Institute for Fiscal Studies

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

IFS Working Paper W14/10

Barbara Sianesi

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

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

May 2014

Barbara Sianesi

Institute for Fiscal Studies

Abstract: One of the most powerful critiques of the use of randomised experiments in the social sciences is the possibility that individuals might react to the randomisation itself, thereby rendering the causal inference from the experiment irrelevant for policy purposes. In this paper we set out a theoretical framework for the systematic consideration of “randomisation bias”, and provide what is to our knowledge the first empirical evidence on this form of bias in an actual social experiment, the UK Employment Retention and Advancement (ERA) study. Specifically, we empirically test the extent to which random assignment has affected the process of participation in the ERA study. We further propose a non-experimental way of assessing the extent to which the treatment effects stemming from the experimental sample are representative of the impacts that would have been experienced by the population who would have been exposed to the program in routine mode. We consider both the case of administrative outcome measures available for the entire relevant sample and of survey-based outcome measures. For the case of survey outcomes we extend our estimators to also account for selective non-response based on observed characteristics. Both for the case of administrative and survey data we further extend our proposed estimators to deal with the non-linear case of binary outcomes.

JEL codes: C21, J18, J38

Keywords: Social experiments, randomisation bias, sample selection, treatment effects, matching methods, reweighting estimators

Acknowledgments: I would like to acknowledge the helpful suggestions from Richard Blundell, Costas Meghir and seminar participants at the IFS, the Work Pensions and Labour Economics 2012 conference, the 2012 Annual EALE Conference, the 2012 4th Joint IZA/IFAU Conference on Labor Market Policy Evaluation, as well as Electra Small for her hard data work, Ingun Borg, Mike Daly, Christine Daniels, Richard Dorsett, Lesley Hoggart, Phil Robins, Jim Riccio and Stacy Sharman for detailed comments on previous drafts. Two editors and three referees further provided many insightful suggestions. The usual disclaimer applies. Financial support from ESRC Research Grant ES/I02574X/1 is gratefully acknowledged.

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. They are in fact the most reliable partial-equilibrium method for evaluating whether a program works, on average, for its participants – provided certain conditions are met.[1] An overarching label for such identifying conditions is the "no randomisation bias" assumption (first identified by Heckman, 1992; see also Heckman *et al.*, 1999), which rules out that random assignment *per se* has affected potential outcomes[2], as well as the program participation process. One of the most powerful critiques of the use of randomised experiments in the social sciences is thus the possibility that individuals might react to the randomisation itself, thereby making it impossible to infer the behaviour in real life from that observed in the experimental conditions. To our knowledge there is however no empirical evidence on the existence and scope of randomization bias, and no approach has been put forward for dealing with it.

This paper is the very first to tackle the issue of randomization bias with real experimental data. Specifically, we are in the unusual position to empirically test the extent to which random assignment has affected the process of participation in the experiment. We further propose a non-experimental way of assessing the extent to which the treatment effects stemming from the experimental sample are representative of the impacts that would have been experienced by the population who would have been exposed to the program in routine mode.

We offer a theoretical framework for the systematic consideration of randomisation bias, and then develop an analysis framework based on the conditional independence assumption (CIA). We consider both the case of administrative outcome measures available for the entire 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 standard CIA required to identify the average treatment effect of interest is made up of two parts. One part remains testable under the experiment 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. For the case of survey outcomes we extend our estimators to also account for selective non-response based on observed characteristics. Both for the case of administrative and survey data we further extend our proposed estimators to deal with the non-linear case of binary outcomes.

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

[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.

over 16,000 individuals being randomly assigned, the ERA study represented the largest randomised trial of a social program in the UK. The trial was set up to test the effectiveness of offering time-limited support once in work, in the form of advisory services and 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 mandated to enter the New Deal 25 Plus (ND25+) program, and lone parents who volunteered for the New Deal 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 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 the average impact of offering ERA services and incentives for all those who would have been eligible to receive such an offer in the absence of randomisation, the experimental evaluation can provide unbiased impact estimates only for the experimental sample – those who reached the randomisation stage and agreed to be randomly assigned. The concern is that the ERA study participants may be a selective group, potentially rendering the corresponding experimental impacts irrelevant for policy purposes. Note that it was the experimental set-up *per se* which gave rise to diverted customers and formal refusers, as 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 who would have been exposed to the offer of ERA had it not been evaluated via random assignment. Randomisation bias would then be present if the average effect for the experimental group is different from the average effect which would have arisen had the program been run in routine mode. Non-participation in the ERA study can thus be seen as poten-

[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.

tially introducing randomisation bias in the experimental estimate for the impact of offering ERA eligibility on the eligible population.[4]

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 place *after* treatments have been assigned.[5] This type of non-participation is also separate from entry effects[6]; 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 empirically measure the extent to which randomisation has affected the participation process. This is because (1) the treatment is the bestowing of an eligibility (to advisory services and financial incentives); (2) the parameter of interest is the average impact of offering this eligibility (the intention to treat effect); and (3) in routine mode the offer of this eligibility would have covered a well-defined and observed population: all ND25+ and NDLP entrants in the six districts over the intake window.

The key objective of the paper is to recover the causal effect of making the ERA package available for the full eligible population, i.e. for all those who would have been exposed to the ERA offer in the absence of randomisation. We thus 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, and subsequently test for the presence of randomisation bias by assessing whether this impact for the eligible group differs from the experimental impact estimated on the subgroup of ERA study participants. Given that the first-year experimental impacts are at times substantial, it is important to assess whether such estimates suffer from randomisation bias.

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 of administrative outcomes focus on matching and reweighting techniques under the CIA assumption that we observe all the ERA outcome-relevant characteristics that drive selection into the ERA

[4] An alternative but, as we discuss in Section 3.2, possibly less pertinent way to consider this issue is as a threat to the external validity of the experimental estimates.

[5] 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).

[6] 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.

study.[7] 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 strong. We show however that this CIA has two parts. One part is testable under the experiment.[8] The other part is the requirement that individuals were not diverted or did not formally refuse based on residual unobserved idiosyncratic ERA impacts conditional on arbitrarily heterogeneous impacts according to our rich set of observed characteristics – a highly plausible assumption as we argue in Section 5.4. Under this weak condition we can thus formally test the validity of the standard CIA assumption and, should it be rejected, correct the non-experimental estimates from selection bias. Of course, the corrected estimates are equivalent to those derived directly under our identifying assumption of no selection into the ERA study based on unobserved impacts.

An interesting by-product of testing the standard CIA assumption is that we can 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 and 1999, and Frölich, 2004; to some extent Hotz *et al.*, 2005; and for a caveat, Dolton and Smith, 2011, Lechner and Wunsch, 2011, and Biewen *et al.*, forthcoming). We found that the claim that histories variables can capture labour-market relevant unobservables was not borne out in our data. 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 in terms of no-treatment outcomes.

Our main findings highlight the power of our suggested strategy for the case of ERA, where the additional information from the experimental set-up consistently altered the conclusions arising from standard non-experimental methods in terms of employment outcomes. Specifically, non-experimental methods relying on the standard CIA based on the available data appeared to suggest that the experimental estimates underestimate the average impact that the full ND25+ eligible population would have experienced absent randomisation, while being representative of the effect for all NDLP eligibles. For both intake groups, however, once the non-experimental estimates were corrected to reflect failure of the test, the experimental estimates were shown to have a tendency to actually overestimate the average effect that all eligibles would have experienced. This overturning of the conclusion could only be reached by a judicious combination of both non-experimental methods and the experimental set-up.

The experimental results in terms of survey earnings were by contrast found to be unbiased estimates of the impact on all eligibles, even after addressing survey and item non-response: the at

[7] 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.

[8] This test exploits the controls in a way akin to the third test in Heckman and Hotz (1989).

times sizeable gain for responding participants was found to be a reliable representations of what the impacts would have been for the full group of eligibles in the absence of randomisation.

In conclusion, we found evidence that the 26.6% non-participation rate in the ERA experiment has introduced some randomisation bias in the experimental estimates in terms of employment outcomes, but has not affected the experimental estimates in terms of survey earnings.

The remainder of the paper is organised as follows. Section 2 describes the ERA study, outlines how non-participation in the study has come about and summarises the available qualitative evidence. Section 3 sets out a theoretical framework for randomisation bias, first in general terms and then showing how it simplifies in the ERA set-up. Section 4 describes the sample definitions and data content and provides some basic descriptives. Our methodological approaches are outlined in Section 5. After setting out the analysis framework in Section 5.1, we discuss how to assess and deal with randomisation bias when administrative outcomes are observed for the entire relevant sample (Section 5.2) and when only survey outcomes are available (Section 5.3), in both situations considering the case of continuous as well as binary outcomes. Section 5.4 provides an in-depth discussion of the plausibility of our identifying assumption. The results of all the empirical analyses are presented in Section 6, while Section 7 concludes.

2. The ERA intervention and non-participation in the ERA study

2.1 The Employment Retention and Advancement (ERA) study

ERA was conceived as an ambitious ‘next step’ in UK welfare-to-work policy. The existing New Deal programs were uniquely focused on helping non-working people on benefits find jobs, with no guidance offered to them once they entered work. By contrast, ERA’s remit was to help break the ‘low-pay-no-pay cycle’ so common among low-wage workers by concentrating on *in-work* support to help stabilise (“Retention”) and improve (“Advancement”) their work situations. While still unemployed, ERA offered job placement assistance as done by the regular New Deal programs. However once in work (and for up to two years), ERA offered a combination of financial incentives and access to employment counselling services.[9]

The ERA study was set-up to test the effectiveness of offering this package of post-placement support in helping individuals retain and progress in work. It is important to stress that ERA was the offer of a package of in-work support, and that the ERA experiment was a randomisation of eligibility design, in which eligible individuals (New Deal entrants) were randomly selected to be offered

[9] Specifically ERA offered (a) access to employment-related assistance by an adviser; (b) eligibility to a retention bonus of £400 three times a year for staying in full-time work 75% of the time, and to training tuition assistance (up to £1,000) and a training bonus (also up to £1,000) for completing training whilst at least part-time employed, and (c) access to emergency payments from a Discretionary Fund to overcome short-term barriers to remain in work.

the new treatment. Specifically, the program group was offered (or became eligible to) the new ERA services and incentives, while the control group continued to receive their usual New Deal services. For the two New Deal groups, the ERA study thus aimed at assessing the average effect of *offering* eligibility to ERA services and incentives (against the benchmark treatment of the standard New Deal). This average treatment effect (*ATE*) for the eligibles of making such a package available is unconditional on the actual take-up of any element of this offer[10] and should thus be interpreted as an intention-to-treat effect. For many purposes, this is the policy-relevant parameter, as it is informative on how the *availability* of ERA services and incentives affects individual outcomes.

