Dealing with randomisation bias in a social experiment exploiting the randomisation itself: the case of ERA

Barbara Sianesi

Dealing with randomisation bias in a social experiment exploiting the randomisation itself: The case of ERA

July 2013

Barbara Sianesi

Institute for Fiscal Studies

Abstract: We highlight the importance of randomisation bias, a situation where the process of participation in a social experiment has been affected by randomization *per se*. We illustrate how this has happened in the case of the UK Employment Retention and Advancement (ERA) experiment, in which over one quarter of the eligible population was not represented. Our objective is to quantify the impact that the ERA eligible population would have experienced under ERA, and to assess how this impact relates to the experimental impact estimated on the potentially selected subgroup of study participants. We show that the typical matching assumption required to identify the average treatment effect of interest is made up of two parts. One part remains testable under the experiment even in the presence of randomisation bias, and offers a way to correct the non-experimental estimates should they fail to pass the test. The other part rests on what we argue is a very weak assumption, at least in the case of ERA. We implement these ideas to the ERA program and show the power of this strategy. Further exploiting the experiment we assess the validity in our application of the claim often made in the literature that knowledge of long and detailed labour market histories can control for most selection bias in the evaluation of labour market interventions. Finally, for the case of survey-based outcomes, we develop a reweighting estimator which takes account of both non-participation and non-response.

Address for correspondence: Institute for Fiscal Studies, 7 Ridgmount Street, WC1E 7AE London, UK. E-mail: barbara_s@ifs.org.uk.

1. Introduction

Randomised social experiments are generally hailed as the gold standard in program evaluation. Under certain conditions, they are in fact the most reliable method for evaluating whether a program works, on average, for its participants.[1] An overarching label for such necessary identifying conditions is the "no randomisation bias" assumption (Heckman, 1992; Heckman *et al.*, 1999), which rules out that random assignment *per se* may affect average treatment and no treatment outcomes[2], as well as the program participation process.

In this paper we are in the unusual position to empirically assess part of the "no randomisation bias" assumption: the possibility that the program participation process has changed because of the presence of randomization. We develop a framework for the analysis of the consequences of this type of bias under the selection-on-observables assumption. We consider both the case of administrative outcome measures for the relevant sample and of survey-based outcome measures. With administrative outcomes we highlight how the randomisation itself can actually offer ways to support non-experimental methods in addressing the shortcoming it gave rise to. Specifically, we show that the typical matching assumption required to identify the average treatment effect of interest is made up of two parts. One part remains testable under the experiment even in the presence of randomisation bias, and offers a way to correct non-experimental estimates that fail to pass the test. The other part rests on what we argue is a very weak assumption, at least in our application. We thus showcase the usefulness of a judicious combination of both non-experimental methods and the experimental set-up in overcoming the latter's shortcoming when administrative outcome data are available. We additionally exploit the experiment to assess the validity in our application of the claim often made in the literature that knowledge of long and detailed histories can control for most selection bias in the program evaluation. For the case of survey outcomes we extend our estimators dealing with non-participation to account for selective non-response based on observable characteristics.

The issue which motivated the paper arose in the recent Employment Retention and Advancement (ERA) demonstration, which ran in six districts across the UK between 2003 and 2007. With over 16,000 individuals being randomly assigned, the ERA study represented the largest randomised trial of a social program in the UK. The experiment was set up to test the effectiveness of a package of time-limited support once in work, combining advisory services with a new set of financial incentives rewarding sustained full-time work and the completion of training whilst employed. Eligible for this initiative were long-term unemployed over the age of 25 who were mandated to enter the New Deal 25 Plus (ND25+) program, and lone parents who volunteered for the New Deal

---

[1] For a discussion and appraisal of randomised experiments, see e.g. Burtless (1995) and Heckman and Smith (1995).

[2] As such it also rules out control group contamination, whereby control group members engage in a different type or intensity of alternative programs from what they would have done in the absence of the experiment.

for Lone Parents (NDLP) program.[3] In the first follow-up year, the employment chances of both groups remained largely unaffected, while a sizeable experimental impact was found in terms of earnings, especially for the NDLP group (see Hendra *et al*., 2011 for the final appraisal of ERA).

Since ERA offered of a package of support once in work, *all* individuals flowing into ND25+ and NDLP in the six evaluation districts during the one-year intake window should automatically have become eligible to be offered ERA. It has however emerged that only parts of the target population have entered the evaluation sample: some eligibles actively refused to be randomly assigned (the "formal refusers"), while some were somehow not even offered the possibility to participate in random assignment and hence in ERA (the "diverted customers"). A sizeable fraction of the eligibles – 23% of ND25+ and 30% of NDLP – were thus not represented in the experiment.

While the policymaker would be interested in assessing the average impact of offering ERA services and incentives for all those eligible to receive such an offer, the experimental evaluation can provide unbiased impact estimates only for the ERA study participants – those who reached the randomisation stage and agreed to participate in the demonstration. The concern is that this subgroup may be a selective one, not representative of the eligible population in the ERA districts who would have been eligible for ERA had it been an official national policy.[4] The fact that ERA was a study and involved random assignment has significantly altered how the intake as a whole was handled, as well as the nature of the adviser/New Deal entrant interaction in a way that would not have been the case if ERA had been normal policy. Indeed it was the set-up of the experimental evaluation *per se* which gave rise to diverted customers and formal refusers – these eligible individuals were denied or 'refused' participation in something which in normal circumstances one could not be denied or one could not 'refuse': becoming *eligible* for financial incentives and personal advice. Randomisation can thus be viewed as having affected the process of participation in ERA, resulting in an adviser-selected and self-selected subgroup which is potentially different from the sample of New Deal entrants that would have been exposed to ERA had it not been evaluated via random assignment. Non-participation can thus be seen as potentially introducing randomisation bias in the experimental estimate for the impact of offering ERA eligibility on the eligible population.[5]

Note that non-participation in the ERA study, which takes place before random assignment, is a distinct problem from non- or partial compliance (no-shows, drop-outs, non-take up), which takes

---

[3] These two groups represent 83% of all ERA study participants. We do not consider the third target group due to its conceptually different set-up coupled with lack of data.

[4] ERA as a normal policy would be envisaged as an integral, seamless component of the New Deal program in which *any* New Deal participant would automatically be enrolled upon entering work.

[5] An alternative but, as we discuss in Section 2.3, possibly less pertinent way to consider this issue is as a threat to the external validity of the experimental estimates.

place *after* treatments have been assigned.[6] This type of non-participation is also separate from entry effects[7]; the extrapolation to other populations beyond the pilot areas (see e.g. Hotz *et al.*, 2005, for the extrapolation of experimental results to other sites); and attrition (loss of experimental sample in the collection of outcome information).

The beauty of the ERA study is that it offers the rare chance to actually measure the extent of randomisation bias. This is because (1) the treatment is the offer of ERA support and incentives, (2) the whole population of ND25+ and NDLP entrants in the six districts was eligible for this offer (and would be eligible under an official policy) and (3) such entrants are identified in the data.

The key objective of the paper is to recover the causal effect for the full eligible population of making the ERA package available. Given the substantial first-year experimental impacts, especially for the NDLP group, it is important to assess how robust the findings are to non-participation bias. We thus first use non-experimental methods to quantify the impact that the full eligible population would have been likely to experience in the first follow-up year had they been offered the chance to participate in ERA, and subsequently assess how this impact for the eligible group relates to the experimental impact estimated on the subgroup of study participants.

With "experimentation on a context-specific subpopulation", Manski (1996) advocates bounding the treatment effect of substantive interest (done for ERA in Sianesi, 2010), as in general very strong assumptions on entry into the study are needed for point identification. Our analyses focus on matching and reweighting techniques under the conditional independence assumption (CIA) that we observe all outcome-relevant characteristics that drive selection into the ERA study.[8] While our data include demographics, information on the current unemployment spell, extremely detailed labour market histories over the previous three years and local factors, the CIA needed to identify the average treatment effect on the non-treated (the non-participants in our case) is admittedly a strong assumption. We show however that this matching assumption has two parts, and that one part is indeed testable under the experiment despite randomisation bias. The other part is the requirement that individuals were not diverted or did not formally refuse based on residual unobserved idiosyn-

---

[6] The set-up and aims of Dubin and Rivers (1993) are opposite to the ones in the current paper. In their set-up, refusal to participate in the wage subsidy experiment happened after random assignment (to the program group). While their experiment thus directly recovers the intention to treat (it also includes non-take up of the subsidy by participants themselves), the authors aim to tease out the impact on the participants. Their formal refusers could be viewed as the program group "no-shows" considered by Bloom (1984), and indeed the approach followed by Dubin and Rivers builds upon the Bloom estimator. Note also that the non-participants in the ERA experiment were not exposed to ERA, and thus no link can be made to the literature on "dropouts" (see e.g. Heckman *et al.*, 2000).

[7] The new ERA incentives and entitlement rules could affect individual behaviour so as to gain eligibility (see Moffitt, 1992). Some long-term unemployed could e.g. be induced to delay exit from unemployment in order to reach the start of ND25+-with-ERA, or more lone parents could be induced to volunteer for NDLP-with-ERA. The composition of New Deal entrants if ERA were an established intervention would thus be different from the one during the experiment.

[8] The only other paper we are aware of which considers this kind of non-participation in an experiment, Kamionka and Lacroix (2008), relies on a duration model under different distributional assumptions on unobserved heterogeneity.

cratic ERA impacts conditional on arbitrarily heterogeneous impacts according to our rich set of observed characteristics – a highly plausible assumption as we argue in Section 4.2. In our specific set-up, we can thus formally test the standard matching assumption under the very weak condition of no selection into the ERA study based on unobserved impacts. Furthermore, in cases where the experimental data do reject the standard CIA, information from the experiment can be used to correct the non-experimental estimates under the assumption of no selection on unobserved impacts.

Exploiting the experiment we also assess the validity in our application of the claim often made in the literature that knowledge of long and detailed labour market histories can control for most selection bias in the evaluation of labour market interventions (see e.g. Dehejia and Wahba, 1999, Heckman and Smith, 1999, Heckman *et al.*, 1998, Heckman *et al.*, 1999, and Frölich, 2004, to some extent Hotz *et al.*, 2005, and for a caveat, Dolton and Smith, 2011, and Lechner and Wunsch, 2011).

Our main findings tell a consistent story. Those non-experimental estimates of average employment and earnings ERA impacts for all eligibles which pass the CIA test are found to be statistically indistinguishable from the corresponding experimental effect for the participants. Non-experimental estimates that fail to pass the test indicate that the experimental estimates significantly underestimate the average impact that the full eligible population would have experienced had they been offered the chance to participate in ERA. However, once these non-experimental estimates are corrected to take into account failure of the test, the story changes back to one of representativeness of the experimental estimate for the effect on all eligibles. When combined, experimental data and non-experimental methods thus suggest that despite the sizeable share of non-participants, the ERA experiment does not seem to have suffered from randomisation bias in terms of year-1 impacts.

The claim often made in the literature that histories variables can capture labour-market relevant unobservables was not borne out in our data, at least in the no-treatment case, which is indeed the case of interest when using non-experimental comparison groups for estimating treatment effects. Additionally, in contrast to Dolton and Smith (2011), the way of summarising labour market histories did not make the slightest difference in reducing selection bias.

The remainder of the paper is organised as follows. In Section 2 we outline how non-participation in the ERA experiment has come about, summarise the available qualitative evidence and discuss ways to view the issue of non-participation. Section 3 describes the sample definitions and data content, and provides some basic descriptives. Our methodological approaches and the type of analyses we perform are described in Section 4. The results of all the empirical analyses are presented and discussed in Section 5, while Section 6 concludes.

2. Non-participation in the ERA study: The issues

2.1 How non-participation came about

The demonstration was set-up to test the effectiveness of ERA, the offer of a package of support. While still unemployed, ERA offered job placement assistance, as done by the regular New Deal programs. Once in work and for up to two years, ERA offered both the support of an adviser to help retain and progress in work, as well as eligibility to a retention bonus of £400 three times a year for staying in full-time work 75% of the time, to training tuition assistance (up to £1,000), to a bonus (also up to £1,000) for completing training whilst at least part-time employed, and to access emergency payments to overcome short-term barriers to remain in work.

In an ideal scenario, all individuals in the six evaluation districts who would be eligible for ERA if it were an official policy would have been randomly assigned to either the program or control group. Departures from this ideal situation have arisen from two sources:

1. intake process: not all eligible individuals have been offered the possibility to participate in random assignment and hence in ERA (the "diverted customers"); and

2. individual consent: some individuals who were offered the chance to take part in the experimental evaluation actively refused to do so (the "formal refusers").

