American Statistical Association*, 74(366), 318-328.
- Rubin, D.B. (1987) *Multiple Imputation for Non-response in Surveys*, Wiley, New York.
- Rubin, D.B. (1997), Assignment to Treatment Group on the Basis of a Covariate, *Journal of Educational Statistics*, 2(1), 1-26.
- Schafer, J. (1997) *Analysis of Incomplete Multivariate Data*, Chapman and Hall.
- Sianesi, B. (2001a). An Evaluation of the Active Labour Market Programmes in Sweden. Working Paper No. 2001:5, Institute for Labour Market Policy Evaluation, Uppsala, Sweden.
- Sianesi, B. (2001b). Psmatch: Perform Various Types of Propensity Score Matching, STATA Ado file.
- Sianesi, B. (2004). An Evaluation of the Swedish System of Active Labour Market Programmes in the 1990s, *Review of Economics and Statistics*, Vol. 86, No. 1, pp. 133-155.
- Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London; New York, Chapman and Hall.

Smith, J. and Todd, P. (2005), Does Matching Overcome LaLonde's Critique of Non-experimental Estimators? *Journal of Econometrics*, 125(1-2): 305-353.

Todd, P.A. (1999). A Practical Guide to Implementing Matching Estimators, mimeo.

Vella, F. (1998). Estimating Models with Sample Selection Bias, *Journal of Human Resources*, 33(1), 127-169.

Verbeek, M. and Nijman, T. (1992), Testing for Selectivity Bias in Panel Models, *International Economic Review*, 33(3), 681-703.

Wooldridge, J. (2002) Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification, CEMMAP working paper CWP11/02.