2.2 Non-participation in the ERA study

As discussed above, the ERA treatment was the bestowing of an eligibility (to advisory services, financial incentives and a discretionary fund), which under normal conditions – that is in the absence of the randomised trial and/or if it were an official policy – would cover all New Deal entrants. Thus in an ideal scenario, the ERA study would have performed randomisation among *all* individuals starting the New Deal in the six evaluation districts over the intake window.[11] Departures from this ideal situation have however arisen from two sources:

1. intake process: not all eligible individuals have been offered the possibility to be randomly assignment and hence become eligible to 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 was offered ERA services and incentives, or to the control group, who only received the New Deal program whilst unemployed.

[10] It of course always remained up to individuals to decide whether and to what extent to avail themselves of the ERA elements. For instance, around 15% of the program group reported that they had had no contact *at all* with employment office staff during the year following their randomisation into the ERA group. Furthermore, some program group members may simply not (or no longer) be aware of some of the ERA features, as testified by the 1-year follow-up survey according to which around one quarter of the program group who had not heard of the retention bonus and as many as half or more (49% for NDLP and 57% for ND25+) who were not aware of the training bonus.

[11] It may help to think of how ERA would have been evaluated in the absence of random assignment. The standard way to evaluate programs in the UK is the pilot vs. comparison area-based scheme. ERA would have been introduced in the six 'pilot' districts, the New Deal inflow in those areas would have been viewed as the treated group, for whom the evaluator would have chosen a (possibly matched) comparison group from the New Deal inflow in other (possibly matched) comparable districts (see e.g. the evaluation of the Educational Maintenance Allowance by Dearden *et al.*, 2009). In the standard design there would have been no self-selection issues, since the entire eligible population in the pilot areas would be covered by the ERA offer. For the two New Deal groups, random assignment was thus implicitly deployed to control for area effects, that is for systematic differences in the eligible population between areas, as well as for area-specific macroeconomic effects or trends.

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 considerable discretion in how they organised the ERA intake and random assignment processes.[12] 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 become eligible to 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 the study be mandatory.[13]

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. According to the ERA implementation study (Walker *et al.*, 2006, p.26-17), "this incentive structure was real and widely recognised", so that "when New Deal 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 ERA advisers and off their caseloads) and those with good prospects to refuse (because they would keep them on their caseloads and get credit for a place-

[12] 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.

[13] 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).

ment). When 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." Job entry targets had thus 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. 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.

2. How willing were individuals to be randomly assigned?

Individuals who were given the option to participate in random assignment could formally refuse[14] 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.[15]

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, es-

[14] Signing: "*I do not consent to taking part in this research scheme or to being randomly assigned.*"

[15] 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).

pecially 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 the ERA study 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 both how large and how selective the non-participating groups were. Results are summarised in Section 4.4, 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.

3. Randomisation bias: A theoretical framework

3.1 A general framework[16]

Consider the general set-up of a program, which can either run in routine mode (*RCT*=0) or is implemented along with a randomised trial (*RTC*=1). Interest is in evaluating the average treatment effect on the treated (*ATT*). Randomisation bias can naturally be defined as a situation where the fact that a program is being implemented alongside a randomised trial gives rise to an average treatment effect on the treated, *ATT*(1), which is different from the average effect on the treated which would have arisen had the program been run in routine mode, *ATT*(0).

For now, imagine that eligible individuals can decide whether to participate or not in the program, with *D*(*RCT*) being a binary indicator denoting participation. The notation highlights how individual participation decisions can potentially depend on whether the program is being evaluated by random assignment or is run in routine mode. A first channel for randomization bias to operate is thus that the proportion of participants differs under randomization and in routine mode, i.e. $P(D_i(1)=1) \neq P(D_i(0)=1)$, *and* that such differences between the populations participating under the two scenarios also translate into a difference in the corresponding *ATT*'s.

The second channel of randomisation bias is one whereby randomisation *per se* may affect program impacts. To formalise it, potential outcomes[17] are allowed to depend on *RCT*: $Y_{1i}(1)$ denotes the outcome of individual *i* if *i* were to participate in the program implemented along a randomised trial and $Y_{1i}(0)$ is the outcome of individual *i* if *i* were to participate in the program implemented in routine mode, and $Y_{0i}(RCT)$ denote the corresponding no-treatment outcomes, similarly allowed to

[16] I am indebted to a referee for suggestions on how to set up a theoretical framework to think about randomisation bias.

[17] For the potential outcome framework see Rubin (1974).

differ according to whether the program is under randomisation or not. For example, $Y_1(RCT)$ will depend on *RCT* if the presence of random assignment disrupts the normal operation of the program, or if knowing that it is an experiment changes the behaviour of the program group; while $Y_0(RCT)$ will depend on *RCT* in the presence of control group substitution (a situation in which 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), or if control group members are disappointed from being randomised out in a way that affects their no-treatment outcomes.

The full definition of randomisation bias (Heckman, 1992; Heckman *et al.*, 1999) is thus:

$ATT(1) \equiv E(Y_{1i}(1) - Y_{0i}(1) \mid D_i(1)=1) \neq E(Y_{1i}(0) - Y_{0i}(0) \mid D_i(0)=1) \equiv ATT(0)$

In the following, we ignore the possibility that random assignment *per se* has affected impacts, in other words we assume $Y_{ki}(1) = Y_{ki}(0) = Y_{ki}$ for k=0,1 and for all i. This seems a very reasonable assumption in the case of ERA, as it is not easy to envisage how the initial randomisation process could have disrupted the treatment, which is *eligibility* to financial incentives and advisory services.[18] As to no-treatment outcomes, control group substitution can safely be ruled out as the control group was simply not eligible to ERA-style in-work incentives and support elsewhere. Discouragement effects were similarly highly unlikely as the control group didn't effectively know what ERA was about (see Section 2.2). In any case any initial disappointment that might have arisen from being randomised out of a mystery treatment would have long faded by the time we look at labour market outcomes one year later.

Our definition of randomization bias thus simplifies to:

$$ATT(1) \equiv E(Y_{1i} - Y_{0i} \mid D_i(1)=1) \neq E(Y_{1i} - Y_{0i} \mid D_i(0)=1) \equiv ATT(0) \qquad (1)$$

Let us now look in more detail at participation decisions. According to the potential participation behaviour, the population of all eligibles can be partitioned into four groups (similarly to the LATE framework of Imbens and Angrist, 1994):

- The *compliers* are those who decide not to participate *because* of randomisation: for them, $D_i(1) < D_i(0)$. Compliers arise e.g. if some eligibles withhold consent to be randomly assigned or, in case participation in the program entails some cost (e.g. in the form of preparatory activities), if some decide not to participate because they do not value the probability of obtaining the treatment as much as the certainty of obtaining it (threat of service denial).
- The *defiers* are those who are induced to participate by randomization: for them, $D_i(1) > D_i(0)$. Defiers could e.g. arise if intake criteria need to be changed in order to accommodate

[18] The ERA intervention was indeed completely new for advisers, who after some initial struggles became more and more proficient in delivering it as time went by. But these are to be regarded as standard teething problems in implementing a radically novel type of treatment (helping customers once in work), problems which are quite distinct from the initial randomisation process and which would equally have arisen had ERA been introduced without randomisation.

random assignment. Specifically, if it is required to maintain the program group at routine levels, the need to create a control group would entail the need to expand the pool of accepted participants. The additional recruits compared to normal operation (*RCT*=0) would be defiers.

- The *always-takers* would participate in any case: $D_i(1) = D_i(0) = 1$.
- The *never-takers* would not participate under either scenario: $D_i(1) = D_i(0) = 0$.

Two problems need to be overcome in order to assess the presence of randomisation bias in (1):

1. While we observe the participants under randomisation (the *D*(1)=1 group made up of the always-takers and the defiers), we do not in general observe those who would have participated under normal operation (the *D*(0)=1 group, encompassing the always-takers and the compliers).
2. Even if we could identify the individuals in the *D*(0)=1 group, the experiment would not allow for the identification of the average effect that the program would have had for them.

In short, when all that is available is a randomised trial, both the group of treated under normal operation (*D*(0)=1) and the treatment effect for this group (*ATT*(0)) are in general unobserved as they entail some counterfactual components. However since both *P*(*D*(0)=1) and *ATT*(0) are the relevant parameters to make a decision about whether or not to implement the program in routine mode (Heckman and Smith, 1998), obtaining biased estimates from the randomised trial is a real issue.

3.2 Randomisation bias in the ERA study

In the ERA set-up, however, the first problem – i.e. the identification of the *D*(0)=1 group – greatly simplifies. This is because, as discussed in Section 2.1:

(1) ERA was not a voluntary program where individuals needed to come forward to apply in order to receive the treatment[19], but was the bestowing of an eligibility (eligibility to advisory services, to financial incentives and to a discretionary fund);

(2) the parameter of interest is the average impact of *offering* this eligibility; and

(3) in routine mode the offer of this eligibility would have covered a well-defined (and indeed observed) population: all New Deal entrants in the given districts over the given time window. The *D*(0)=1 group that would have been exposed to ERA under normal operation is thus observed.

Note thus that by construction in the ERA set-up there were no never-takers, as under routine conditions one cannot refuse to be eligible to something they are eligible to. There were also no defiers, i.e. there were no New Deal entrants who would be eligible to ERA only in the demonstration (as there was no possibility – nor indeed any need – to manipulate entry into the New Deal in order to create the pool to be randomly assigned). It follows that the observed group of ERA study partici-

[19] New Deal entrants would never be required to apply in order to become eligible for ERA support; they would just *be* eligible for (or subject to) the ERA incentives and services by virtue of starting the New Deal (indeed the expectation was that even whilst unemployed in the New Deal individuals' behaviour would be affected by ERA, e.g. waiting for a better match for improved retention and/or for a full-time job rather than taking the first low-pay low-stay job).

pants (the $D(1)=1$ group) is made up of the always-takers only, and that the compliers are identified as the complement to the full eligible population. That is, the compliers are those eligibles who were not included in the ERA study: the diverted customers and the formal refusers. The fact that ERA was a study and involved random assignment has thus created a pool of eligible individuals who were denied or artificially allowed to 'refuse'[20] participation in something which in routine circumstances one could not be denied or one could not 'refuse': becoming *eligible* for financial incentives and personal advice.[21] Finally, the (observed) $D(0)=1$ group of New Deal entrants in the evaluation district and over the intake window is thus made up of the ERA study participants (the always-takers) and of the groups of formal refusers plus diverted customers (the compliers).