Diverted customers and formal refusers make up the group of the ERA non-participants, those who whilst eligible for ERA have not been included in the experimental sample. The ERA study participants are those who were eligible for ERA, were offered the chance to participate in the study *and* agreed to take part. These are those making up the evaluation sample, i.e. those who were subsequently randomly assigned either to the program group, who received ERA services and incentives, or to the control group, who only received the baseline New Deal program whilst unemployed.

2.2 What is known about non-participation in the ERA study

Qualitative work conducted as part of the ERA evaluation has shed interesting light on the origins of non-participation by looking closely at the assignment and participation process in ERA at selected sites (Hall *et al.*, 2005, and Walker *et al.*, 2006). Based on detailed observations and interviews with staff and individuals, the authors conjecture that it is quite unlikely for ERA non-participants to be a random subgroup of the two eligible New Deal groups. The discussion of what is known about non-participation from this qualitative work is organized in two parts.

1. Ensuring that staff randomly assigned all eligible individuals

The six districts could exercise significant discretion in how they organised the ERA recruitment, intake and random assignment processes.[9] Although the expectation was that the intake staff, be it an ERA or a New Deal Adviser, would encourage *all* eligible individuals – and encourage all of them equally hard – to consent to be randomly assigned and have a chance to participate in ERA, staff could use discretion on two fronts: what individuals to tell about ERA, directly determining the extent of diverted customers, and in what terms to present and market ERA to individuals, thus affecting their likelihood to become formal refusers. As to the latter, the abstract notion that staff would use the same level of information and enthusiasm in recruiting all eligible individuals was particularly hard to implement in practice. Discretion in their choice of marketing strategy could take various forms: how 'hard' to sell ERA; what features of the program to mention – in particular whether and in what terms to mention the retention bonus, or whether to selectively emphasise features (e.g. the training bonus) to make ERA more appealing to the particular situation of a given individual; and how far to exploit the misunderstanding that participation in ERA be mandatory.[10]

But why and under what circumstances would caseworkers want to apply such discretion? Advisers were given individual-level targets for how many people they moved into work and were accordingly rewarded for job entries. This incentive structure seems to have led advisers conducting the intake process to use their own discretion in deciding what individuals to sell random assignment or how hard to sell it in order to 'hang onto' those whom they perceived as likely to move into work quickly. Job entry targets had an asymmetric influence on the incentives of New Deal and of ERA advisers: where the intake was conducted by New Deal advisers, job-ready individuals would be more likely to be diverted from ERA; where ERA advisers were doing the intake, they would be less likely to be diverted.[11] It thus appears quite unlikely that non-participants, and especially diverted customers, be random subgroups of the eligible population; rather, these were people whom advisers had a vested interest in not subjecting to ERA.

---

[9] In some districts, it was the New Deal advisers who conducted the intake and randomisation, with the ERA advisers being responsible for working with the ERA program group only after random assignment had taken place. In other districts, both types of advisers were responsible for conducting intake interviews and randomisation. These models did not necessarily apply at the district level, since within a particular district, different offices and staff members sometimes used somewhat different procedures. The intake and randomisation procedures further varied over time, in the light of experience and depending on the situation and needs of the district or even single office.

[10] It additionally became apparent that probably owing to their greater knowledge of and enthusiasm for ERA, ERA advisers tended to give clearer explanations of ERA than New Deal advisers (Walker *et al.*, 2006, Appendix F).

[11] "Overall, when New Deal Personal Advisers undertook the interviewing, they had reason to encourage people with poor job prospects to join ERA (because in many cases they would move on to ASAs and off their caseloads) and those with good prospects to refuse (because they would keep them on their caseloads and get credit for a placement). When ASAs [ERA advisers] were involved in conducting intake interviews, they could have benefited from encouraging customers with poor employment prospects to refuse ERA and people with good prospects to join." (Walker *et al.*, 2006, p.26). In conclusion: "While [this] incentive structure was real and widely recognised, it is impossible to assess with any degree of precision how strong an effect it had on marketing strategies (and, thus, on the resulting make-up of the groups of customers who ended up being randomly assigned)" (p.27).

2. How willing were individuals to be randomly assigned?

Individuals who were given the option to participate in random assignment could formally refuse[12] and thus be excluded from the experimental sample. It is not fully clear how much individuals actually knew about what they were refusing – according to observations at intake and interviews with the unemployed themselves after those sessions, not much.[13]

The qualitative work highlighted how recruitment to ERA greatly differed between the two New Deal groups. While lone parents on NDLP were all volunteers to that program and thus mostly responded favourably to ERA too, ND25+ participants were more difficult to recruit. The reasons for formal refusal that were identified included being puzzled by how the additional offer of ERA fitted in the mandatory participation in ND25+, having been unemployed for long periods of time and thus finding it difficult to envisage what might happen after they obtained a job, an outcome that they and their advisers thought rather unlikely anyway, and feeling close to getting a job in the near future and not wanting to stay in touch with the office. It thus appears that the group of formal refusers, and in particular those amongst the more problematic ND25+ group, might be far from random, and instead selected on (predicted) non-ERA outcomes. Some staff further identified specific attitudes and traits as good predictors that individuals, particularly among those mandated to start ND25+, would decline participation: a strong antipathy to government, feeling alienated from systems of support and governance, being resistant to change or taking risks, 'preferring to stick with what they know', reacting against the labour market, and enjoying being able to refuse to do something in the context of a mandatory program. A further possible reason for refusal was being engaged in benefit fraud. Overall, the qualitative evidence suggests that those who declined to join may, in fact, differ in important respects from those who agreed to participate. Formal refusers, especially those amongst the more problematic ND25+ group, appeared to have weaker job prospects and poorer attitudes than the average New Deal entrant.

As mentioned, caseworkers could decide how to sell ERA in order to steer individuals' refusal decisions. When New Deal advisers undertook the intake interviews, they could benefit if job-ready individuals refused to participate in ERA and those with bad prospects consented. Conversely, when ERA advisers were leading the intake process, they could benefit if individuals with bad job prospects formally refused, while those with good prospects agreed.

While the insights provided by these in-depth case studies were based on only very few observations and thus could not be safely generalised, Goodman and Sianesi (2007) thoroughly explored

---

[12] Signing: "*I do not consent to taking part in this research scheme or to being randomly assigned.*"
[13] Walker *et al*. (2006) conclude that "very few customers could be described as understanding ERA, and all of them had already been assigned to the program group and therefore had been given further details about the services available" and "there was a consensus among the Technical Advisers who conducted both the observations and the interviews with customers […] that most customers truly did not have a good appreciation of ERA." (p.43).

both how large and how selective the non-participating groups were. Results are summarised in Section 3.4, highlighting how the non-participation problem is a relevant one, both in terms of its incidence (26.6% of all eligibles) and of the diversity of the excluded groups.

2.3 How to view non-participation

Non-participation in the ERA study can be considered in different ways.

If the parameter of interest is the impact of the ERA offer for the sample of participants, non-participation can be viewed as a potential threat to the external validity of the experimental estimate (Cook and Campbell, 1979). External validity in this case relates to the extent to which the conclusions from the experiment would generalise to the whole eligible population (in the six evaluation districts).[14] This is how Kamionka and Lacroix (2008) cast the problem of non-participation in the Canadian Self-Sufficiency Entry Effects Demonstration, in which some eligibles could either not be contacted at baseline or refused to take part in the experiment. While the latter are the counterparts of the formal refusers in the ERA study and by construction arose because of randomisation, it does not in fact seem appropriate to argue that random assignment *per se* gave rise to the first type of non-participation in the Canadian experiment.

If by contrast the parameter of substantive interest is the impact of the ERA offer for all eligibles (in the six districts), non-participation can be viewed as having potentially introduced bias in the experimental estimate for the parameter of interest. As argued in the introduction, it seems pertinent to view this "non-participation bias" as randomisation bias. Had there been no random assignment, there would have been no need to ask for consent to participate in the experiment, and there would thus have been no formal refusers.[15] As to the diverted customers, there are always individuals who do not know about policies available to them and whose advisors do not know either or do not think they would benefit (indeed, around one quarter of the ERA program group "had not heard of the retention bonus" according to the 1-year follow-up survey). Such "no shows" would however still need to be included in the evaluation sample to assess the intention to treat – the casual effect for the eligibles of making such a package available. By contrast, the way the experimental evaluation was set up on the ground in the specific case of ERA created diversion incentives for advisors which normally would not have been there, worsened by a situation where no outside information on ERA was available.[16] The fact that ERA was a study and involved random assignment

---

[14] We are always only concerned with the *current* experimental evaluation, i.e. the eligible group within the six ERA districts over the study intake window. There is the wider generalisability question that has a national rollout in mind and which relates to how the experimental results obtained in the six districts would generalise to the rest of the country.
[15] As mentioned, individuals were not refusing ERA – of which they had no real knowledge – but to take part in the research or be randomly assigned.
[16] This is of course not to say that randomisation necessarily entails randomisation bias, and in principle an experiment could have been devised in such as way as to avoid non-participation (e.g. not asking for consent, randomising offices

has thus created a pool of eligible individuals who were denied or 'refused' participation in something which in normal circumstances one could not be denied or one could not 'refuse': becoming *eligible* for financial incentives and personal advice. It is in this sense that non-participation can be seen as having potentially introduced randomisation bias in the experimental estimate for the impact of offering ERA eligibility on the eligible population.

3. Data and sample

3.1 Data

A number of data files have been put together for the analysis. The administrative data held by the Department for Work and Pensions (DWP) on ND25+ and NDLP entrants provided us with the sampling frame. We extracted files for all cases identified as having entered these New Deal programs in the six districts over the relevant random assignment period, as detailed below. We have further exploited the New Deal extract files for information about past program participation as well as a number of other relevant individual characteristics.

We have then merged these files with the Work and Pensions Longitudinal Study (WPLS). This relatively recently released, spell-level dataset contains DWP information about time on benefits and HMRC records about time in employment and, what became available only later in the evaluation, tax year earnings. These administrative records have been used to construct both detailed labour market histories and outcome measures. We have further combined the administrative data with information from the ERA evaluation dataset on the participation decision and the outcome of random assignment. We have finally merged in local-area level data (Census, travel-to-work and super-output area data). In section 3.3 we summarise the extensive variables we have selected and derived from all of these sources.

3.2 Sample

To define our sample of ERA eligibles, we need to clearly define the criteria determining eligibility and identify the relevant individuals in the data.[17] We consider as *eligible* for ERA:

1. those who became mandatory for ND25+ during the period when the respective district was conducting random assignment *and* who subsequently also started the Gateway still within the relevant random assignment intake window; and

---

rather than individuals, or even changing the incentive structure). Randomisation bias can however happen in a given experimental set-up – and it has happened in the ERA set-up.

[17] See Goodman and Sianesi (2007) for a description of how problem cases were handled and what adjustments were performed on the ERA experimental sample.

2. those lone parents who were told about NDLP (had a work-focussed interview and/or expressed an interest in NDLP) during the period when the respective district was conducting random assignment *and* who subsequently also volunteered for NDLP still within the relevant random assignment intake window.

The random assignment window was actually district- and intake group-specific, since one district started conducting random assignment later than the others and some districts stopped conducting random assignment for some groups earlier.[18]

We also consider ERA impacts on earnings collected from the first ERA customer survey. This survey covers the experiences of a sample of ERA participants during the first 12 months following individuals' dates of random assignment. When looking at survey outcomes, we consider the intersection of the random assignment and survey intake windows. There is in fact very good overlap, with only 5.6% of the full eligible sample being lost when imposing consistent intake criteria with those used to select the survey sample.

Table 1 provides sample breakdowns by participation status and survey status. Non-participation was substantially lower amongst the ND25+ group (23% of all eligibles) than the NDLP group (over 30%). We observe survey outcomes for around one third of study participants.

Table 1   Sample breakdown by target group

	ND25			NDLP		
Eligibles	7,796	100.0%		7,261	100.0%	
– Study non-participants	1,790	23.0%		2,209	30.4%	
– Study participants	6,006	77.0%	100.0%	5,052	69.6%	100.0%
– with survey outcome	1,840		30.6%	1,745		34.5%
– without survey outcome	4,166		69.4%	3,307		65.5%

3.3 Outcomes and observable characteristics

We assess ERA impacts on employment and earnings during a 12-month follow-up period using both administrative and survey measures.

Administrative data on employment is available from WPLS records for the *full* sample of ERA eligibles in the six districts, i.e. including the non-participants. We consider the probability of having ever been in employment and the total number of days in employment, counting the 12-month follow-up period from the moment individuals flowed in (i.e. from the moment ND25+ entrants started the Gateway, or lone parents volunteered for NDLP).