In general, the causal effect of randomisation on program participation choices, $P(D(1)=1) - P(D(0)=1)$, is unidentified, as one cannot observe individual participation decisions both in the presence and in absence of randomisation (an instance of Holland's, 1986, "fundamental problem of causal inference"). However in the ERA case, as we have seen, an eligible individual's participation status in the absence of randomisation is known, as they would all be eligible. $P(D(0)=1)$ is thus equal to 1, and, due to the experiment, $P(D(1)=1)$ is identified by the probability that an eligible is an ERA study participant. The causal impact of randomisation on participation patterns can thus be directly estimated by the observed share of non-participants. The incidence of non-participation by intake group and district is shown in Section 4.4.

It is important to stress that even if one were to show that the ERA study participants are different (in terms of observed and/or unobserved characteristics) from the full eligible population, this would not *per se* necessarily entail the presence of randomisation bias; as per condition (1), one would also need to show that the corresponding treatment effects differ as well. In other words, non-participation would have introduced randomisation bias in the experimental estimate for the impact of offering ERA eligibility on the eligible population if the average effect for the study participants is different from the average effect that the full eligible population would have experienced. Of course, even the ERA set-up faces the second problem of Section 3.1, as the experiment does not allow for the identification of the average effect that the program would have had on the full eligible population. We turn to these identification issues in Section 5.

An alternative way to view non-participation is as a potential threat to the external validity of the experimental estimate (Cook and Campbell, 1979). In this case, the parameter of interest is not the impact of the ERA offer for all eligibles (in the six districts), but the impact of the ERA offer for

[20] 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.

[21] This is of course not to say that randomisation necessarily entails changes in participation patterns, and in principle an experiment for ERA could have been devised in such as way as to avoid non-participation (e.g. not asking for consent, randomising offices rather than individuals, or even changing the incentive structure). Changes in participation patterns can however happen in a given experimental set-up – and have happened in the ERA set-up.

the sample of ERA study participants, and the issue is the extent to which such conclusions from the experiment would generalise to the whole eligible population (in the six evaluation districts).[22]

Finally note that we are always only concerned with the *current* experimental evaluation, i.e. with 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.[23]

4. Data and sample

4.1 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 this data with information from the ERA evaluation dataset (specifically, the Basic Information Form, BIF) on the participation decision and the outcome of random assignment (program/control group) for those who agreed to be randomly assigned.

Lastly, local-area level data (Census, travel-to-work and super-output area data) was merged in.

In section 4.3 we summarise the extensive variables we have derived from all of these sources.

[22] 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.

[23] To assess the impact of offering ERA in routine mode to all eligibles in the UK, one would need to address complex issues such as (a) compositional differences in the population of eligibles (in terms of different New Deal inflows in the rest of the country compared to the ones in the six districts and/or different inflows at a time in the future under different macro and local conditions; as well as arising from entry effects into the New Deal, e.g. more lone parents volunteering for NDLP in order to become eligible for ERA or some jobseekers over 25 not leaving unemployment in order to reach the start of ND25+ and hence ERA eligibility); (b) different ERA treatment (advisers would have become more adapt at delivering ERA and an increased knowledge and awareness of ERA among the eligibles would underlie the intention-to-treat effect); and (c) general equilibrium effects arising from the scaling up of a small pilot program.

4.2 Sample

To define our sample of ERA eligibles, we need to define the criteria determining eligibility and identify the relevant individuals in the data.[24] 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
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.[25]

The ERA study participants are directly identified by having signed their consent on the BIF; they make up the evaluation sample which was randomly assigned between a program group who was offered ERA services and incentives and a control group who was not. The formal refusers are also directly identified from the decision variable on the BIF. Diverted customers are those individuals who did start either the NDLP or ND25+ program during the intake period, but who for some reason were not added to the BIF and thus to the ERA evaluation sample. In addition to caseworkers exercising discretion as to whom they told about ERA (see Section 2.2), there is also the (undocumented) possibility that some individuals informally refused *before* having their BIF filled out. The diverted customers can only be identified residually, as those eligibles (as defined above) who did not appear on the BIF file. Formal refusers and diverted customers together form the group of non-participants (in the ERA study).

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.

[24] See Goodman and Sianesi (2007) for a very detailed description.

[25] The random assignment window was actually district- and intake group-specific, since one district started conducting random assignment later and some districts stopped earlier for some groups. Specifically, 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).

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%

4.3 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 been employment after 12 months 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).

Survey data on labour earnings in the 12-month follow-up period is available for a sample of participants. This measure offers a clean definition of employment (including all part-time work and self-employment; note it does not include the ERA bonuses) 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 different 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 sensitivity analyses of survey-based estimates.

All our outcomes of interest – employment probabilities and durations, and earnings – are related to labour market performance. As listed in Table 2, we have put together an extensive collection of individual, office and local area characteristics that are most likely to affect individuals' labour 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, disability 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

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

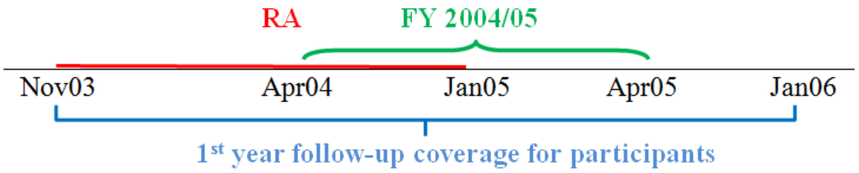


Table 2 Summary of observed characteristics

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 for 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

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 program) 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 experimented 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 years.[26] 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 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 (for more details see Appendix A5).

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.

[26] 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 least 5 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.

4.4 Descriptive analysis

As seen in Section 3.2, the incidence of non-participation can be viewed as the causal effect of randomisation on participation. 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.

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*

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 were 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.

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[27], 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 selective differences in outcome-relevant characteristics, 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 considerably worse employment outcomes among ND25+ entrants.

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

5. Methodological approaches

5.1 *Analysis framework*

The population of interest are those eligible to be offered ERA services and incentives, i.e. all New Deal entrants in the six evaluation districts over the intake window. We implicitly condition on this population throughout. The binary variable Q captures selection into the ERA study: Q=0 denotes those individuals who despite being eligible have not been randomly assigned (these are the compliers of Section 3.2, made up of the diverted customers and formal refusers), and Q=1 denotes the ERA study participants (the always-takers). The participants make up the experimental sample 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.

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”. ERA participants with missing survey outcome information (S=0), whatever the reason, are referred to as “non-respondents”.

Let $p \equiv P(Q=0)$ be the probability of non-participation among the eligibles (or the causal effect of randomisation on participation), which is directly identified in the data (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. Potential outcomes represented this way are (a) invoking SUTVA[28], (b) not affected by participation in the ERA study ($Y_{1Qi} = Y_{1i}$ and $Y_{0Qi} = Y_{0i}$ for Q=0, 1) and (c) not affected by randomisation *per se* ($Y_{ki}(RCT=1) = Y_{ki}(RCT=0) = Y_{ki}$ for k=0,1 – as justified in Section 3.1).

Our parameter of interest (see also Section 2.1) 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:

[28] 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.

$ATE = (1-p)\cdot ATE_1 + p\cdot ATE_0$. (2)

The parameter of interest *ATE* is thus given by a weighted average of the mean impact on participants and of the average impact that the non-participants would have experienced, with weights given by the relative share of participants and non-participants within the eligible pool.

Under some conditions (randomisation has not disrupted the program, there has been no control group substitution 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, as due to the randomness of *R* within the *Q*=1 group we have:

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

By contrast, ATE_0 and hence *ATE* are unobserved. Following on from the discussion in Section 3.2, we define randomisation bias to be present if the average effect for the study participants is different from the average effect that the full eligible population would have experienced had ERA been implemented in routine mode, that is if: $ATE_1 \neq ATE$.

We discuss how to assess and deal with randomisation bias in experimental studies when follow-up information on the outcomes of the non-participants is available (administrative outcomes – Section 5.2) and when it is not (survey outcomes – Section 5.3), in both situations considering continuous as well as discrete outcomes.

5.2 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 (*ATNT*), given that, as in the standard case, the no-treatment outcome of the non-treated (i.e. the non-participants) is observed. Equation (2) thus becomes:

$ATE = (1-p)\cdot ATE_1 + p\cdot\{E(Y_1 \mid Q=0) - E(Y \mid Q=0)\}$. (3)

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. The conditional independence assumption (CIA)[29] that allows us to directly identify this missing 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.

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

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 participates in the ERA study 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 $E(Y_1 | Q=0)$ is straightforward:

$$E(Y_1 | Q=0) = E[E(Y_1 | Q=0, X) | Q=0] = \text{(CIA-1)} = E[E(Y_1 | Q=1, X) | Q=0]$$

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

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 $p(X) \equiv P(Q=0 | X)$.

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 respect to the latter.

Specifically, consider the CIA in terms of non-ERA outcomes:

(CIA-0) $E(Y_0 | Q=0, X) = E(Y_0 | Q=1, X)$.

This condition is not needed for identification, given that – as is the case for the *ATNT* – we do observe the non-ERA outcomes of the non-participants. However – and in contrast to the standard *ATNT* case – randomisation allows us to directly test (CIA-0), i.e. that the non-ERA outcomes of the non-participants are the same, on average, as those of observationally equivalent participants. This test is implemented by testing whether $E(Y | Q=0, X) = E(Y | R=0, X)$, i.e. whether once controlling for observables, the non-participants and the control group (a representative sample of the participants for whom non-ERA outcomes are observed) experience the same average outcome.[30] 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

[30] 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.

groups there remain statistically significant differences conditional on the observables, this provides evidence of residual selection based on unobserved characteristics related to no-treatment outcomes.

Were we to reject the CIA in terms of Y_0, how would this relate to the plausibility of our identifying assumption of the CIA in terms of Y_1? We discuss this issue separately for continuous and for discrete outcomes.

5.2.1 Dealing with selection on unobserved characteristics: Continuous outcomes

To better understand how (CIA-1) and (CIA-0) relate to one another, 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) $E(\beta \mid Q{=}0, X) = E(\beta \mid Q{=}1, X)$.

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).

If one were to assume (CIA-β), (CIA-1) and (CIA-0) would thus imply each other.

This can also be seen in a more parametric model for observed outcomes which allows impacts to be heterogeneous across individuals in both observable $b(X_i)$ and unobservable b_i dimensions[31]:

$$Y_i = m(X_i) + \{b(X_i) \cdot R_i\} \cdot Q_i + \{b_i \cdot R_i\} \cdot Q_i + u_i. \quad (4)$$

In this model, (CIA-1) amounts to $(u_i, b_i) \perp Q_i \mid X_i$, and thus requires conditional independence to hold both in terms of Y_0-relevant unobserved characteristics u (no "omitted variable bias"; CIA-0) and in terms of unobserved idiosyncratic impacts b (no Roy model; CIA-β).[32]

Under (CIA-β), one can thus directly test the validity of the standard (CIA-1) assumption 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, note that under (CIA-β) the unobserved counterfactual conditional on X can be written as:

$$\begin{aligned} E(Y_1 \mid Q{=}0, X) &= \text{(CIA-}\beta\text{)} = E(Y_1 \mid Q{=}1, X) + \{E(Y_0 \mid Q{=}0, X) - E(Y_0 \mid Q{=}1, X)\} \\ &= E(Y \mid R{=}1, X) + \{E(Y \mid Q{=}0, X) - E(Y \mid R{=}0, X)\}. \end{aligned}$$

The ERA counterfactual for non-participants with characteristics X is thus identified by the average ERA outcome of program group members with the same X – the conditional matching estimate under (CIA-1) – plus a correction term that reflects a potential failure of (CIA-0) and captures by how much non-participants and participants with the same value of X still differ in terms of their average non-ERA outcome. It then follows that:

[31] The model reflects the fact that a participating (Q=1) individual receives the impact only with a 50-50 chance (i.e. if randomised into the program group).

[32] The distinction between selection on unobserved characteristics and selection on unobserved impacts is at the heart of Heckman *et al.* (1999), who highlight important implications of violations of either assumption for the properties of standard evaluation methods.

$E(Y_1 \mid Q{=}0) = E_X[E(Y \mid R{=}1, X) \mid Q{=}0] + \{E(Y \mid Q{=}0) - E_X[E(Y \mid R{=}0, X) \mid Q{=}0]\}$.

A violation of (CIA-0) thus introduces an overall bias term in the matching estimate of the average ERA outcome for the non-participants based on the average observed ERA outcome of observationally equivalent participants. In our set-up this bias term is identified in the data.

The expression for the average impact for the non-participants simplifies to:

$$ATE_0 = E_X[E(Y \mid R{=}1, X) \mid Q{=}0] - E_X[E(Y \mid R{=}0, X) \mid Q{=}0]. \qquad (5)$$

Estimation can be carried out by matching the non-participants twice, once to the program group and once to the control group, imposing common support across both terms.

Conceptually, one can arrive at the estimator in (5) in two ways. One, as done so far, is by directly focusing on the only unidentified term in equation (3), $E(Y_1|Q{=}0)$, invoke the corresponding (CIA-1) assumption and correcting, under (CIA-β), for any bias in terms of no-treatment outcomes. There is an interesting parallel between this way of proceeding and the standard difference-in-differences estimator, as in both cases an identifying assumption on the difference (idiosyncratic impact or trend conditional on X) is made to correct for an observed violation of such an assumption on the level (see Appendix A2 for more details). The other way is to ignore that only part of the ATE_0 needs to be identified, and directly identify ATE_0 under (CIA-β):

ATE_0 =(CIA-β)= $E_X[E(Y_1-Y_0|Q{=}1,X) \mid Q{=}0]$ =(RA)= $E_X[E(Y|R{=}1,X)|Q{=}0] - E_X[E(Y|R{=}0, X)|Q{=}0]$.[33]

Either way, the ultimately identifying assumption is (CIA-β); in Section 5.4 we argue its plausibility in the ERA experiment.

5.2.2 Dealing with selection on unobserved characteristics: Binary outcomes

The ways to deal with violations of (CIA-0) based on invoking (CIA-β) we have examined so far would not be appropriate for a binary outcome, as there would be (negative) dependence between base levels (Y_0) and differences (Y_1–Y_0). We could now interpret model (4) as a linear probability model, where the probability of, say, being employed is modelled as a linear function of a set of covariates including the offer of ERA support. This linear specification is however controversial, as linearity in the error term is particularly unrealistic for outcome variables with bounded support and predictions from the model cannot be guaranteed to fall within the 0-1 interval.

An alternative way to proceed proposed by Blundell *et al.* (2004) for their difference-in-differences analysis is to focus instead on latent dependent variables within the popular index models, and to assume linearity in the index. Following this idea, we can explore the potential of specifying our CIA-type assumptions on the *latent* variable Y^* which determines observed employment

[33] Indeed, one could further ignore that only (part of) the ATE_0 needs to be identified and identify under (CIA-β) the ATE directly: ATE =(CIA-β)= $E_X[E(Y_1-Y_0|Q{=}1,X)]$ =(RA)= $E_X[E(Y|R{=}1,X)] - E_X[E(Y|R{=}0,X)]$. Estimation of the ATE can then be carried out by matching the eligibles twice, once to the program group and once to the control group. This is not the most efficient way to proceed, however, as it implicitly involves extra matching steps to estimate the experimentally available ATE_1 (with $E(Y \mid R{=}k)$ being estimated by $E_X[E(Y \mid R{=}k, X) \mid Q{=}1]$ for k=0,1).

status Y according to: $Y_i = 1(Y_i{}^*>0)$. We can now think of applying model (4) to Y^*, modelling an individual's underlying, or latent, employability as a function of many factors (like skills, fixed costs for instance in terms of young children, etc), among which is the offer of ERA support ($R \cdot Q$):

$$Y_i^* = m(X_i) + \{b(X_i) \cdot R_i\} \cdot Q_i + \{b_i \cdot R_i\} \cdot Q_i + u_i. \qquad (4^*)$$

To extend to a non-linear setting our approach for dealing with randomisation bias, we thus follow what Blundell and Costa Dias (2009) and Lechner (2011) suggest for difference-in-differences models for a binary outcome.

Specifically, let the potential outcome equations be:

$Y_0 = 1(Y_0{}^*>0)$ and $Y_1 = 1(Y_1{}^*>0)$.

We are thus assuming that the conditional expectation of the binary potential outcome variables is related to the conditional expectation of the latent outcome variables in the following way:

$E(Y_0 \mid Q, X) = H[E(Y_0{}^* \mid Q, X)]$ and $E(Y_0{}^* \mid Q, X) = H^{-1}[E(Y_0 \mid Q, X)]$

$E(Y_1 \mid Q, X) = H[E(Y_1{}^* \mid Q, X)]$ and $E(Y_1{}^* \mid Q, X) = H^{-1}[E(Y_1 \mid Q, X)]$,

where the function $H(.)$ is assumed to be strictly monotonously increasing and invertible. H plays the role of a typical link function, which e.g. in a Probit model would be the cumulative standard normal distribution function Φ.

Just as these authors assume the difference-in-differences identifying assumption of common trends at the level of the expectations of the latent no-treatment outcome variable, we now invoke our CIA assumptions at the level of the expectations of the latent potential outcome variables.

We start with the latent-variable version of our identifying (CIA-1):

(CIA-1*) $\quad E(Y_1{}^* \mid Q=1, X) = E(Y_1{}^* \mid Q=0, X)$

which holds if and only if (CIA-1), i.e. $E(Y_1 \mid Q=1, X) = E(Y_1 \mid Q=0, X)$, holds.

Given that $E(Y_1 \mid Q=1, X) = E(Y \mid R=1, X)$, under (CIA-1*) we can then proceed fully non-parametrically the way done for continuous outcomes and recover the missing treatment counterfactual for the non-participants by matching them to observationally similar program group members.

We could alternatively use a Probit model. In this case, $H = \Phi$ and $E(Y \mid X=x, D=d) = \Phi(x'\theta_d)$. Under this parametric version of (CIA-1*) we would then estimate the counterfactual as:

$$E(Y_1 \mid Q=0) = \sum_{i\epsilon\{Q=0\}} \frac{\Phi(x_i{}'\theta_{R=1})}{N_0}.$$

Let us consider the equivalent of (CIA-0):

(CIA-0*) $\quad E(Y_0{}^* \mid Q=1, X) = E(Y_0{}^* \mid Q=0\ X)$

i.e. $\quad H^{-1}[E(Y_0 \mid Q=1, X)] = H^{-1}[E(Y_0 \mid Q=0, X)]$,

which holds if and only if (CIA-0), i.e. $E(Y_0 \mid Q=1, X) = E(Y_0 \mid Q=0, X)$, holds.

The test for (CIA-0*) is thus equivalent to the one for (CIA-0). As for continuous outcomes, testing whether $E(Y \mid R=0,X) = E(Y \mid Q=0,X)$ can thus be performed non-parametrically via matching.

Alternatively, we can assume a Probit model and test whether $E(Y \mid X{=}x, R{=}0) = E(Y \mid X{=}x, Q{=}0)$ by testing whether $\Phi(x'\theta_{R=0}) = \Phi(x'\theta_{Q=0})$.

So far we have shown that, as expected, moving to latent variables does not change the estimator under (CIA-1) or the testing strategy for (CIA-0) that were discussed before Section 5.2.1. Both of these are non-parametric in nature and thus apply to either continuous or binary outcomes, the only difference is that in the latter case one might wish to implement a parametric Probit version.

Things however diverge once we consider the implications for (CIA-1*) should (CIA-0*) fail. In particular, we will no longer be able to implement the estimator for discrete outcomes with correction (or, equivalently, directly) under (CIA-β^*) using non-parametric methods.[34]

We start again by first exploring the relationship between (CIA-1*) and (CIA-0*), noting how (CIA-1*) is equivalent to (CIA-0*) and (CIA- β^*), the latter ruling out selection into the ERA study based on the realised idiosyncratic ERA gains in terms of individual *latent* employability:

(CIA-β^*) $\quad E(Y_1^* - Y_0^* \mid Q{=}1, X) = E(Y_1^* - Y_0^* \mid Q{=}0\ X)$.