---

[18] Random assignment was conducted between 1 Nov 2003 and 31 Oct 2004, with the exceptions of North West England (3 Jan 2004 to 31 Jan 2005) and the NDLP intake in Wales (1 Nov 2003 to 21 Aug 2004).

Survey data on earnings in the 12-month follow-up period is available for a sample of partici-pants. This measure offers a clean definition of earnings (including all part-time work and self-employment) over a comparable horizon for each individual, i.e. over the year since their individual random assignment date. This was the only earnings information originally available. Subsequently, administrative earnings information became available for all eligibles. However, these data do not include self-employment spells, nor do they systematically capture all part-time workers with low earnings. Furthermore, earnings are related to fiscal years, thus covering different horizons for dif-ferent individuals in relation to random assignment. Indeed, for a relevant share of our sample (65% of ND25+ and 59% of NDLP eligibles), 2004/05 fiscal year earnings partially cover *pre*-treatment periods (see Figure 1). Nonetheless, there is scope to use this administrative information for sensi-tivity analyses of survey-based estimates.

All our outcomes of interest – employment probabilities and durations, and earnings – are re-lated to labour market performance. As listed in Table 2, we have put together an extensive collec-tion of individual, office and local area characteristics that are most likely to affect individuals' la-bour market outcomes, and that might potentially have affected selection into the ERA sample. Note that all of these variables have to be defined both for the ERA study participants and non-participants, which required us to derive such information from administrative data sources alone.

In addition to demographic characteristics (gender, age, ethnicity, partner and children, disabil-ity and illness), we have summarised information on an individual's current unemployment spell, including in particular indicators of a very recent/current employment spell, how long it took them to start the Gateway or volunteer for NDLP once having become mandatory for it or being told about it, and of whether ND25+ entrants volunteered for the Gateway ahead of time. We have also created variables capturing the extent of past participation in voluntary employment programs (as a crude indicator of willingness to improve one's circumstances), in the ND25+ (a mandatory pro-gram) and in Basic Skills (a program designed to address basic literacy, numeracy and IT skills).

We have further constructed three years' worth of labour market history in terms of time in employment, on active benefits (JSA and compensation whilst on a labour market program) and on inactive benefits (Income Support and Incapacity Benefits). As highlighted in the table, we experi-mented with different ways of capturing these histories. The parsimonious 'summary' consists of a series of dummy variables capturing the proportion of time employed (zero, less than 25%, 25 to 50%, more than 50%) and the proportion spent on benefits (zero, less than 50%, more than 50%, 100%)s, separately on active and inactive benefits. 'Employment dummies' are 36 monthly dummy variables indicating whether the individual had a positive number of days employed at any time during each of the 36 months pre-inflow. The 'sequence dummies' follow Card and Sullivan (1988) in building a series of dummy variables, each capturing a labour market *sequence* over the past 3

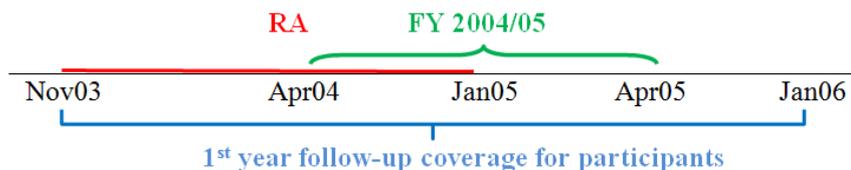Figure 1: Timeline of Random Assignment (RA) and 2004/05 tax year coverage

Table 2   Summary of observed characteristics

ERA district	
Inflow month	District-specific month from random assignment start when the individual started the ND25 Gateway or volunteered for NDLP
Demographics	Gender, age, ethnic minority, disability, partner (ND25+), number of children (NDLP), age of youngest child (NDLP)
Current spell	Not on benefits at inflow (NDLP), employed at inflow (indicator of very recent/current employment), time to show up (defined as the time between becoming mandatory for ND25+ and starting the Gateway or between being told about NDLP and volunteering for it), early entrant into ND25+ program (Spent <540 days on JSA before entering ND25+)
Labour market history (3 years pre-inflow)	Past participation in basic skills, past participation in voluntary programs (number of previous spells on: NDLP, New Deal for Musicians, New Deal Innovation Fund, New Deal Disabled People, WBLA or Outreach), past participation in ND25+; Active benefit history (JSA and compensation from NDYP, ND25+, Employment Zones and WBLA and Basic Skills), inactive benefit history (Income Support and Incapacity Benefits), employment history: (1) parsimonious summary (2) monthly employment dummies (3) dummies for sequences of employment/benefits/neither states (4) dummies for ever employed in 12m window at any time in the past
Local conditions	Total New Deal caseload at office, share of lone parents in New Deal caseload at office, quintiles of the index of multiple deprivation, local unemployment rate

years.[19] As it turned out, though the specific combinations differ for the two intake groups, the first 22 (out of 48) combinations cover in both cases exactly 90% of the sample. Lastly, a series of dummies for being 'ever employed' during a 12-month window at any time in the past (specifically, between 1+k and 12+k months pre-inflow, with k=0, 3, 6, 9, 12, 15, 18, 21, 24).

The Census has provided us with information on local labour market conditions (travel-to-work area unemployment rates) and on the deprivation of the area the individual lives in (index of local deprivation). We have also constructed information at the office level (total New Deal caseload and

---

[19] The sequence is defined according to status over 3 adjacent periods. For ND25+: 1 to 18 months (most would be on JSA); 19 to 27 months and 28 to 36 month pre-inflow. For NDPL: 1 to 12 months, 13 to 24 months and 25 to 36 months pre-inflow. State can be in the first period: always on benefits, employed for at least one month, anything else; in the second period: always on benefits, employed for at least 5 months, no employment and no benefits for at 5 five months, anything else; and in the third period: always on benefits, employed for at least 5 months, no employment and no benefits always, anything else.

share of lone parents in such caseload), aimed at capturing office-specific characteristics that might impact on the probability of participation in the study as well as on subsequent outcomes.

Despite offering such rich and detailed information, the administrative data do not contain information on education, which thus remains an unobservable, together with "innate ability" or work commitment. The previous literature has however indicated the potential for detailed and flexibly modelled labour market histories (like those we have constructed) to help proxy such unobserved traits and thus to eliminate much of the selection bias (see e.g. Dehejia and Wahba, 1999, Heckman and Smith, 1999, Heckman *et al.*, 1998, Heckman *et al.*, 1999, and Frölich, 2004, and to some extent Hotz *et al.*, 2005). Recent work by Dolton and Smith (2011) has however qualified this claim. While finding support for the widely recognised importance of controlling for pre-program outcome measures – and to do so in a flexible way – in order to reduce selection bias, they claim that even then important unobservables have remained unaccounted for. These conclusions were reached by noting how much their non-experimental impact estimates change: conditioning on histories in a flexible rather than parsimonious way reduces impact estimates to more *a priori* 'reasonable' values, and further conditioning on a number of survey measures of attitudes towards work for a subset of their sample has a large effect on the impact estimates, highlighting how even flexibly modelled histories did not fully capture these otherwise unobserved factors.

We are in a position to assess the validity of both conjectures in a more formal way, as the specific nature of our set-up and data – randomisation coupled with administrative outcome data for the non-participants – allows us to perform a number of tests not generally available. The sensitivity tests outlined below lend themselves to formally quantify how much selection bias is reduced by controlling for detailed as opposed to parsimonious histories, as well as whether controlling for histories is indeed enough to remove all selection bias.

3.4 Descriptive analysis

Table 3 shows that as to incidence, non-participation overall was lower amongst ND25+ (23%) than NDLP entrants (over 30%). In terms of composition, 9% of all ND25+ eligibles have been diverted and 14% formally refused. By contrast, over one quarter (26.4%) of all eligible NDLP entrants appear to have been diverted, while only 4% formally refused.

There was also marked variation in the incidence and composition of non-participation according to ERA district, with some clear outliers in terms of performance. In the East Midlands almost half of all eligible NDLP entrants did not take part in ERA, most of whom diverted customers. The performance of Scotland and North West England is particularly remarkable, with not one single diverted customer among the ND25+ group, while North East England stands out with over one quarter of the ND25+ eligible population being formal refusers.

Table 3: Breakdown by district (%)

| | ND25+ | | | NDLP | | |
	Non-participants	*Diverted Customers*	*Formal Refusers*	Non-participants	*Diverted Customers*	*Formal Refusers*
All	23.0	*9.4*	*13.6*	30.4	*26.4*	*4.0*
Scotland	8.7	*0.0*	*8.7*	5.3	*2.5*	*2.8*
NE England	34.9	*8.8*	*26.1*	29.2	*28.2*	*1.0*
NW England	14.6	*0.0*	*14.6*	6.2	*2.5*	*3.7*
Wales	20.7	*9.6*	*11.1*	23.6	*20.1*	*3.6*
East Midlands	27.5	*16.8*	*10.7*	47.1	*41.2*	*5.9*
London	25.8	*14.8*	*11.1*	31.0	*26.1*	*4.9*

Goodman and Sianesi (2007) uncovered a very strong role of office affiliation in determining both ERA offer and consenting choice, though as expected it was stronger in the former. Most of the explained variation in ERA offer, acceptance and participation was accounted for by an individual's district, office affiliation and inflow month[20], underscoring the key role played by local practices. Individual employment prospects, as well as attitudes towards and past participation in government programs were however also found to matter, leaving only a residual role to demographic characteristics (see also Appendix Table A1).

In the absence of randomisation bias, the control group and the non-participants should experience similar outcomes, as neither of them has been offered ERA services. However, Goodman and Sianesi (2007) have found non-participants to be somewhat higher performers than participants in terms of labour market outcomes among NDLP entrants, but to have significantly worse employment outcomes among ND25+ entrants.

4. Methodological approaches

4.1 *Analysis framework*

The population of interest are those eligible to be offered ERA services and incentives. We implicitly condition on this population throughout. The binary variable Q captures selection into the ERA study, with $Q=0$ denoting individuals who despite being eligible have not been randomly assigned, and $Q=1$ denoting the study participants. The participants make up the experimental group which was randomly assigned between a program group who was offered ERA ($R=1$) and a control group who was not ($R=0$). The problem to be addressed is changes in participation patterns introduced by the experimental evaluation, given that due to diversion and refusal to be randomly assigned, the population under the experiment ($Q=1$) does not correspond to the eligible population, made up by the ($Q=1$) and ($Q=0$) groups.

---

[20] Over time, formal refusal rates fell for both intake groups, likely to reflect increased adviser experience in selling ERA and the permission to mention ERA financial incentives.

Further, let S denote the availability of a survey-based outcome measure conditional on ERA participation. Specifically, $S=1$ when survey outcomes such as earnings are observed; this happens only for that subsample of participants who (1) were randomly selected to be surveyed, (2) could be contacted, (3) accepted to take the survey and (4) answered the earnings question. For short, we refer to them as "respondents". We refer to "non-respondents" ($S=0$) those ERA participants with missing survey outcome information, whatever the reason.

Let $p \equiv P(Q=0)$ be the probability of non-participation among the eligibles, directly identified in the data by the proportion of non-participants among the eligibles (see Table 1).

Denote the observed outcome by Y and define two potential outcomes for each eligible individual i: Y_{1i} the outcome if i were offered ERA services and Y_{0i} the outcome if i were not offered ERA services.[21] In addition to SUTVA[22], we need to assume that these treatment and no-treatment outcomes among the eligibles are not affected by whether an individual participates in the study or not, i.e. participants and non-participants may be drawn from different parts of the distributions of observed and unobserved characteristics, but the mere fact of having the chance to participate in the experiment does not change the relationship between characteristics on the one hand and treatment and no-treatment outcomes on the other. This requires the potential outcomes of individual i not to be indexed by Q: $Y_{1Qi} = Y_{1i}$ and $Y_{0Qi} = Y_{0i}$ for $Q=0, 1$.

Our parameter of interest is the average treatment effect (*ATE*) of offering ERA on the full ERA eligible population in the six districts, defined as the average outcome for all those eligible for ERA if they were offered ERA services compared to the average outcome for all those eligible for ERA if they were not offered ERA services: $ATE \equiv E(Y_1 - Y_0)$.

Denote the average impact of ERA on the participants by $ATE_1 \equiv E(Y_1 - Y_0 \mid Q=1)$ and on the non-participants by $ATE_0 \equiv E(Y_1 - Y_0 \mid Q=0)$. The three impacts are then linked according to:

$$ATE = (1-p)\cdot ATE_1 + p\cdot ATE_0 \tag{1}$$

Equation (1) states that the parameter of interest *ATE* is given by a weighted average of the impact on study participants and of the impact the non-participants would have experienced, with weights given by the relative share of participants and non-participants within the eligible pool.

While ATE_0 and hence *ATE* are unobserved, under some conditions (randomisation has not disrupted the program, there has been no control group contamination and outcomes are observed for all or a random sample of the participants), the available experimental data identifies ATE_1, the effect of ERA for participants in the experiment. Due to the randomness of R within the $Q=1$ group:

---

[21] For the potential outcome framework see Rubin (1974); for a review of the evaluation problem see e.g. Heckman *et al.* (1999).

[22] The stable unit-treatment value assumption (SUTVA, Rubin, 1980) requires that an individual's potential outcomes as well as treatment choice do not depend on the treatment choices of other individuals in the population. The former rules out general equilibrium effects in the ERA study, the latter is satisfied in the experiment.

$ATE_1 \equiv E(Y_1|Q=1) - E(Y_0|Q=1) = E(Y_1|Q=1, R=1) - E(Y_0|Q=1, R=0) = E(Y|R=1) - E(Y|R=0).$

If selection into the study has taken place, the composition of participants will be different from the composition of the eligible population, and impacts estimated on participants will not in general be representative of the impacts that the eligible population would have experienced.[23]

We consider how to deal with randomisation bias in experimental studies when follow-up information on the outcomes of the non-participants is available (administrative outcomes – Section 4.2) and when it is not (survey outcomes – Section 4.3).

4.2 Follow-up data on the non-participants (administrative outcomes)

In case of administrative data on the outcomes of all eligibles, ATE_1 is identified by the experimental contrast and recovering ATE_0 is akin to recovering the average treatment effect on the non-treated, given that, as in the standard case, the no-treatment outcome of the non-treated (i.e. the non-participants) is observed. Equation (1) thus becomes:

$$ATE = (1-p) \cdot ATE_1 + p \cdot \{E(Y_1 \mid Q=0) - E(Y \mid Q=0)\}. \tag{2}$$

As in a typical matching context, to estimate ATE_0 and hence ATE, we thus only need to identify $E(Y_1 \mid Q=0)$, the average outcome that the non-participants would have experienced had they been offered ERA services and incentives. The conditional independence assumption[24] that allows us to directly identify this counterfactual is that given observed attributes X, non-participants would have experienced the same average ERA outcome as participants:

(CIA-1) $E(Y_1 \mid Q=0, X) = E(Y_1 \mid Q=1, X).$

To give (CIA-1) empirical content, we require common support, i.e. overlap in the distribution of the observed characteristics X between participants and non-participants:

(CS) $P(Q=1 \mid X) > 0$ for all X in the support of the eligibles.

Specifically, the experimental evaluation cannot provide estimates of the impact of ERA for individuals with observed characteristics $\tilde{X}$ if no participant displays those values. Thus although there may be eligibles with characteristics $\tilde{X}$, if selection into the ERA experiment is such that nobody with characteristics $\tilde{X}$ is offered ERA or consents to take part so that $P(Q=1|\tilde{X})=0$, the effect for this subset of eligibles is not non-parametrically identified.

Under random assignment (RA) and (CIA-1), identification of $E(Y_1 \mid Q=0)$ is achieved as:

$E(Y_1 \mid Q=0) = E[E(Y_1 \mid Q=0, X) \mid Q=0] = (CIA-1) = E[E(Y_1 \mid Q=1, X) \mid Q=0]$

---

[23] Under two alternative conditions the ATE_1 based on experimental data would still provide unbiased estimates of the ATE even in the presence of a non-negligible share of non-participants. The first case is one of homogeneous ERA impacts, that is $Y_{1i} - Y_{0i} = \beta$ for all eligible individuals i. The second case is one where impacts are heterogeneous, but the decisions of eligibles or caseworkers on participation in the ERA study are not affected by the realised individual gain from receiving ERA, i.e. if $P(Q=1 \mid Y_1 - Y_0) = P(Q=1)$, then $ATE_1 = ATE_0 = ATE$.

[24] Known also as selection-on-observables, unconfoundedness, ignorability or exogeneity. For a recent review of methods relying on this assumption, see Imbens (2004).

$$=(\text{RA})= E[E(Y_1 \mid R=1, X) \mid Q=0] =(\text{CS})= E[E(Y \mid R=1, X) \mid Q=0]$$

where (CS) ensures that there are participants (program group members) for each X for which there are non-participants, so that the last term can be estimated from the data.

As for implementation, each non-participant can be matched to one or more similar program group member(s) based on the propensity score $p(X) \equiv P(Q=0 \mid X)$.

Compared to standard OLS regression, matching methods are non-parametric, allowing both ERA impacts and non-ERA outcomes to depend on observables in arbitrary ways. They additionally highlight the actual comparability of groups by offering ways to assess balancing of observables between matched samples. Like OLS, however, they rule out selection on unobservables. Due to our unique set-up we are however in a position to perform some tests in this respect.

To better understand assumption (CIA-1), note that from the definition of impacts $\beta \equiv Y_1 - Y_0$, we have that $Y_1 = Y_0 + \beta$, so that (CIA-1) is equivalent to assuming:

(CIA-0) $E(Y_0 \mid Q=0, X) = E(Y_0 \mid Q=1, X)$ and

(CIA-β) $E(\beta \mid Q=0, X)$ $= E(\beta \mid Q=1, X)$. (3)

Assumption (CIA-1) is thus made up of two assumptions: no residual selection into the ERA study based on unobserved characteristics affecting non-ERA outcomes (CIA-0) and no residual selection into the study based on unobserved idiosyncratic realised impact components (CIA-β).

Coupled with randomisation, observing the outcomes of the non-participants allows us to directly test condition (CIA-0) that the no-treatment outcomes of the non-participants are the same, on average, as those of observationally equivalent participants. This test is implemented by testing whether $E(Y \mid Q=0, X) = E(Y \mid R=0, X)$, i.e. whether once controlling for observables, the non-participants and the control group (a representative sample of the participants for whom no-treatment outcomes are observed) experience the same average outcome. This test can be performed by running a regression, on the pooled sample of controls and non-participants, of observed outcome on the observables, plus a dummy variable for participation in the ERA study. To minimise all sensitivity to the specification of how the observables should enter the outcome equation and affect differences between the two groups, one can instead perform matching (matching to each non-participant one or more similar control group member) and test for the equality of mean outcomes of the two matched groups. If in the comparison of the outcomes of these two groups there remain significant differences conditional on the observables, this provides evidence of residual selection based on unobservables related to no-treatment outcomes.

The test set-up can further be used to guide the choice of matching method as well as of how to summarise the observables, in particular labour market histories. The idea is to calibrate such decisions on balancing observed outcomes between non-participants and matched controls.

Under (CIA-β), one can thus directly test the validity of the standard (CIA-1) matching as-

sumption by testing (CIA-0); additionally, if (CIA-0) – and hence (CIA-1) – fail, one can correct the matching estimates from selection bias. To see how, consider a violation of (CIA-0), so that non-participants and participants with the same value of X still differ in terms of their average non-ERA outcome by $\alpha(X)$ (note that by how much they differ is allowed to depend on the value of X):

$$E(Y_0 \mid Q=0, X) = E(Y_0 \mid Q=1, X) + \alpha(X) \quad \text{with } \alpha(X) \neq 0 . \tag{4}$$

Consider now average ERA outcomes for participants and non-participants with observables X:

$$E(Y_1 \mid Q=k, X) = E(Y_0 \mid Q=k, X) + E(\beta \mid Q=k, X) \quad \text{for } k=0,1.$$

Because of (4) we now have that:

$$E(Y_1 \mid Q=0, X) = E(Y_0 \mid Q=1, X) + \alpha(X) + E(\beta \mid Q=0, X)$$
$$=(\text{CIA-}\beta)= E(Y_0 \mid Q=1, X) + E(\beta \mid Q=1, X) + \alpha(X) = E(Y_1 \mid Q=1, X) + \alpha(X)$$
$$=(\text{RA})= E(Y \mid R=1, X) + \alpha(X)$$

$$E(Y_1 \mid Q=0) = E[E(Y \mid R=1, X) \mid Q=0] + E[\alpha(X) \mid Q=0]$$

A violation of (CIA-0) thus introduces an overall bias term $E[\alpha(X) \mid Q=0]$ in the matching esti-mate of the average ERA outcome for the non-participants $E(Y_1 \mid Q=0)$ based on the average ob-served ERA outcome of observationally equivalent participants. In our set-up this bias term is how-ever identified in the data. Using the definition of $\alpha(X)$ in (4), we can correct the matching estimate by its bias in terms of no-treatment outcomes:

$$E(Y_1 \mid Q=0) = E(Y \mid Q=0) + E[E(Y \mid R=1, X) \mid Q=0] – E[E(Y \mid R=0, X) \mid Q=0].$$

The expression for the average impact for the non-participants under (CIA-β) simplifies to:

$$ATE_0 = E[E(Y \mid R=1, X) \mid Q=0] – E[E(Y \mid R=0, X) \mid Q=0].$$

Estimation of the adjusted ATE_0 can be carried out by matching the non-participants twice, once to the program group and once to the control group. Common support should be imposed on the non-participants across both terms.

Irrespective of whether the (CIA-0) test is passed, the impact of ERA for the eligibles can thus be identified despite randomisation bias under the (CIA-β) assumption of no selection into the ERA study based on realised unobserved individual gains, once allowing for arbitrarily heterogeneous impacts based on a rich set of observed characteristics. [25] This assumption is quite plausible in the ERA set-up. As highlighted by the qualitative research (see Section 2.2), formal refusers (like in-deed consenters) had no substantive knowledge of what ERA was or would entail, hence no possi-bility to try to predict their idiosyncratic gain from it. As to diverted customers, the qualitative evi-

---

[25] Equation (2) highlights how only $E(Y_1 \mid Q=0)$ needs to be identified. One could nonetheless ignore this fact and iden-tify the ATE_0 directly using (CIA-β), obtaining as expected the same expression in the main text:
$ATE_0 =(\text{CIA-}\beta)= E[E(Y_1 – Y_0 \mid Q=1, X) \mid Q=0] =(\text{RA})= E[E(Y \mid R=1, X) \mid Q=0] – E[E(Y \mid R=0, X) \mid Q=0].$
Indeed, one could further ignore that only (part of) the ATE_0 needs to be identified and identify the ATE directly:
$ATE =(\text{CIA-}\beta)= E[E(Y_1 – Y_0 \mid Q=1, X)] =(\text{RA})= E[E(Y \mid R=1, X)] – E[E(Y \mid R=0, X)].$ Estimation of the ATE can be carried out by matching the eligibles twice, once to the program group and once to the control group.

dence is that the main source of diversion was an incentive structure causing advisers to divert based on the outcome they predicted for the job-seeker. It is possible of course that a minority of advisers diverted those whom they thought would not benefit from ERA, but in this case the question is how successful they were, on average, in predicting the unobserved individual-specific impact component, over and above impact heterogeneity based on our rich set of observables and which our estimation methods leave completely unrestricted. Given also that ERA was a completely new treatment for advisors as well (who never before had to support job-seekers once they had entered work) and given how research has shown that caseworkers are not particularly good at systematically predicting outcomes and indeed counterfactual outcomes (e.g. Frölich, 2001 and Lechner and Smith, 2007), the assumption of no refusal nor diversion based on realised unobserved idiosyncratic ERA impacts would seem particularly plausible in this case.

In presenting the results it is convenient to have a common metric across groups and outcomes (days in employment, employment probability and, for some subgroups, earnings); any violation of (CIA-0) can be expressed as a fraction $\theta_0 \equiv E(Y \mid Q{=}0)/E[E(Y \mid R{=}0, X) \mid Q{=}0]$, i.e. non-participants are found to experience a fraction θ_0 of the average non-ERA outcome of their observationally equivalent non-participants. The value of θ_0 reveals different types of selection processes: if $\theta_0{<}1$ ($\theta_0{>}1$), non-participants would have experienced on average lower (higher) outcomes under ERA than observationally equivalent participants, while θ_0 equals one when (CIA-0) and hence – provided (CIA-β) is met – (CIA-1) are satisfied. θ_0 and $\alpha_0 \equiv E[\alpha(X) \mid Q{=}0]$ are related as $\theta_0 = \alpha_0/E[E(Y \mid R{=}0, X) \mid Q{=}0] + 1$, where obviously $\theta_0{=}1$ if and only if $\alpha_0{=}0$.