In terms of the model for latent Y^* in (4*), (CIA-0*) amounts to $E(u \mid Q, X) = E(u \mid X)$, (CIA-$\beta^*$) to $E(b \mid Q, X) = E(b \mid X)$ and (CIA-1*) to $E(u \mid Q, X) = E(u \mid X)$ and $E(b \mid Q, X) = E(b \mid X)$.

So if we assume (CIA-β^*), failure of (CIA-0*) implies failure of (CIA-1*).

Rewriting (CIA-β^*) to show the missing counterfactual, we have:

$E(Y_1^* \mid Q{=}0, X) \quad = E(Y_1^* \mid Q{=}1, X) + \{E(Y_0^* \mid Q{=}0, X) - E(Y_0^* \mid Q{=}1, X)\}$, which is equivalent to

$H^{-1}[E(Y_1 \mid Q{=}0, X)] \quad = H^{-1}[E(Y_1 \mid Q{=}1, X)] + H^{-1}[E(Y_0 \mid Q{=}0, X)] - H^{-1}[E(Y_0 \mid Q{=}1, X)]$

$\qquad = H^{-1}[E(Y \mid R{=}1, X)] \quad + H^{-1}[E(Y \mid Q{=}0, X)] \quad - H^{-1}[E(Y \mid R{=}0, X)]$.

Hence:

$E(Y_1 \mid Q{=}0, X) = H\{\ H^{-1}[E(Y \mid R{=}1, X)] +\ H^{-1}[E(Y \mid Q{=}0, X)] - H^{-1}[E(Y \mid R{=}0, X)]\ \}$.

Assuming a Probit model we obtain:

$E(Y_1 \mid Q{=}0, X) = \Phi\{\Phi^{-1}(\Phi(x'\theta_{R=1})) + \Phi^{-1}(\Phi(x'\theta_{Q=0})) - \Phi^{-1}(\Phi(x'\theta_{R=0}))\} = \Phi(x'\theta_{R=1} + x'\theta_{Q=0} - x'\theta_{R=0})$.

So that, under (CIA-β^*):

$$E(Y_1 \mid Q{=}0) = \sum_{i\epsilon\{Q=0\}} \frac{\Phi(x_i'\theta_{R=1} + x_i'\theta_{Q=0} - x_i'\theta_{R=0})}{N_0}\ .$$

It is straightforward to show that (a) assuming (CIA-β^*) and imposing the correction to the ATE_0 identified under (CIA-1*) leads to exactly the same result as imposing (CIA-β^*) straight away, and (b) if (CIA-0*) holds, results under parametric-(CIA-1*) and under (CIA-β^*) coincide.

Note that all three estimators for binary outcomes only differ in how they estimate the missing counterfactual $E(Y_1 \mid Q{=}0)$; the corresponding ATE_0 is then estimated by subtracting the observed average outcome of the non-participants. In the non-linear case, the observed no-treatment outcome

[34] This is another parallel with difference-in-differences models, which are functional form dependent (see e.g. the discussion in Lechner, 2011).

of the non-participants no longer cancels out as was the case for the bias-corrected estimator for the ATE_0 with a continuous outcome in (5). Table 4 summarises all our estimators for ATE_0.

Table 4: Estimators for ATE_0 for binary and continuous administrative outcomes

Binary outcomes	Continuous outcomes
Non-parametric (CIA-1$^{(*)}$) $ATE_0 = E[E(Y \mid R=1, X) \mid Q=0] - E(Y \mid Q=0)$	
Parametric (CIA-1*) $ATE_0 = \sum_{i\epsilon\{Q=0\}} \frac{\Phi(x_i{}'\theta_{R=1})}{N_0} - E(Y \mid Q=0)$	
Parametric (CIA-β^*) $ATE_0 = \sum_{i\epsilon\{Q=0\}} \frac{\Phi(x_i{}'\theta_{R=1} + x_i{}'\theta_{Q=0} - x_i{}'\theta_{R=0})}{N_0} - E(Y \mid Q=0)$	Non-parametric (CIA-β) $ATE_0 = E[E(Y \mid R=1, X) \mid Q=0] - E[E(Y \mid R=0, X) \mid Q=0]$

5.3 No follow-up data 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 4.3).

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 reweighting 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.[35]

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:

[35] 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.

(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[36]:

$$
\begin{aligned}
ATE &\equiv E(Y_1 - Y_0) = E[E(Y_1 - Y_0 \mid X)] = \text{(CIA-}\beta\text{)} = E[E(Y_1 - Y_0 \mid Q{=}1, X)] \\
&= \text{(RA)} = E[E(Y_1 \mid R{=}1, X)] - E[E(Y_0 \mid R{=}0, X)] \\
&= \text{(NR)} = E[E(Y_1 \mid R{=}1, S{=}1, X)] - E[E(Y_0 \mid R{=}0, S{=}1, X)] \\
&= E[E(Y \mid R{=}1, S{=}1, X)] - E[E(Y \mid R{=}0, S{=}1, X)] \qquad (6)
\end{aligned}
$$

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 A3 for the derivation):

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

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

Alternatively, the weights can be constructed via matching[37], 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.

5.3.1 Binary outcomes

Whilst we do not consider survey-based binary outcomes in our empirical analysis, this section highlights the main issues involved in the identification of ATE in such a situation.

We keep the assumptions (NR) on non-response. In this non-linear case, however, the (CIA-β) assumption made on the latent outcome variables, (CIA-β^*), is not enough for identification, as the latent model does not extend to the difference in potential outcomes. Instead, we need to invoke a stronger set of conditions which imply (CIA-β^*):

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

[36] 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.

[37] 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 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 $P(R{=}k \;\&\; S{=}1 \mid Q{=}0 \vee Q{=}1, X)$.

(CIA-1*) $E(Y_1^* \mid Q, X) = E(Y_1^* \mid X)$

Consider the average conditional no-treatment latent outcome:

$E(Y_0^* \mid X)$ =(CIA-0*)= $E(Y_0^* \mid Q{=}1, X) = H^{-1}[E(Y_0 \mid Q{=}1, X)]$ =(RA)= $H^{-1}[E(Y_0 \mid R{=}0, X)]$
=(NR)= $H^{-1}[E(Y_0 \mid R{=}0, S{=}1, X)] = H^{-1}[E(Y \mid R{=}0, S{=}1, X)]$

Hence, $H^{-1}[E(Y_0 \mid X)] = H^{-1}[E(Y \mid R{=}0, S{=}1, X)]$, or $E(Y_0 \mid X) = E(Y \mid R{=}0, S{=}1, X)$.

Assuming $H = \Phi$ for a Probit model, we obtain that $E(Y_0 \mid X) = \Phi(x'\theta_{R=0,S=1})$.

Similarly, under (CIA-1*), (RA) and (NR) we obtain that $E(Y_1 \mid X) = \Phi(x'\theta_{R=1,S=1})$.

The conditional *ATE* is thus identified as $ATE(X) = \Phi(x'\theta_{R=1,S=1}) - \Phi(x'\theta_{R=0,S=1})$, and the *ATE* as:

$$ATE = \frac{1}{N} \sum_{i\epsilon\{Q=1\}\cup\{Q=0\}} \{\Phi(x_i'\theta_{R=1,S=1}) - \Phi(x_i'\theta_{R=0,S=0})\}$$

The corresponding estimator for continuous outcomes in (6) only needed to invoke (CIA-β) and could rely on fully non-parametric methods (e.g. first and second terms estimated via matching).

5.3.2 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.[38] Furthermore, under the weak assumption (CIA-β), we could take the *ATE*

[38] Before considering non-participation, one can focus on the experimental March-May sample and (after having checked that, as one would expect, randomisation still holds, i.e. at least the observables are balanced between the pro-

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 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 subgroup 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[39] (though some evidence can be gleaned from test (B) in footnote 38). 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.

5.4 Plausibility of the identifying assumption

As seen for both continuous and binary outcomes, irrespective of whether (CIA-$0^{(*)}$) holds, the impact of ERA for all eligibles can be identified under (CIA-$\beta^{(*)}$). It is thus critical to discuss the plausibility of the assumption that participation in the ERA study was not based on realised unobserved idiosyncratic ERA impacts.

There is scant empirical evidence as to the extent to which unemployed job-seekers and their advisers are *ex ante* able to systematically predict the individual realised unobserved idiosyncratic impacts from a labour market program. All the available evidence, however, points to the insurmountable difficulties individuals face in estimating counterfactuals (indeed not just *ex ante*, but even *ex post*). Specifically, Smith *et al.* (2013) find that *ex post* participant evaluations are unrelated to the econometric impact estimates. As to caseworkers, Bell and Orr (2002) find that their *ex ante* evaluations of which participant will benefit most from the program have no predictive content, while Frölich (2001) and Lechner and Smith (2007) find that caseworkers' program allocation choices fail to maximize participants' subsequent employment prospects.

gram and control March-May subgroups) compare the experimental contrast in terms of administrative earnings for the survey respondents among the March-May group to (A) the experimental impact estimate in terms of administrative earnings for the full March-May group in order to assess non-response bias (for the March-May group) in terms of characteristics that affect administrative earnings; and to (B) the experimental contrast in terms of survey earnings for the respondents among the March-May group in order to garner evidence on whether administrative and survey earnings essentially measure the same impact despite not necessarily covering the same components.

[39] Indeed, administrative and survey earnings measures often differ substantially even when they nominally capture exactly the same components of earnings (see the discussion in the Bound *et al.* (2001).

As for the specific ERA set-up, drawing upon the meticulous and in-depth implementation analysis allows us to make the case for (CIA-$\beta^{(*)}$) even more compelling.

As to the formal refusers, we know from careful interviews with clients and staff (cf. Section 2.2) that, at the time of random assignment, New Deal entrants really had no clue of what ERA would entail, as the description of the program was purposefully left extremely vague in order to avoid disappointing control group members.[40] 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).[41] Given thus that formal refusers had no substantive knowledge of what ERA was or would entail, they had no possibility to try *ex ante* to predict their idiosyncratic gain from it.