4.3 No follow-up information on the non-participants (survey outcomes)

In some situations only survey outcome information might be available; in the case of ERA, administrative earnings became available only later on in the evaluation. Even then, administrative earnings have a much less clean definition, both as some components are not captured and as they pertain to different amounts of time on the program for different individuals; indeed for a subgroup of the eligibles such information is pre-treatment (see Section 3.3).

Focus on survey outcomes raises two additional issues: not only treatment but now also no-treatment outcomes of the non-participants are unobserved, and in the presence of non-random survey/item non-response among participants, ATE_1 itself will in general be unobserved. In case of survey outcomes, only p is directly identified in equation (1): $ATE = (1{-}p){\cdot}ATE_1 + p{\cdot}ATE_0$.

What is also identified in the data is the experimental contrast on the responding participants, $\Delta_{S=1} \equiv E(Y \mid S{=}1, R{=}1) - E(Y \mid S{=}1, R{=}0)$, which will not necessarily be equal to ATE_1.

This problem is akin to attrition and involves reweighing the outcomes of the responding participants (responding program and control groups) on the basis of the characteristics X of the full

eligible group (i.e. full program group, full control group and non-participants) to make them representative – in terms of observables X – of the full eligible population.[26]

Assume that, once conditioning on observables X, eligibles do not select into the ERA study based on their realised idiosyncratic unobserved impact component:

(CIA-β) $E(Y_1 - Y_0 \mid Q=1, X) = E(Y_1 - Y_0 \mid Q=0, X)$

We allow for selective non-response, provided selection into the responding sample happens only in terms of observable characteristics:

(NR) $E(Y_1 \mid R=1, S=1, X) = E(Y_1 \mid R=1, S=0, X)$ and

$E(Y_0 \mid R=0, S=1, X) = E(Y_0 \mid R=0, S=0, X)$

Assumption (NR) rules out selection on outcome-relevant unobservables into responding to the earnings question given random assignment status. In other words, conditional on random assignment status and characteristics X, non-response is unrelated to potential outcomes, i.e. program (control) group members with characteristics X who respond and who don't respond would experience on average the same ERA (non-ERA) outcome.

Under random assignment (RA), (CIA-β) and (NR), identification of ATE is achieved as[27]:

$ATE \equiv E(Y_1 - Y_0) = E[E(Y_1 - Y_0 \mid X)] =$(CIA-$\beta$)$= E[E(Y_1 - Y_0 \mid Q=1, X)]$

$=$(RA)$= E[E(Y_1 \mid R=1, X)] - E[E(Y_0 \mid R=0, X)]$

$=$(NR)$= E[E(Y_1 \mid R=1, S=1, X)] - E[E(Y_0 \mid R=0, S=1, X)]$

$= E[E(Y \mid R=1, S=1, X)] - E[E(Y \mid R=0, S=1, X)]$ (3)

To derive the empirical counterpart we consider weighting and matching estimators. The former directly weights the outcomes of the (responding) participants so as to reflect the distribution of observables in the original eligible population (see Appendix 1 for the derivation):

$ATE = E[\omega_1(X) \cdot S \cdot R \cdot Y - \omega_0(X) \cdot S \cdot (1-R) \cdot Y]$,   where

$\omega_k(X) \equiv \dfrac{P(Q=1)}{P(Q=1|x)} \dfrac{P_{RS|Q}(k,1|1)}{P_{RS|Q,X}(k,1|1,x)}$ for $k=0, 1$

Alternatively, the weights can be constructed via matching[28], with the advantages that the exact specifications of the propensity score and response probabilities are not needed and that one can assess the extent of the actual comparability achieved between groups.

---

[26] See Wooldridge (2002) for weighting estimators to deal with incidental truncation problems such as attrition under the CIA and Huber (2012) for weighting estimators to deal with different forms of attrition in randomised experiments.

[27] An alternative set of assumptions to (RA) and (NR) yielding the same expression for the ATE are the external validity of the impact for respondents given X, $E(Y_1–Y_0|Q=1,X) = E(Y_1–Y_0|Q=1,S=1,X)$, and that random assignment keeps holding given X within the responding sample, $E(Y_k|S=1,R=1,X) = E(Y_k|S=1,R=0,X)$ for $k=0,1$.

[28] To derive the terms $E[E(Y \mid R=k, S=1, X)]$ for $k=0,1$, match each eligible individual in the $Q=0$ and $Q=1$ groups, to individuals in the subgroup of responding $R=k$ members and calculate the weight that gets assigned to each individual in the latter subgroup (this weight will be larger than one). Reweigh the outcomes in the latter subgroup using these

Sensitivity analysis

We have proposed exploiting the experiment to test for the presence of unobserved characteristics driving selection into the ERA study when outcome data is available for all eligibles. While this is not the case for survey outcomes, in the ERA evaluation we can nonetheless consider two specific subgroups for whom some robustness analysis can meaningfully be carried out.

The "***post-April group***" is made up of those eligibles who started the New Deal or ERA from April 2004 onwards. For these individuals, representing 35% of ND25+ and 41% of NDLP eligibles, the 2004/05 fiscal year administrative earnings data represent outcomes (see Figure 1). This group thus offers the chance to carry out the (CIA-0) test in terms of (administrative) earnings. Additionally, it can be used to glean guidance on how best to construct the set of matching variables X, as the way of summarising labour market histories that produces the best balancing in the (CIA-0) test can then be used in the weighting and matching estimators for survey earnings. Of course, both uses of this subgroup potentially suffer from an issue of external validity.

The "***March-May group***" is made up of those eligibles who started the New Deal or ERA around the start of the 2004/05 tax year, which we approximate as the three months March to May 2004. For these individuals, representing 25% of both ND25+ and NDLP eligibles, tax year 2004/05 earnings closely correspond to earnings in the 1-year follow up period, in that they cover (roughly) the same horizon (see Figure 1). This subgroup too lends itself to testing (CIA-0) on administrative earnings.[29] Furthermore, under the weak assumption (CIA-β), we could take the *ATE* for this group in terms of administrative earnings as the 'truth', and check against it the performance of the proposed matching and weighting estimators for survey-measured earnings, which in addition to selection into the study have to deal with non-response. Specifically, we can compare

---

weights and take their average over this subgroup, i.e. use the matched outcome to estimate $E(Y_k)$. One can match on the basis of the propensity score $P(R=k \ \& \ S=1 \mid Q=0 \lor Q=1, X)$.

[29] Before turning to the issue of non-participation, one can focus on the experimental sample and perform the following additional tests on the March-May group (after having checked that, as one would expect, randomisation still holds, i.e. at least the observables are balanced between the program and control March-May subgroups):

(A) If the experimental contrast in terms of administrative earnings for the survey respondents among the March-May group is not significantly different from the experimental impact estimate in terms of administrative earnings for the full March-May group, there is no evidence of non-response bias (for the March-May group) in terms of characteristics that affect administrative earnings. Specifically, there is evidence of internal validity (the responding program and control group members have maintained comparability to one another so that the experimental contrast recovers the average impact for respondents) and of external validity (the impact for the responding March-May subsample is representative of the impact for the full March-May sample).

(B) If the experimental contrast in terms of administrative earnings for the survey respondents among the March-May group is not significantly different from the experimental contrast in terms of survey earnings for the respondents among the March-May group, there is evidence that administrative and survey earnings essentially measure the same impact despite not necessarily covering the same components.

(C) Given the above results that administrative and survey measures essentially measure the same impact and there is no non-response bias, the experimental impact estimate in terms of administrative earnings for the full March-May group should not be significantly different from the experimental contrast in terms of survey earnings for the respondents among the March-May group.

the *ATE* estimate for the March-May group in terms of administrative earnings to the *ATE* estimate for the March-May group in terms of survey earnings, which was derived from its responding sub-group taking account of non-response. While potentially informative, this sensitivity check might at best provide corroborative evidence. First, while the subgroup was chosen to align the horizons over which the two types of earnings are measured, nothing can be done to force the two measures to capture exactly the same components (though some evidence can be gleaned from test (B) in foot-note 29). Additionally, there could once again be external validity issues in extending any conclusion from the May-March group to the full sample. Finally, implementation-wise the group might be too small to allow one to discriminate with enough precision between different estimates. Widening the temporal window beyond three months to define a larger group would yield a gain in precision but also result in increasingly different horizons covered by administrative and survey earnings, reflecting the standard trade-off between bias and variability.

5. Implications of non-participation for the experimental estimates

This section presents all empirical results, first those relating to employment outcomes measured by administrative data (Section 5.1), then those relating to yearly earnings measured, for the most part, by survey information (Section 5.2).

An overarching comment which applies to all our results is that matching has always performed extremely well in balancing the observables, both when estimating impacts (see table A3 in the Appendix) and when assessing the (CIA-0) condition. Also, while common support was always imposed, it never led to the loss of more than 1-2% of the group of interest.

5.1 Employment

ND25+ group

The first column of Table 4 presents the benchmark experimental on the average ERA impact for ND25+ participants (ATE_1) in terms of time in employment and employment probability in the first follow-up year. The table displays both the raw experimental contrast and the regression-adjusted estimate controlling in various ways for the observables in Table 2. Although randomisation has worked very well in the ERA experiment so that the program and control groups are well-balanced in terms of such characteristics, controlling for them can increase the precision of the experimental estimate by reducing the residual outcome variance. This seems to be largely the case, particularly for days in employment, where the experimental impact becomes significant following the regression adjustment. Furthermore, the adjustment allows one to control for differences in observables between program and control groups that have occurred by chance. This also seems to matter some-

Table 4: Employment outcomes for ND25+: Experimental point estimates of the average impact for participants (ATE_1) and residual bias in terms of non-ERA outcomes for different ways of constructing labour market histories

| | ATE_1 | CIA-0 test | | θ_0 |
		OLS	Matching	
DAYS EMPLOYED				
Raw	4.0	-9.4***		0.834
All other X's plus				
summary	4.6*	-7.9***	-9.7***	0.829
monthly employment	4.8**	-7.6***	-9.4***	0.835
ever employment	5.0**	-7.6***	-9.4***	0.835
sequence	4.8**	-7.9***	-8.8***	0.843
summary + monthly employment	4.8**	-7.7***	-9.2***	0.837
summary + ever employed	5.0**	-7.7***	-9.3***	0.837
summary + sequence	4.8**	-8.0***	-8.8***	0.843
EVER EMPLOYED				
Raw	0.014	-0.062***		0.808
All other X's plus				
summary	0.017	-0.044***	-0.056***	0.825
monthly employment	0.017	-0.043***	-0.053***	0.831
ever employment	0.019*	-0.042***	-0.053***	0.831
sequence	0.017	-0.043***	-0.052***	0.835
summary + monthly employment	0.017	-0.044***	-0.052***	0.835
summary + ever employed	0.019*	-0.043***	-0.053***	0.831
summary + sequence	0.017	-0.044***	-0.053***	0.833

Notes: 'Raw' are outcome differences between non-participants and participants. 'OLS' and 'Matching' are adjusted differences. Non-participants are observed to experience a fraction θ_0 of the average non-ERA outcome of observationally equivalent participants. See Section 3.3 for the description of how labour market histories have been constructed.

what, with point estimates always increasing once conditioning on observables.

A small positive effect of ERA of an extra 4 to 5 days in employment has been uncovered for the participants, while their employment chances in the follow-up year have remained basically unaffected. But what effect would the full eligible group have experienced, on average?

Before turning to this question, we consider the results from testing the (CIA-0) condition that, controlling for our rich set of observables, participants and non-participants experience the same average non-ERA outcome. Table 4 reports the OLS and matching results from comparing the outcomes of the two groups conditional on different ways of summarising labour market histories, as well as the corresponding θ_0. The overall conclusions are twofold. First, there remain large and significant imbalances in employment outcomes, with non-participants being on average 8-10 fewer days in employment and 4-5pp less likely to be employed than observationally equivalent participants. Second, how past labour market history is measured makes no difference at all. In contrast to Dolton and Smith (2011), more detailed, sophisticated and flexible ways of capturing histories do not yield *any* gain in making non-participants and controls look similar in terms of their no-treatment outcome. As we will see, both these conclusions apply for the NDLP group as well (see

Table 7).[30] The claim often made in the literature (see e.g. Dehejia and Wahba, 1999, Heckman and Smith, 1999, Heckman *et al.*, 1998, Heckman *et al.*, 1999, and Frölich, 2004, and to some extent Hotz *et al.*, 2005) that histories variables can capture labour-market relevant unobservables is thus not borne out in our data, at least in the no-treatment case, which is indeed the case of interest when using non-experimental comparison groups for estimating treatment effects.

To shed further light, Table 5 reports the results of the same test performed at the district level, together with the district-level experimental impact and the incidence of non-participation, diverted customers and formal refusers (cf. Table 1). Interestingly, the overall imbalance in terms of days employed is purely driven by NE England, the district in which a striking 26% of ND25+ entrants formally refused to participate in the experiment. The non-balancing overall in terms of employment probability is found to be driven by NE England again, as well as by East Midlands – the two districts with the highest incidence of non-participants (34.4% and 27.5%).

Once these 'offending' districts are excluded, the remaining pooled districts still display some significant differences in outcomes between non-participants and controls; controlling for observables does however make these differences vanish. Incidentally, the reported results are based on the parsimonious summary of labour market histories; as was the case for the overall group, district-level imbalances were not sensitive at all to the way histories are constructed, and for NE England and especially East Midlands, controlling for histories and other observables did not reduce the raw difference by much, if at all.

Table 6 thus presents the experimental ERA impact estimate for the participants (ATE_1) and the non-experimental estimates for the non-participants (ATE_0) and the full eligible group (ATE) in different sets of districts: first, in those districts which passed the (CIA-0) test, then in all districts and finally in the individual districts which failed to pass the (CIA-0) test. In the latter two cases, we present estimates of the ATE_0 (and hence of the ATE) that ignore the mismatch in non-ERA outcomes between non-participants and observationally similar participants, as well as estimates that have been adjusted to correct for such mismatch.[31]

On the pooled districts which pass the (CIA-0) test, the effects for the participants and for all eligibles are found to be very similar and not significantly different from one another.

If we ignored that the (CIA-0) test failed overall and only corrected for differences in observed characteristics between participants and non-participants in estimating the effect of ERA on the full eligible population, we would conclude that the experimental estimate significantly underes-

---

[30] For the subgroup of eligibles randomised or flowing in after April 2004, administrative earnings from the 2003/04 fiscal year represent pure *pre*-treatment information. We assessed whether the addition of pre-treatment earnings could help in passing the (CIA-0) test, but again it did not make any difference.

[31] The overall ATE is calculated using a district-weighted average of the estimators for the ATE_0, where the correction has only been applied to the district-specific ATE_0's of the districts failing the (CIA-0) test.

Table 5: ND25+: Non-participation probabilities (p), experimental impact (ATE_1) and (CIA-0) test results by district

| | p | *Formal refusers* | *Diverted customers* | ATE_1 | CIA-0 test | | | | θ_0 |
					Raw	θ_{raw}	**OLS**	**Matching**	
DAYS EMPLOYED									
All	23.0	*13.6*	*9.4*	4.6*	4.0	0.834	-7.9***	-9.7***	0.829
Scotland	8.7	*8.7*	*0.0*	8.6	-17.3	0.690	-8.2	-8.3	0.828
NE England	34.9	*26.1*	*8.8*	-10.3	-33.9***	0.565	-23.6***	-27.8***	0.616
NW England	14.6	*14.6*	*0.0*	7.5	-7.0	0.864	-1.8	-3.0	0.937
Wales	20.7	*11.1*	*9.6*	-13.6	-12.0	0.816	-16.3	-7.9	0.864
East Midlands	27.5	*10.7*	*16.8*	8.0	-4.3	0.934	-5.7	-7.7	0.885
London	25.8	*11.1*	*14.8*	8.9**	-3.6	0.915	-3.7	-2.8	0.932
no NE Eng	21.0	*11.5*	*9.5*	6.5***	-5.7*	0.894	-4.7	-5.3	0.901
EVER EMPLOYED									
All	23.0	*13.6*	*9.4*	0.017	-0.062***	0.808	-0.044***	-0.056***	0.825
Scotland	8.7	*8.7*	*0.0*	0.047	-0.096*	0.726	-0.039	-0.041	0.861
NE England	34.9	*26.1*	*8.8*	-0.036	-0.191***	0.541	-0.149***	-0.172***	0.571
NW England	14.6	*14.6*	*0.0*	0.033	-0.024	0.915	0.010	0.010	1.038
Wales	20.7	*11.1*	*9.6*	-0.035	-0.027	0.923	-0.017	-0.004	0.987
East Midlands	27.5	*10.7*	*16.8*	0.031	-0.073**	0.817	-0.060**	-0.071**	0.819
London	25.8	*11.1*	*14.8*	0.022	-0.017	0.929	-0.009	-0.010	0.958
no NE,EM	18.8	*11.8*	*7.0*	0.023*	-0.041**	0.858	-0.007	-0.011	0.956

Notes: 'Raw' are outcome differences between non-participants and participants. 'OLS' and 'Matching' are adjusted differences. Non-participants are observed to experience a fraction θ_{raw} of the average non-ERA outcome of participants and a fraction θ_0 of the average non-ERA outcome of observationally equivalent participants. Estimates shown control for parsimonious history summaries.

Table 6: Employment outcomes for ND25+: Average ERA impacts for participants (ATE_1), non-participants (ATE_0) and non-participants, as expected, falls and the adjusted estimates for the ATE fall back in line all eligibles (ATE)

	p	ATE_1	ATE_0	ATE	$ATE_1 \neq ATE$
DAYS EMPLOYED					
All but NE England	0.210	6.5**	9.7***	7.2***	no
All districts	0.230	4.6*	10.1***	5.9***	*
All districts, adjusted			6.3**	5.0**	no
NE England	0.349	-10.3	8.1	-3.9	**
NE England, adjusted			-15.4*	-12.1	no
EVER EMPLOYED					
All but NE England and E Midlands	0.188	0.023*	0.026*	0.024**	no
All districts	0.230	0.017	0.056***	0.026**	***
All districts, adjusted			0.007	0.015	no
NE England	0.349	-0.036	0.092**	0.009	***
NE England, adjusted			-0.062**	-0.045	no
E Midlands	0.275	0.031	0.083***	0.045**	*
E Midlands, adjusted			0.013	0.026	no

Notes: Kernel matching with epanechnikov kernel (bandwidth of 0.06); statistical significance based on boot-strapped bias-corrected confidence intervals (1000 replications); $ATE_1 \neq ATE$: bootstrap-based statistical significance of the difference; *** significant at 1%, ** at 5%, * at 10%.

timates how much ERA can improve the employment outcomes of all eligibles. Specifically, non-participants would appear to enjoy more than double an increase in days employed (10) than do participants (4.6), resulting in an overall ATE of 6 days, significantly different from ATE_1. Similarly, the effect for the non-participants would appear to be a highly significant, 5.6pp increase in employment probability, compared to an insignificant 1.7 increase for participants. The ATE for all eligibles would correspondingly have been a significant increase of 2.6pp, again significantly different from ATE_1.

When by contrast the correction is applied to reflect failure of the test, the effect for the non-participants falls (as expected, given the sign of the residual imbalance in the test), and the adjusted estimates of the ATE fall back in line with the corresponding estimates of the ATE_1.

Exactly the same pattern is found for the individual districts which fail to pass the (CIA-0) test: ignoring failure of the test would lead to the wrong conclusion that that the effect for the participants significantly underestimates the effect for the eligibles, while appropriately adjusting the estimates once again confirms that the impact for the participants is representative of the impact that ERA would have had on its full eligible group.

NDLP group

Table 7 displays the experimental estimates for the NDLP participants in terms of employment outcomes. As was the case for the ND25+ group, no experimental impact could be detected on the probability of being employed in the follow-up year; for the NDLP participants, however, employment durations too have remained completely unaffected.

Before turning to our estimates for all eligibles, we again consider the results of the (CIA-0) test (Table 7). Perhaps surprisingly, for both employment outcomes, there are no raw differences in the average no-treatment outcomes of non-participants and participants. Non-participants are a non-significant 3.8 more days in employment and 0.01pp more likely to be employed during the follow-up year than controls. Large and significant differences however emerge once controlling for observables: non-participants are now 10-12 fewer days and 4pp less likely to be employed than observationally equivalent controls. One must however control for relevant pre-treatment characteristics as there are significant imbalances in the raw groups, e.g. 21.7% of non-participants are employed (and 13.1% are not on benefits) at inflow, compared to only 13.3% (and 7%) of participants, and 47.8% of non-participants were never employed in the 3 pre-inflow years against 50.5% of participants (see also Appendix Table A1 for marginal effects).

Table 8 shows that the overall result of no differences in raw outcomes masks a striking diversity by district, with positively selected non-participants in NW England 'cancelling out' the negatively selected non-participants in Scotland. Interestingly, these two districts experienced only a small incidence of non-participation (5-6% of all eligibles), but as we see have excluded a highly selective group. Non-participants in the other districts represented a far higher proportion of the eligibles (24%, 29%, 31% and 47%), however they did not experience significantly different non-ERA outcomes, on average, from the participants.

We exclude those districts where balancing the observables using matching fails the (CIA-0) test: NE England, and for days employed, Scotland as well. We also exclude the East Midlands in both cases, as inclusion of its negatively selected participants representing almost half of the eligibles would result in matching failing the test at any significance level.

Table 9 presents estimates of the three causal parameters of interest for the set of districts that passed the (CIA-0) test, for all districts and for individual districts that failed the test, in the latter two cases also showing the non-experimental estimates corrected for their bias in terms of the mean no-treatment outcome.

Table 7: Employment outcomes for NDLP: Experimental point estimates of the average impact for participants (ATE_1) and residual bias in terms of non-ERA outcomes for different ways of constructing labour market histories

| | ATE_1 | CIA-0 test | | θ_0 |
		OLS	Matching	
DAYS EMPLOYED				
Raw	-0.1		3.8	1.033
All other X's plus				
summary	-2.2	-10.4***	-11.2**	0.914
monthly employment	-2.4	-10.2***	-10.2**	0.921
ever employment	-2.5	-11.0***	-12.1**	0.907
sequence	-2.4	-10.8***	-11.7**	0.910
summary + monthly employment	-2.7	-10.6***	-11.1**	0.915
summary + ever employed	-2.2	-10.8***	-12.4**	0.906
summary + sequence	-2.1	-10.4***	-11.2**	0.914
EVER EMPLOYED				
Raw	0.003		0.004	1.009
All other X's plus				
summary	-0.006	-0.041***	-0.041**	0.928
monthly employment	-0.007	-0.040***	-0.042**	0.925
ever employment	-0.007	-0.041***	-0.042**	0.925
sequence	-0.007	-0.043***	-0.044**	0.922
summary + monthly employment	-0.008	-0.042***	-0.044**	0.922
summary + ever employed	-0.007	-0.042***	-0.045***	0.921
summary + sequence	-0.007	-0.042***	-0.043**	0.924

Notes: 'Raw' are outcome differences between non-participants and participants. 'OLS' and 'Matching' are adjusted differences. Non-participants are observed to experience a fraction θ_0 of the average non-ERA outcome of observationally equivalent participants. See Section 3.3 for the description of how labour market histories have been constructed.

The story that emerges for the NDLP group is quite compelling, irrespective of the adjustment: the employment effect of ERA in terms of either employment probability or duration would have been the same – and statistically indistinguishable from zero – for the non-participants as for the experimental group. For the NDLP group, the experimental estimate of no ERA impact on employment outcomes is thus representative of the average effect for all eligibles. (The only exception is the negative impact on employment durations in East Midlands, but again the evidence is that participants, non-participants and all eligibles would have experienced the same effect).

Table 8: NDLP: Non-participation probabilities (p), experimental impact (ATE_1) and (CIA-0) test results by district

	p	Formal refusers	Diverted customers	ATE_1	Raw	θ raw	CIA-0 test OLS	Matching	θ_0
DAYS EMPLOYED									
All	30.4	4.0	26.4	-2.2	3.8	1.003	-10.4***	-11.2**	0.914
Scotland	5.3	2.8	2.5	9.6	-75.0***	0.478	-71.1***	-64.2**	0.490
NE England	29.2	1.0	28.2	0.0	2.7	1.023	-14.7	-18.8*	0.864
NW England	6.2	3.7	2.5	21.1**	38.4*	1.336	31.6*	27.6	1.224
Wales	23.6	3.6	20.1	-16.6	20.3	1.141	-4.9	-7.6	0.955
East Midlands	47.1	5.9	41.2	-15.5**	4.9	1.044	-11.1*	-10.7	0.916
London	31.0	4.9	26.1	-3.5	12.9	1.127	-3.4	-6.4	0.947
no Sctl, NE, EMidls	23.4	4.3	19.1	-1.9	13.3*	1.117	-2.9	-9.2	0.931
EVER EMPLOYED									
All	30.4	4.0	26.4	-0.006	0.004	1.009	-0.041***	-0.041**	0.928
Scotland	5.3	2.8	2.5	0.041	-0.130	0.786	-0.063	-0.056	0.895
NE England	29.2	1.0	28.2	-0.020	-0.003	0.994	-0.063**	-0.071*	0.880
NW England	6.2	3.7	2.5	0.063*	0.165**	1.319	0.130*	0.130	1.242
Wales	23.6	3.6	20.1	-0.044	0.031	1.049	-0.052	-0.038	0.946
East Midlands	47.1	5.9	41.2	-0.036	-0.001	0.998	-0.049**	-0.043	0.923
London	31.0	4.9	26.1	0.000	0.046	1.105	-0.026	-0.030	0.942
no NE, EMidls	21.3	4.1	17.2	0.009	0.033	1.066	-0.018	-0.030	0.947

Notes: 'Raw' are outcome differences between non-participants and participants. 'OLS' and 'Matching' are adjusted differences. Non-participants are observed to experience a fraction θ_{raw} of the average non-ERA outcome of participants and a fraction θ_0 of the average non-ERA outcome of observationally equivalent participants. Estimates shown control for parsimonious history summaries.