As to the diverted customers, the qualitative evidence singles out the main source of diversion as being an incentive structure causing advisers to divert based on the very short-term non-ERA employment outcome (Y_0) they predicted for the job-seeker. Specifically, all advisers were under pressure to meet job-entry targets within a 'work-first' office ethos, being rewarded personally (and at the office-level) based on how many customers they placed into a job – indeed, any job. Advisers thus had only one major factor driving their decisions as to whether to divert customers and/or steer refusal: a job-seeker's likelihood to find any job as quickly as possible, i.e. non-ERA short-term employment probability. This diversion incentive applied to *any* adviser performing the intake, i.e. to ERA advisers as well. It has to be noted that while ERA had been designed as a primarily post-employment support package, a view towards advancement would ideally begin 'from day one', with advisers ideally encouraging customers to wait for a better match for improved job retention and/or for a full-time job, rather than taking the first low-pay low-stay job that came their way. However, as mentioned, ERA advisers were subject to the same job-entry targets; the competing operational priorities facing ERA[42] were so strong that six months into the program it was recognised (Hall *et al.*, 2005) that "the plan to begin advancement advice 'from day one' has not been realised" (p.23), "ERA advisers are finding it especially difficult to switch to a proactive way of working", "over the next three years they will need to make the transition from 'helping people get jobs' to 'making a sustained contribution to establishing their customers in decent well-paid job'",

[40] The customer fact sheet given out at intake simply read: "Customers randomly selected to receive Employment Retention and Advancement Services will be working closely with an Advancement Support Adviser to find, retain and advance in work. Advancement Support Advisers will provide help for up to 33 months as part of a package of support".

[41] Indeed, the discussion within the Project Team arose about 'informed consent' (or 'informed decline'), given that customers did not fully understand what they had consented to (or indeed, refused).

[42] "ERA is seen as cutting across the grain of [the employment offices'] work-first ethos" (p.25).

and that “this transition is crucial and the evidence of this report is that it will need increased input and resources to be delivered at a strength that will make the difference intended” (p.26).

Further support to the assumption that diversion was not driven by realised impacts is provided by the fact that ERA was a totally new program for advisers as well: never before had they been asked to help customers once in work. Indeed six months into the program the Project Team realised that there was essentially no ERA treatment being delivered on the ground. The implementation study by Hall *et al.* (2005) noted that “there is at this stage little evidence that the content of pre-employment services differs much between program and control group members” and that “no district achieved any real focus on advancement during the early period of ERA” (p.23), concluding the final section on “The future of ERA” with: “There is evidence that raises real concerns about whether [the employment offices] can deliver ERA in the way the design requires”.[43]

To summarise, violating the identifying (CIA-$\beta^{(*)}$) assumption would require asking individuals to systematically predict, *ex ante*, the realised idiosyncratic impacts of a mystery program (New Deal entrants) or of a completely new program (advisers) – and to do so over and above any completely general heterogeneity in impacts based on the full set of rich observed characteristics ($b(X)$ in models (4) and (4^*)). The assumption of no refusal or diversion based on realised unobserved idiosyncratic ERA impacts would thus seem particularly defensible in the ERA case

6. Empirical evidence on randomisation bias

This section presents all our empirical results, first those relating to employment outcomes measured by administrative data (Section 6.1), then those relating to yearly earnings measured, for the most part, by survey information (Section 6.2).

An overarching comment which applies to all the matching results is that our chosen estimator (kernel matching with an Epanechnikov kernel and a bandwidth of 0.06 as in Heckman *et al.*, 1997) has always performed extremely well in balancing the observables, both when estimating impacts (see Appendix Table A4) 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.

6.1 Employment

6.1.1 ND25+ group

The first column of Table 5 presents the experimental estimates of 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

[43] In response, the Project Team intervened with, among others, training events and advancement workshops for advisers, so that over time the ERA treatment was effectively brought about.

estimate controlling 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 and also control for differences in observables that have occurred by chance. Indeed for days in employment, the experimental impact becomes significant and the point estimates increase following the regression adjustment (see also Appendix Table A5.1).

A small positive effect of ERA of an extra 4-5 days in employment and a 2.2pp higher employment probability at the end of the first year has been uncovered for the participants.

But what effect would the full eligible group have experienced, on average?

Table 5: Employment outcomes for ND25+: Experimental point estimates of the average impact for participants (ATE_1) and residual bias in terms of non-ERA outcomes

	DAYS EMPLOYED			EMPLOYED AT MONTH 12		
		(CIA-0) test			**(CIA-0) test**	
	ATE_1	**OLS**	**Matching**	ATE_1	**Probit**	**Matching**
Raw	4.0	-9.4***		0.022**	-0.038***	
Conditional on *X*	4.6*	-7.9***	-9.7***	0.022**	-0.029**	-0.035***

Notes: ‘Raw’ are outcome differences between non-participants and participants. ‘OLS’, marginal effect from ‘Probit’ and ‘Matching’ are adjusted differences, controlling i.a. for parsimonious histories (see Appendix Table A5 for results based on different ways of constructing labour market histories).

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 5 reports the OLS/Probit and matching results from comparing the outcomes of the two groups conditional on observables. The overall conclusions are twofold. First, there remain large and statistically significant imbalances in employment outcomes, with non-participants being on average 8-10 fewer days in employment and around 3pp less likely to be employed than observationally equivalent participants. Second, how past labour market history is measured makes no difference at all (see Appendix A5 for results based on different ways of summarising labour market histories and for a more in-depth discussion). In contrast to Dolton and Smith (2011), but in line with Biewen *et al.* (forthcoming), more detailed, more 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. Both of these conclusions apply for the NDLP group as well (see Appendix Table A5.1). For the subgroup of eligibles flowing in after April 2004 (the subgroup for whom 2003/04 fiscal year earnings represent pure *pre*-treatment information), even the addition of pre-treatment earnings did not make any difference (Appendix Table A5.2). 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 for the no-treatment case, which is indeed the case of interest when using non-experimental comparison groups for estimating treatment effects.

Having shown that the ERA study participants are different from the non-participants (in terms of observed as well as unobserved characteristics) is not enough to infer that randomisation bias is present; what is also needed is that such uncovered compositional differences effectively translate into an experimental impact which is different from the average effect that the full eligible population would have experienced had ERA been implemented in routine mode. It is thus the comparison of the experimental estimates of the ATE_1 to the estimates of the ATE which is informative of the presence and direction of randomisation bias.

To this end, Table 6 presents the experimental impact estimate for the participants (ATE_1) and two sets of non-experimental estimates for the non-participants (ATE_0) and for all eligibles (ATE). The first set relies on the (CIA-$1^{(*)}$) assumption and ignores the mismatch documented in Table 5 in the non-ERA outcomes of the non-participants and observationally similar participants. As seen in Section 5.2, these estimates are biased under our identifying assumption (CIA-$\beta^{(*)}$). The second set of estimates has been adjusted to correct for such mismatch and corresponds to the estimates directly obtained under (CIA-$\beta^{(*)}$).

Table 6: Employment outcomes for ND25+: Average ERA impacts for participants (ATE_1), non-participants (ATE_0) and all eligibles (ATE)

		ATE_1	ATE_0	ATE	$ATE_1 \neq ATE$
DAYS EMPLOYED	Unadjusted	4.6*	10.1***	5.9***	*
	Adjusted		0.5	3.7	**
EMPLOYED M=12	Unadjusted	0.022**	0.045***	0.027***	*
	Adjusted		0.014	0.020*	**

Notes: Unadjusted estimates ignore failure of the (CIA-$0^{(*)}$) test. Share of non-participants is 0.23. Kernel matching with Epanechnikov kernel (bandwidth of 0.06); statistical significance based on bootstrapped bias-corrected confidence intervals (1000 replications); $ATE_1 \neq ATE$: bootstrap-based statistical significance of the difference; *** significant at 1%, ** at 5%, * at 10%.

The difference between the ATE_1 and the adjusted ATE is found to be statistically different from zero, with the experimental estimates having a tendency to *overestimate* the average effect that all eligibles would have enjoyed absent randomisation: participants experience an extra 4.6 days in employment thanks to ERA (significant at 10%), while no statistically significant impact could be detected for all eligibles; participants also enjoy a 2.2pp increase in employment probability (significant at 5%), compared to a 2.0pp gain for all eligibles (significant at 10%). A certain extent of randomisation bias thus seems to affect the ND25+ experimental estimates in terms of employment outcomes.

Table 6 further shows that had we simply relied on the standard assumption for the *ATNT*, (CIA-1$^{(*)}$), we would have reached quite a different conclusion. If we ignored that the (CIA-0$^{(*)}$) test failed 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 *underestimates* how much ERA would have improved the employment outcomes of all eligibles absent randomisation. Specifically, the employment gain for the non-participants (10 days and 4.5pp) would appear to be more than double that of participants (4.6 days and 2.2pp), resulting in highly statistically significant overall *ATE*'s of 6 days and 2.7pp, which are statistically different from the ATE_1's. Relying on the estimate of the *ATNT* based on invoking the standard (CIA-1$^{(*)}$) assumption and ignoring its rejection under (CIA-$\beta^{(*)}$) would thus have led to an erroneous conclusion as to the direction of randomisation bias.

6.1.2 NDLP group

Table 7 shows that, in contrast to the case of ND25+, the ERA offer has left the duration and incidence of employment of NDLP participants completely unaffected during the follow-up year.

Before turning to our estimates for all eligibles, we again consider the results of the (CIA-0) test (Table 7; see also Appendix Table A5.1). Perhaps surprisingly, for both employment outcomes there are no statistically significant raw differences in the average no-treatment outcomes of non-participants and participants. Large and statistically significant differences however emerge once controlling for observables, with non-participants being now 10-12 fewer days and 4pp less likely to be employed than observationally equivalent controls. We have however to control for relevant pre-treatment characteristics as there are sizeable 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 A1 for marginal effects).

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

	DAYS EMPLOYED			EMPLOYED AT MONTH 12		
		(CIA-0) test			**(CIA-0) test**	
	ATE_1	**OLS**	**Matching**	ATE_1	**Probit**	**Matching**
Raw	-0.1	3.8		-0.007	-0.003	
Conditional on *X*	-2.2	-10.4***	-11.2**	-0.014	-0.040***	-0.039**

Notes: 'Raw' are outcome differences between non-participants and participants. 'OLS', marginal effect from 'Probit' and 'Matching' are adjusted differences, controlling i.a. for parsimonious histories (see Appendix Table A5 for results based on different ways of constructing labour market histories).

Table 8 presents estimates of the three causal parameters of interest. For the ATE_0 and ATE, both the unadjusted non-experimental estimates and the estimates corrected for the bias in terms of mean no-treatment outcomes are shown.