Table 9: Employment outcomes for NDLP: Average ERA impacts for participants (ATE_1), non-participants (ATE_0) and all eligibles (ATE)

	p	ATE_1	ATE_0	ATE	$ATE_1 \neq ATE$
DAYS EMPLOYED					
All but Scotland, NE Eng, EMidls	0.234	-1.9	-3.8	0.5	no
All districts	0.304	-2.2	-2.1	-2.2	no
All districts, adjusted			-8.8*	-4.2	no
Scotland	0.053	9.6	72.1	12.9	no
Scotland, adjusted			19.4	10.1	no
NE England	0.292	0.0	5.7	1.7	no
NE England, adjusted			-14.5	-4.2	*
East Midlands	0.471	-15.5**	-4.4	-10.3	no
East Midlands, adjusted			-15.3*	-15.4**	no
EVER EMPLOYED					
All but NE Eng, EMidls	0.213	0.011	0.007	0.010	no
All districts	0.304	-0.006	0.015	0.000	no
All districts, adjusted			-0.007	-0.007	no
NE England	0.292	-0.020	0.033	-0.005	no
NE England, adjusted			-0.040	-0.026	no
East Midlands	0.471	-0.036	0.020	-0.009	**
East Midlands, adjusted			-0.022	-0.029	no

Notes: Kernel matching with epanechnikov kernel (bandwidth of 0.06); statistical significance based on boot-strapped bias-corrected confidence intervals (1000 replications); $ATE_1 \neq ATE$: bootstrap-based statistical significance of the difference; *** significant at 1%, ** at 5%, * at 10%.

5.2 Earnings

For both intake groups, the experiment highlights a sizeable and statistically significant gain in terms of average earnings in the first follow-up year: £445 for the ND25+ group and an even more substantial £788 for the NDLP group (see Table 10). These adjusted experimental contrasts are based on the survey sample with non-missing earnings information. Slightly less than half (49%) of the New Deal ERA study participants were randomly selected to take part in the first-year follow-up survey. Not all the selected individuals could however be located, accepted to be take the survey, or could be interviewed. Response rates remained high though: 87% among the NDLP and 75% among the ND25+ fielded samples. Of these respondents, 10% have however missing information on yearly earnings. Thus, for only one third of all ERA study participants do we observe earnings (31% in the ND25+ and 35% in the NDLP group). It thus follows that earnings information is available for one quarter of the ERA eligibles (23.6% of the ND25+ and 24.1% of the NDLP eligibles).

The survey sample was randomly chosen, and while there is good evidence (see Dorsett *et al.*, 2007, Appendix G) that the respondents to the survey did not differ dramatically from

Table 10: Survey earnings: Experimental contrast for respondents ($\Delta_{S=1,X}$) and impact on all eligibles (*ATE*)

		ND25+		NDLP	
$\Delta_{S=1,X}$		445.4**	$\Delta_{S=1,X} \neq ATE$	788.1***	$\Delta_{S=1,X} \neq ATE$
ATE	Weighting	579.6**	not sig	762.1***	not sig
	Matching	551.2***	not sig	708.5***	not sig

Notes: $\Delta_{S=1,X}$ is the experimental contrast ignoring potential non-response bias, adjusted for *X*.
Matching estimator: kernel matching with epanechnikov kernel (bandwidth of 0.06), estimates pertain to those non-participants satisfying both support conditions. Statistical significance based on bootstrapped bias-corrected confidence intervals (1000 replications): *** significant at 1%, ** at 5%, * at 10%.

the non-respondents – both in terms of baseline characteristics and administrative outcomes – no analysis has been performed on item non-response, i.e. on those 10% of survey sample members who did not respond to the earnings question. In our definition of non-respondents we have lumped survey and item non-respondents, since impact estimates on earnings can only be obtained for our narrower definition of respondents.

To derive estimates of the impact of ERA for all eligibles in terms of survey-based earnings, we thus apply the weighting and matching approaches accounting for non-response outlined in Section 4.3. Table 10 compares the estimated impact for the eligible population to the regression-adjusted experimental contrast calculated on the responding participants.

Once non-response and non-participation are taken into account using either method, point estimates increase for the ND25+ group and remain largely stable for the NDLP group. The two non-experimental methods produce quite close point estimates to each other, which do not significantly differ from the adjusted experimental contrast on respondents. For either intake group, we thus find that the sizeable and significant earnings impact uncovered for survey respondents extends to all eligibles.

In the case of survey outcomes, in addition to the arguably weak assumption of no selection into the study based on the realised unobserved idiosyncratic gain (once allowing for arbitrarily heterogeneous impacts according to observed characteristics), we have to invoke additional assumptions about the response process. We now turn to presenting the results from the sensitivity analyses we have suggested based on two special subgroups (see Table 11).

For both the Post-April and March-May inflow subgroups of the ND25+ and NDLP intake groups, the (CIA-0) test in terms of administrative earnings is passed, i.e. no significant differences in non-ERA earnings remain between non-participants and matched participants.[32]

---

[32] The way of summarising labour market histories for the Post-April group that produced the best balancing was then used to obtain the estimates in terms of survey earnings for the full sample in Table 10.

Table 11: Sensitivity analyses for earnings outcomes

(i) **ND25+**

(CIA-0) test in terms of 2004/05 earnings (admin)

	History	Raw	θ raw	OLS	Matching	θ_0	N
Post-April group	monthly employment	-147	0.937	-240	-208	0.910	2,723
March-May group	summary+month. emp.	-465*	0.776	-275	-109	0.938	1,935

Full March-May group

		p	ATE_1	ATE_0	ATE_{adm}	$ATE_1 \neq ATE_{adm}$
(a) 2004/05 earnings (admin)		0.248	183.9	531.7**	270.2	not sig

			$\Delta_{S=1,X} \neq ATE_{surv}$	$ATE_{adm} \neq ATE_{surv}$
(b) annual earnings (survey)				
$\Delta_{S=1,X}$		273.1		
ATE_{surv}	Weighting	819.6	not sig	not sig
	Matching	700.4**	not sig	not sig

(ii) **NDLP**

(CIA-0) test in terms of 2004/05 earnings (admin)

	History	Raw	θ raw	OLS	Matching	θ_0	N
Post-April group	summary	210	1.087	-82	-69	0.976	3,002
March-May group	summary	323	1.132	-10	52	1.019	1,845

Full March-May group

		p	ATE_1	ATE_0	ATE_{adm}	$ATE_1 \neq ATE_{adm}$
(a) 2004/05 earnings (admin)		0.320	375.9	621.8	454.7*	not sig

			$\Delta_{S=1,X} \neq ATE_{surv}$	$ATE_{adm} \neq ATE_{surv}$
(b) annual earnings (survey)				
$\Delta_{S=1,X}$		736.1		
ATE_{surv}	Weighting	759.9	not sig	not sig
	Matching	566.0	not sig	not sig

Notes: 'Raw' are earnings differences between non-participants and participants. 'OLS' and 'Matching' are adjusted differences. Non-participants are observed to experience a fraction θ_{raw} of the average non-ERA earnings of participants and a fraction θ_0 of the average non-ERA earnings of observationally equivalent participants. Incidence of non-participation is p; ATE_1 is the average impact for participants, ATE_0 for non-participants and ATE (either in terms of administrative or survey earnings) for all eligibles.

Table 11(a) shows that ERA has increased average earnings for participants, non-participants and all eligibles among the March-May group, though only the estimates for ND25+ non-participants and for all NDLP eligibles manage to reach statistical significance. What is of interest, however, is that the impact for participants in terms of administrative earnings is representative of the impact for all eligibles. This has indeed been a recurrent finding: for administrative outcomes that either directly pass the (CIA-0) test or, if they fail it, once the correction is applied, the ATE_1 is not significantly different from the ATE.

The March-May group lends itself to a more direct robustness check as this is the sub-group for whom fiscal year earnings in 2004/05 correspond to yearly earnings in the 1-year follow up period, the same horizon covered by survey earnings.[33] Assumption (CIA-β) identifies the *ATE* for the March-May group in terms of administrative earnings (ATE_{adm}). Under (CIA-β) and assuming that administrative and survey earnings covering the same horizon essentially measure the same impact (for which we found support as for the survey respondents among the March-May group the experimental contrast in terms of administrative earnings is not significantly different from the one in terms of survey earnings), we can assess how well the proposed matching and weighting estimators based on survey respondents deal with non-response by comparing their earnings estimate of the *ATE* for the March-May group (ATE_{surv}) to the *ATE* estimate for the March-May group in terms of administrative earnings (ATE_{adm}).

Table 11(b) reports the results of this analysis, which unfortunately are not particularly compelling given that the small size of this subgroup (25% of the eligibles) coupled with the use of non-parametric methods makes it difficult to reach statistical significance. The estimates for the March-May group are positive but mostly insignificant. As to our robustness checks, none of the non-experimental estimates based on survey earnings is significantly different from the *ATE* estimated from fiscal year administrative data, or indeed from the adjusted experimental contrast for survey respondents. Implementation-wise the group is too small to allow to discriminate with enough precision between estimates using different non-parametric methods, which highlights the price to be paid in terms of precision when we limit the sample of interest around the start of the fiscal year. On the other hand, widening the temporal window that defines the group would reduce comparability of administrative and survey earnings outcomes. Even though we thus fail to get strong guidance from the March-May group, the picture that emerges is consistent with the experimental impact on survey respondents to be a reliable estimate of the effect that ERA would have had on the annual earnings of the full eligible group – one which in addition to the non-participants includes *all* the participants, i.e. the non-respondents among the participants as well.

---

[33] Both in the case of ND25+ and NDLP, the participants among the March-May group pass the following basic checks. Random assignment as expected keeps holding (at least in terms of balancing the observables) and none of the following estimates is significantly different from one another: the experimental contrast in terms of administrative earnings for the survey respondents among the March-May group, the experimental impact estimate in terms of administrative earnings for the full March-May group and the experimental contrast in terms of survey earnings for the respondents among the March-May group.

6. Summary and conclusions

In this paper we have moved beyond a specific limitation of an experiment by 'climbing on its shoulders': we have drawn on random assignment to validate and if necessary correct the non-experimental methods used to address an issue – non-participation – which had arisen because of randomisation *per se*. Specifically, provided individuals were not diverted or did not formally refuse based on residual unobserved idiosyncratic impact components, the experiment both allows one to test the standard matching assumption required to identify the average treatment effect of interest and forms the basis of a correction for non-experimental estimates that fail the test.

We have shown the power of this strategy for the case of ERA, where the experimental set-up consistently altered the conclusions arising from non-experimental methods in terms of employment impacts for the ND25+ intake group. Non-experimental methods appeared to suggest that the experimental estimates significantly underestimate the average impact that the full eligible population would have experienced. However, once the non-experimental estimates were corrected to reflect failure of the test, the story changed to one of representativeness of the experimental estimate for the impact on all eligibles.

For the NDLP intake group, irrespective of the correction, the absence of employment impacts for the experimental sample was found to extend to the full eligible sample.

For either intake group, the experimental results in terms of survey earnings were found to be unbiased estimates of the impact on all eligibles, even after addressing survey and item non-response: the at times sizeable gain for responding participants was found to be representative of the average impact for the full eligible group.

Experimental impacts in terms of year-1 employment and survey earnings were thus found to be reliable representations of what the impacts would have been for the full group of eligibles had they been offered the chance to participate in ERA. The overall conclusion that despite a non-participation rate of 26.6% the ERA experiment has not suffered from randomisation bias could however only be reached by a judicious combination of both non-experimental methods and the experimental set-up.