If one were to ignore failure of the (CIA-$0^{(*)}$) condition and hence, under CIA-$\beta^{(*)}$), violation of the (CIA-$1^{(*)}$) assumption, one would conclude that the experimental estimate of no ERA impact on employment outcomes is representative of the average effect that the eligibles would have experienced absent randomisation. In particular, the employment effect in terms of either employment duration or probability would have been the same – and statistically indistinguishable from zero – for the experimental group, the non-participants and all eligibles.

This conclusion of an absence of randomisation bias is however again qualified under our preferred estimates which adjust for selection on non-ERA outcome-relevant unobserved characteristics. Specifically, the adjusted estimates of the ATE are statistically different from – and smaller than – the corresponding experimental estimates. While for employment durations neither the experimental ATE_1 nor the adjusted estimate for the ATE reach statistical significance, for employment probability the adjusted ATE of a 2.2pp fall is significant at the 10% level. Under our preferred specification we thus find some weak evidence of randomisation bias, whereby a zero impact for the study participants would not always be representative of the impact that the eligible group would have experienced in the absence of randomisation.

Table 8: Employment outcomes for NDLP: Average ERA impacts for participants (ATE_1), non-participants (ATE_0) and all eligibles (ATE)

		ATE_1	ATE_0	ATE	$ATE_1 \neq ATE$
DAYS EMPLOYED	Unadjusted	-2.2	-2.1	-2.2	no
	Adjusted		-13.4**	-5.6	**
EMPLOYED M=12	Unadjusted	-0.014	0.000	-0.010	no
	Adjusted		-0.039**	-0.022*	**

Notes: Unadjusted estimates ignore failure of the (CIA-$0^{(*)}$) test. Share of non-participants is 0.304. Kernel matching with Epanechnikov kernel (bandwidth of 0.06); statistical significance based on bootstrapped bias-corrected confidence intervals (1000 replications); $ATE_1 \neq ATE$: bootstrap-based statistical significance of the difference; *** significant at 1%, ** at 5%, * at 10%.

6.2 Earnings

For both intake groups, the experiment highlights a sizeable and statistically significant gain in 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 9). 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 or accepted to take the survey. 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 survey respondents did not differ dramatically from 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.

Table 9: 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%.

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 5.3. Table 9 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 point estimates quite close to each other, which are not statistically different from the adjusted experimental contrast on respondents. For either intake group, we thus find that the statistically significant and sizeable 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 10).

Table 10: Sensitivity analyses for earnings outcomes

(i) **ND25+**

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

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

Full March-May group

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

(b) annual earnings (survey)			$\Delta_{S=1,X} \neq ATE_{\text{surv}}$	$ATE_{\text{adm}} \neq ATE_{\text{surv}}$
$\Delta_{S=1,X}$		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	N
Post-April group	summary	210	1.087	-82	-69	3,002
March-May group	summary	323	1.132	-10	52	1,845

Full March-May group

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

(b) annual earnings (survey)			$\Delta_{S=1,X} \neq ATE_{\text{surv}}$	$ATE_{\text{adm}} \neq ATE_{\text{surv}}$
$\Delta_{S=1,X}$		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. 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.

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 statistically significant differences in non-ERA earnings remain between non-participants and matched participants.[44]

Table 10(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-

[44] 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.

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.

The March-May group lends itself to a more direct robustness check as this is the subgroup 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.[45] 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 statistically 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 10(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 statistically insignificant. As to our robustness checks, none of the non-experimental estimates based on survey earnings is statistically different from the *ATE* estimated from fiscal year administrative data, or indeed from the adjusted experimental contrast for survey respondents. The implementation of this sensitivity analysis clearly suffers from the small size of the group, which prevents one to discriminate with enough precision between estimates using different non-parametric methods. Limiting the sample of interest around the start of the fiscal year entails a heavy price in terms of precision. 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.

[45] 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 in the sense that the confidence interval for the difference of the estimates includes zero already at the most conservative level: 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.

7. Conclusions

In this paper we have set out a framework to think about randomisation bias, a form of bias which can potentially render the causal inference from randomised experiments irrelevant for policy purposes. We have also provided what is to our knowledge the first empirical evidence on the presence and extent of randomisation bias in a social experiment, the ERA study.

For both intake groups we have found evidence that non-participation in the ERA study has introduced some randomisation bias in the experimental impact estimates in terms of employment measures, but not in terms of survey earnings.

Provided individuals did not participate in the study based on residual unobserved idiosyncratic impact components, we have shown that we can draw on random assignment to assess the validity of estimates arising from non-experimental methods based on the standard CIA assumption. The power of this strategy was demonstrated for the case of ERA, where the additional information from the experiment consistently overturned the conclusions on employment impacts arising from standard non-experimental methods applied to the available data. Additional findings from the test exploiting the experimental set-up highlighted how the claim often made in the literature that histories variables modelled in a flexible way can capture no-treatment outcome-relevant unobservables is far from being of general validity. Care should thus be taken when considering impact estimates typically obtained using matching methods based on the statement that controlling for detailed histories from administrative data adequately deals with selection.

Finally, it is worth highlighting that we were in a position to assess randomisation bias in the ERA experiment because the treatment was the bestowing of eligibility, and in the absence of randomisation this new eligibility would have covered a well-defined and observed population. Future research efforts should be directed to consider the case of randomised trials of voluntary treatments, in which case both the group of treated under normal operation and the treatment effect for this group are unobserved.

References

Bell, S.H. and Orr, L.L. (2002), "Screening (and creaming?) applicants to job training programs: the AFDC homemaker–home health aide demonstrations", *Labour Economics*, 9, 279–301.

Biewen, M, Fitzenberger, B., Osikominu, R., and Paul, M., "The Effectiveness of Public Sponsored Training Revisited: The Importance of Data and Methodological Choices", forthcoming in *Journal of Labor Economics.*

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

Blundell, R., and M. Costa Dias (2009): "Alternative Approaches to Evaluation in Empirical Microeconomics", *Journal of Human Resources*, 44, 565-640.

Blundell, R., C. Meghir, M. Costa Dias, and J. van Reenen (2004): "Evaluating the Employment Impact of a Mandatory Job Search Program", *Journal of the European Economic Association*, 2, 569-606.

Bound, J., Brown, C. and Mathiowetz, N. (2001), “Measurement error in survey data”, ch 59 in Heckman, J.J. and Leamer, E.E. (eds.) *Handbook of Econometrics*, Volume 5, 3705-3843.

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.

Dearden, L., Emmerson, C., Frayne, C. and Meghir, C. (2009), “Conditional Cash Transfers and School”, *Journal of Human Resources*, 44, 827-857.

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: The early months of implementation. Summary and conclusions”, Department for Work and Pensions Research Report No 265.

Heckman, J.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.J. and Hotz, V.J. (1989), “Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training”, *Journal of the American Statistical Association*, 84, 862-74.

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

Heckman, J.J., and Smith, J. (1998): “Evaluating the Welfare State,” in: S. Strom, ed, *Econometrics and Economics in the 20th Century*, Cambridge University Press, New York.

Heckman, J.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.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.J., Ichimura, H. and Todd, P.E. (1997), "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies*, 64, 605-654.

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

Holland, P. (1986), "Statistics and Causal Inference", *Journal of the American Statistical Association*, 81, 945-970

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.

Imbens, G.W. and Angrist, J.D. (1994), "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62, 446-475.

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. (2011), "The estimation of causal effects by difference-in-difference methods", University of St. Gallen Discussion Paper no. 2010-28.

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.

Smith, J., Whalley, A. and Wilcox, N.T. (2013), "Are program participants good evaluators?", mimeo, April 29.

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: Parallels between the difference-in-differences method and the identification strategy invoking (CIA-β) to correct for a violation of (CIA-0)

Note: For the DiD, the time-subscript for outcomes in the follow-up period is omitted.

	Difference-in-Differences	(CIA-β) to correct for a violation of (CIA-0)
Parameter	Missing counterfactual $E(Y_0 \mid Q=1)$ for the *ATT*	Missing counterfactual $E(Y_1 \mid Q=0)$ for the *ATNT*
	DiD can deal with violations of (CIA-0) only for the *ATT*.[a] This paper is ambitiously trying to perform a similar task but for the *ATNT*.	
Assumptions	(1) Unaffected Sampled Period (USP): no treatment effect in the pre-treatment period (2) Common Trends: the average change in no-treatment outcomes would have been the same for treated and non-treated with the same X: $E(Y_0 - Y_{0,\text{pre}} \mid Q=1, X) = E(Y_0 - Y_{0,\text{pre}} \mid Q=0, X)$ Note this is the same as Bias Stability: the average bias from selection on unobserved characteristics affecting no-treatment outcomes (i.e. the violation of CIA-0) is the same in the pre- and post-treatment periods: $Bias_{\text{pre}}(X) \equiv E(Y_{0,\text{pre}} \mid Q=0, X) - E(Y_{0,\text{pre}} \mid Q=1, X)$ $= E(Y_0 \mid Q=0, X) - E(Y_0 \mid Q=1, X) \equiv Bias(X)$	(1) (Follow-up) potential outcomes are not affect by either - participation in the study ($Y_{1Qi} = Y_{1i}$ and $Y_{0Qi} = Y_{0i}$ for Q=0, 1) nor - randomisation *per se* ($Y_{ki}(RCT=1) = Y_{ki}(RCT=0) = Y_{ki}$ for k=0,1) (2) (CIA-β): the average idiosyncratic impact would have been the same for treated and non-treated with the same X: $E(Y_1 - Y_0 \mid Q=1, X) = E(Y_1 - Y_0 \mid Q=0, X)$
Identification	$E(\mathbf{Y_0} \mid Q=1, X) = E(Y_0 \mid Q=0, X) + \{E(Y_{0,\text{pre}} \mid Q=0, X) - E(Y_{0,\text{pre}} \mid Q=1, X)\}$ $= E(Y \mid Q=0, X) + \{\mathbf{E(Y_{pre} \mid Q=0, X) - E(Y_{pre} \mid Q=1, X)}\}$ conditional matching estimate under (**CIA-0**) **bias in terms of (CIA-0) in the pre-treatment period**	$E(\mathbf{Y_1} \mid Q=0, X) = E(Y_1 \mid Q=1, X) + \{E(Y_0 \mid Q=0, X) - E(Y_0 \mid Q=1, X)\}$ $= E(Y \mid R=1, X) + \{\mathbf{E(Y \mid Q=0, X) - E(Y \mid R=0, X)}\}$ conditional matching estimate under (**CIA-1**) **bias in terms of (CIA-0) in the same follow-up period**
Discussion	Under USP, any non-zero difference in average pre-treatment outcomes of the two groups is an estimate of the bias of the CIA-0 assumption in the pre-treatment period. Assuming that this bias is constant over time, it can be used to correct the estimate of the average no-treatment counterfactual estimated for the treated in the post-treatment period based on the average no-treatment outcome of the matched non-treated.	Assuming that participation in the study and randomisation do not affect potential outcomes allows one to compare non-participants vs controls in the follow-up period (as USP allows DID to compare treated vs non-treated in the pre-treatment period). Assuming that both groups would have experienced the same average idiosyncratic impact, this difference can be used to correct the estimate of the average treatment counterfactual for the non-participants based on the average treatment outcome of the matched program group.
Summary	Assumes that despite the (testable) presence of arbitrary selection into the treatment on levels: $E(Y_{0,\text{pre}} \mid Q=1, X) \neq E(Y_{0,\text{pre}} \mid Q=0, X)$, the trend would have been the same for both groups given X, on average: $E(Y_0 - Y_{0,\text{pre}} \mid Q=1, X) = E(Y_{0t} - Y_{0,\text{pre}} \mid Q=0, X)$ i.e. allows (and tests) for selection on the level but rules out selection on the trend (or difference)	Assumes that despite the (testable) presence of arbitrary selection into the treatment on levels: $E(Y_0 \mid Q=1, X) \neq E(Y_0 \mid Q=0, X)$, the idiosyncratic gain would have been the same for both groups on average: $E(Y_1 - Y_0 \mid Q=1, X) = E(Y_1 - Y_0 \mid Q=0, X)$ i.e. allows (and tests) for selection on the level but rules out selection on gains (or difference)