We also found that the claim often made in the literature that histories variables modelled in a flexible way can capture unobservables relevant to no-treatment outcomes was not borne out in our data. This finding raises serious caveats as to the general validity of impact estimates typically obtained using matching methods based on the statement that controlling for detailed histories from administrative data adequately deals with selection.

References

Bloom, H.S. (1984), "Accounting for no-shows in experimental evaluation designs", *Evaluation Review*, 8, 225–246.

Burtless, G. (1995), "The case for randomised field trials in economic and policy research", *Journal of Economic Perspectives*, 9, 63-84.

Card, D. and Sullivan, D. (1988), "Measuring the effect of subsidized training programs on movements in and out of employment" *Econometrica*, 56, 497-530.

Cook, T.D and Campbell, D.T. (1979), *Quasi experimentation: Design and analysis issues for field settings*, Chicago, Rand McNally.

Dehejia, R., Wahba, S. (1999), "Causal effects in non-experimental studies: re-evaluating the evaluation of training programs", *Journal of the American Statistical Association*, 94, 1053–1062.

Dolton, P. and Smith, J. (2011), "The impact of the UK New Deal for Lone Parents on benefit receipt", IZA Discussion Paper No.5491.

Dorsett, R., Campbell-Barr, V., Hamilton, G., Hoggart, L., Marsh, A., Miller, C., Phillips, J., Ray, K., Riccio, J., Rich, S. and Vegeris, S. (2007), "Implementation and first-year impacts of the UK Employment Retention and Advancement (ERA) demonstration", Department for Work and Pensions Research Report No. 412.

Dubin, J.A., and D. Rivers (1993), "Experimental estimates of the impact of wage subsidies", *Journal of Econometrics*, 56, 219–242.

Frölich, M. (2004), "Program evaluation with multiple treatments", *Journal of Economic Surveys*, 18, 181-224.

Frölich, M. (2001), "Treatment choice based on semiparametric evaluation methods", Discussion Paper 2001-16, Department of Economics, University of St. Gallen.

Goodman, A. and Sianesi, B. (2007), "Non-participation in the Employment Retention and Advancement Study: A quantitative descriptive analysis", Department for Work and Pensions Technical Working Paper No.39.

Hall, N., Hoggart, L., Marsh, A., Phillips, J., Ray, K. and Vegeris, S. (2005), "The Employment Retention and Advancement Scheme: Report on the implementation of ERA during the first months. Summary and conclusions", Department for Work and Pensions Research Report No 265.

Heckman, J. (1992), "Randomization and social policy evaluation", in: C. Manski and I. Garfinkel, eds., *Evaluating welfare and training programs*, Harvard University Press, 201-230.

Heckman, J., and Smith, J. (1995), "Assessing the case for social experiments," *Journal of Economic Perspectives*, 9, 85-110.

Heckman, J. and Smith, J. (1999), "The pre-program dip and the determinants of participation in a social program: Implications for simple program evaluation strategies." *Economic Journal*, 109, 313-348.

Heckman, J., LaLonde, R. and Smith, J. (1999). "The economics and econometrics of active labor market programs", in Orley Ashenfelter and David Card (eds.), *Handbook of Labor Economics,* Volume 3A, 1865-2097.

Heckman, J.J., Hohmann, N. and Smith, J. (2000), "Substitution and dropout bias in social experiments: A study of an influential social experiment", *Quarterly Journal of Economics,* 2, 651–690.

Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998), "Characterising selection bias using experimental data" *Econometrica*, 66, 1017-1098.

Hendra, R., Riccio, J.A Dorsett, R., Greenberg, D.H., Knight, G., Phillips, J., Robins, P.K., Vegeris, S., Walter, J., Hill, A., Ray, K. and Smith, J. (2011), "Breaking the low-pay, no-pay cycle: Final evidence from the UK Employment Retention and Advancement (ERA) demonstration"*,* Department for Work and Pensions Research Report No. 765.

Hotz, V.J., Imbens, G.W. and Mortimer, J.H. (2005), "Predicting the efficacy of future training programs using past experiences at other locations", *Journal of Econometrics*, 125, 241-270.

Huber, M. (2012), "Identification of average treatment effects in social experiments under alternative forms of attrition", *Journal of Educational and Behavioral Statistics*, 37, 443–474.

Imbens, G.W. (2004), "Semiparametric estimation of average treatment effects under exogeneity: A review", *Review of Economics and Statistics*, 86, 4-29.

Kamionka, T. and Lacroix, G. (2008), "Assessing the external validity of an experimental wage subsidy", *Annales d'Economie et de Statistique*, 91-92, 357-384.

Lechner, M. and Smith, J.A. (2007), "What is the value added by caseworkers?", *Labour Economics*, 14, 135-151.

Lechner, M. and Wunsch, C. (2011), "Sensitivity of matching-based program evaluations to the availability of control variables", St. Gallen University Discussion Paper No. 2011-05.

Manski, C.F. (1996), "Learning about treatment effects from experiments with random assignment of treatments", *Journal of Human Resources*, 4, 709-733.

Moffitt, R. (1992), "Evaluation methods for program entry effects", in *Evaluating Welfare and Training Programs*, eds. C. Manski and I. Garfinkel, Harvard University Press.

Rosenbaum, P.R. and Rubin, D.B. (1985), "Constructing a comparison group using multivariate matched sampling methods that incorporate the propensity score", *The American Statistician*, 39, 33–8.

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. (1980), "Discussion of 'Randomisation analysis of experimental data in the Fisher randomisation test'", *Journal of the American Statistical Association,* 75, 591-593.

Sianesi, B. (2010), "Non-participation in the Employment Retention and Advancement Study: Implications for the experimental first-year impact estimates", Department for Work and Pensions Technical Working Paper No.77.

Walker, R., Hoggart, L. and Hamilton, G., with Blank, S. (2006), "Making random assignment happen: Evidence from the UK Employment Retention and Advancement (ERA) Demonstration", Department for Work and Pensions Research Report No. 330.

Wooldridge, J.A. (2002), "Inverse probability weighted M-estimators for sample selection, attrition, and stratification," *Portuguese Economic Journal* 1, 117-139.

Appendices

A1. Marginal effects from probit models of being a non-participant *versus* a participant

	ND25+	NDLP
Scotland	-0.163***	-0.253***
NE England	0.104***	-0.001
NW England	-0.093***	-0.264***
Wales	-0.051***	-0.096***
E Midlands	0.023	0.157***
2nd month of RA	-0.071***	-0.038
3rd month of RA	-0.056**	-0.040
4th month of RA	-0.075***	-0.053**
5th month of RA	-0.067***	-0.073***
6th month of RA	-0.084***	-0.054**
7th month of RA	-0.093***	-0.031
8th month of RA	-0.093***	-0.049*
9th month of RA	-0.087***	-0.090***
10th month of RA	-0.119***	-0.108***
11th month of RA	-0.086***	-0.086***
12th month of RA	-0.114***	-0.107***
13th month of RA	-0.134***	
Female	-0.009	-0.008
Age at inflow	-0.019***	0.009
Missing age	-0.215***	0.265*
Ethnic Minority	0.037**	-0.001
Missing ethnicity	0.012	0.023
Has disability/claims IB at inflow	0.007	-0.004
Has partner, ND25+	-0.010	
2 children, NDLP		-0.007
≥3 children, NDLP		-0.026
Youngest child <1 at inflow, NDLP		-0.009
Youngest child 1-5 at inflow, NDLP		0.021
Not on benefits at inflow, NDLP		0.118***
Early entrant, ND25+	-0.032	
Employed at inflow	0.042*	0.132***
Show up same day	-0.000	0.120
Show up w/in 30 days	-0.029**	-0.059***
Past participation in basic skills	0.007	0.012
Past participation in ND25+: once	0.001	0.082**
Past participation in ND25+: twice	0.011	0.111**
Past participation in ND25+: ≥3	0.044	0.059
Past participation in voluntary programs	-0.039***	0.022
Spent <50% of past 3 yrs on active benefits	-0.008	0.035
Spent >50 & <100% of past 3 yrs on active benefits	-0.006	
Spent 0% of past 3 yrs on inactive benefits	-0.076	-0.053
Spent >0 & <50% of past 3 yrs on inactive benefits	-0.051	0.005
Spent >50 & <100% of past 3 yrs on inactive benefits	-0.084	-0.017
Spent >0 & <25% of past 3 yrs in employment	-0.015	0.011
Spent ≥25% and <50% of past 3 yrs in employment	-0.027*	-0.008
Spent ≥50% of past 3 yrs in employment	-0.075***	-0.048***
Total New Deal caseload at office (100)	-0.002*	-0.004***
Share of lone parents in New Deal caseload at office	0.024	-0.048*
Bottom quintile of local deprivation	0.046	-0.006
2nd quintile of local deprivation	0.050**	0.051
3rd quintile of local deprivation	0.031*	0.020
4th quintile of local deprivation	0.028**	-0.020
TTWA-level unemployment rate	0.681	-1.306
Postcode missing or incorrect	0.417***	-0.061
Observations	7794	7258
Pseudo R squared	0.069	0.121

Notes: * significant at 10%; ** at 5%; *** at 1%;

See Table 2 for list of regressors; parsimonious summary of labour market histories used in the above probits.

A2. Reweighting estimator

As to the first term of expression (3), $E[E(Y \mid R=1, S=1, X)]$

$$= \int E(Y \mid R=1, S=1, x) \frac{f(x)}{f(x \mid R=1, S=1)} f(x \mid R=1, S=1) dx = \int E(\omega_1(x)Y \mid R=1, S=1, x) f(x \mid R=1, S=1) dx$$

$$= E[E(\omega_1(x)Y \mid R=1, S=1, X) \mid R=1, S=1] = E[\omega_1(x) \cdot S \cdot R \cdot Y], \text{ with}$$

$$\omega_1(x) \equiv \frac{f(x)}{f(x \mid R=1, S=1)} = \frac{P(R=1, S=1)}{P(R=1, S=1 \mid x)} = \frac{P(Q=1)P(R=1, S=1 \mid Q=1)}{P(Q=1 \mid x)P(R=1, S=1 \mid Q=1, x)}$$

where $P(R=1, S=1 \mid Q=1)$ is the probability among participants of being randomly assigned to the program group *and* of responding to the earnings question, and $P(R=1, S=1 \mid Q=1, x)$ is the corresponding conditional probability.

$E(Y_1)$ can thus be estimated by reweighing by $\omega_1(x)$ the outcomes of the program group members who responded to the earnings question and averaging them over this subgroup.

Similarly, the second term of expression (3) can be rewritten as:

$$E[E(Y \mid R=0, S=1, X)] = E[E(\omega_0(x)Y \mid R=0, S=1, X) \mid R=0, S=1] = E[\omega_0(X) \cdot S \cdot (1-R) \cdot Y], \text{ with}$$

$$\omega_0(x) \equiv \frac{P(Q=1)P(R=0, S=1 \mid Q=1)}{P(Q=1 \mid x)P(R=0, S=1 \mid Q=1, x)}.$$

A3. Covariate balancing summary indicators before and after matching

	Prob>chi		Pseudo R2		Median bias	
	Before	After	Before	After	Before	After
Administrative outcomes						
ND25+	0.000	1.000	0.069	0.001	4.2	0.6
NDLP	0.000	1.000	0.121	0.001	3.8	0.8
Survey outcomes						
Eligibles vs responding program group						
ND25+	0.000	1.000	0.030	0.005	4.2	1.3
NDLP	0.000	1.000	0.036	0.006	2.9	1.1
Eligibles vs responding control group						
ND25+	0.000	1.000	0.033	0.006	3.9	1.4
NDLP	0.000	1.000	0.042	0.008	3.4	1.1

Notes: Kernel matching estimators.

Prob>chi: *p*-value of the likelihood-ratio test before (after) matching, testing the hypothesis that the regressors are jointly insignificant, i.e. well balanced in the two (matched) groups.

Pseudo R^2: from probit estimation of the conditional probability of being a non-participant (before and after matching), giving an indication of how well the observables explain non-participation.

Median bias: median absolute standardised bias before and after matching, median taken over all the regressors. Following Rosenbaum and Rubin (1985), for a given covariate, the standardised difference *before* matching is the difference of the sample means in the non-participant and participant subsamples as a percentage of the square root of the average of the sample variances in the two groups. The standardised difference *after* matching is the difference of the sample means in the matched non-participants (i.e. falling within the common support) and matched participant subsamples as a percentage of the square root of the average of the sample variances in the two original groups.