[a] For the *ATNT*, the DiD would need a group treated in period 0 for whom the treatment effect disappears in period 1, an implausible scenario (see Lechner, 2011).

A3. Reweighting estimator

As to the first term of expression (6), $E[E(Y \mid R{=}1, S{=}1, X)]$

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

$= E[E(\omega_1(x)Y \mid R{=}1, S{=}1, X) \mid R{=}1, S{=}1] = E[\omega_1(x)\cdot S\cdot R\cdot Y]$, 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 $P(R{=}1,S{=}1|Q{=}1)$ is the probability among participants of being randomly assigned to the program group *and* of responding to the earnings question, and $P(R{=}1,S{=}1|Q{=}1,x)$ 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:

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

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

A4. Matching estimator

The matching estimator used for all the analyses in the paper is the kernel estimator introduced by Heckman *et al.* (1997), who similarly used an Epanechnikov kernel and a bandwidth of 0.06. As seen in the table, this estimator could balance the observables extremely well (that was not the case for e.g. nearest neighbour or Mahalanobis-metric matching). While common support was imposed at the boundaries (i.e. discarding target individuals whose propensity score was larger than the largest propensity score in the group they were being compared to), it never led to the loss of more than 1-2% of the group of interest. Inference has always been based on bootstrapped, bias-corrected confidence intervals (based on 1,000 replications).

Summary indicators of covariate balance 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:

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.

A5. Testing the (CIA-0) condition: Supplementary material and discussion

Table A5.1: Employment outcomes: 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

	Days employed			Employed at month 12		
	ATE_1	(CIA-0) test		ATE_1	(CIA-0) test	
		OLS	Matching		Probit	Matching
ND25+						
Raw	4.0	-9.4***		0.022**	-0.038***	
All other *X*'s plus						
summary	4.6*	-7.9***	-9.7***	0.022**	-0.029**	-0.035***
monthly employment	4.8**	-7.6***	-9.4***	0.023**	-0.028**	-0.031**
ever employment	5.0**	-7.6***	-9.4***	0.022**	-0.028**	-0.034***
sequence	4.8**	-7.9***	-8.8***	0.022**	-0.029**	-0.033**
summary + monthly empl.	4.8**	-7.7***	-9.2***	0.023**	-0.028**	-0.031**
summary + ever employed	5.0**	-7.7***	-9.3***	0.023**	-0.029**	-0.034***
summary + sequence	4.8**	-8.0***	-8.8***	0.022**	-0.029**	-0.033***
NDLP						
Raw	-0.1	3.8		-0.007	-0.003	
All other *X*'s plus						
summary	-2.2	-10.4***	-11.2**	-0.014	-0.040***	-0.039**
monthly employment	-2.4	-10.2***	-10.2**	-0.016	-0.041***	-0.038**
ever employment	-2.5	-11.0***	-12.1**	-0.016	-0.042***	-0.043**
sequence	-2.4	-10.8***	-11.7**	-0.016	-0.041***	-0.040**
summary + monthly empl.	-2.7	-10.6***	-11.1**	-0.017	-0.042***	-0.039**
summary + ever employed	-2.2	-10.8***	-12.4**	-0.015	-0.041***	-0.043**
summary + sequence	-2.1	-10.4***	-11.2**	-0.015	-0.040***	-0.037**

Notes: 'Raw' are outcome differences between non-participants and participants. 'OLS' (for days in employment), marginal effect from 'Probit' (for employed at month 12) and 'Matching' are adjusted differences. See Section 4.3 for the description of how labour market histories have been constructed.

Table A5.2: Days in employment for the sample of inflow from April 2004: Experimental point estimates of the average impact for participants (ATE_1) and residual bias in terms of non-ERA outcomes controlling and not controlling for pre-treatment earnings

	ND25+			NDLP		
	ATE_1	(CIA-0) test		ATE_1	(CIA-0) test	
		OLS	Matching		OLS	Matching
Raw	6.6**	-8.7***			3.2	
With pre-treatment earnings	8.0***	-5.4	-2.9	4.0	-15.0***	-13.6
Without	8.0***	-5.4	-2.9	4.0	-15.0***	-13.6

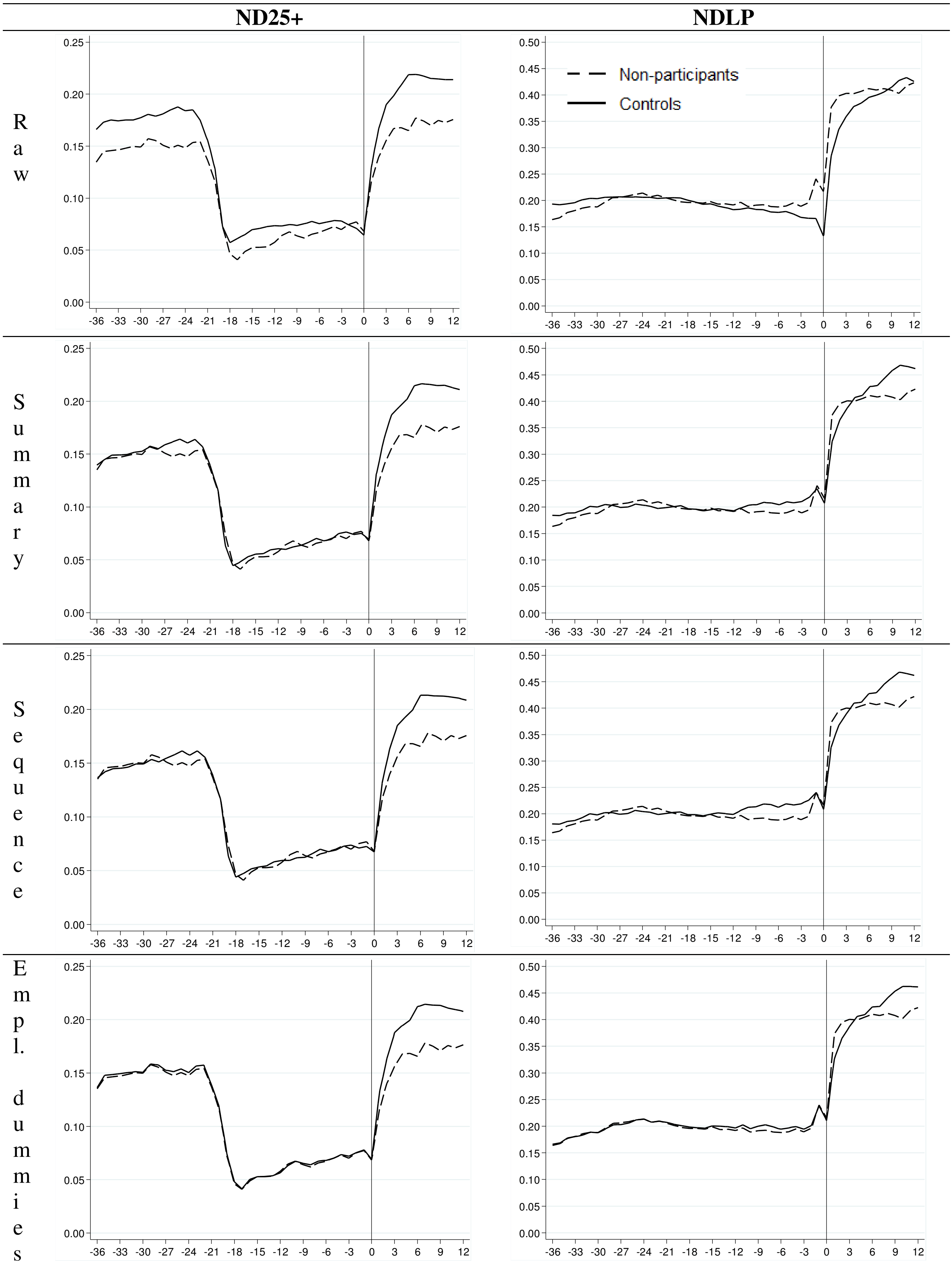
Notes: 'Raw' are outcome differences between non-participants and participants. 'OLS' and 'Matching' are adjusted differences, controlling for all observables in Table 2, the parsimonious summary of labour market histories, and optionally for pre-treatment 2003/04 fiscal year earnings.

A closer look at employment and benefit receipt status over time

Figure A5.1 plots the non-ERA employment rate of participants and non-participants over time, from 36 months before inflow into the ND25+/NDLP program to 12 months post inflow (where only the controls were used to calculate the post-inflow non-ERA outcomes for the participants).

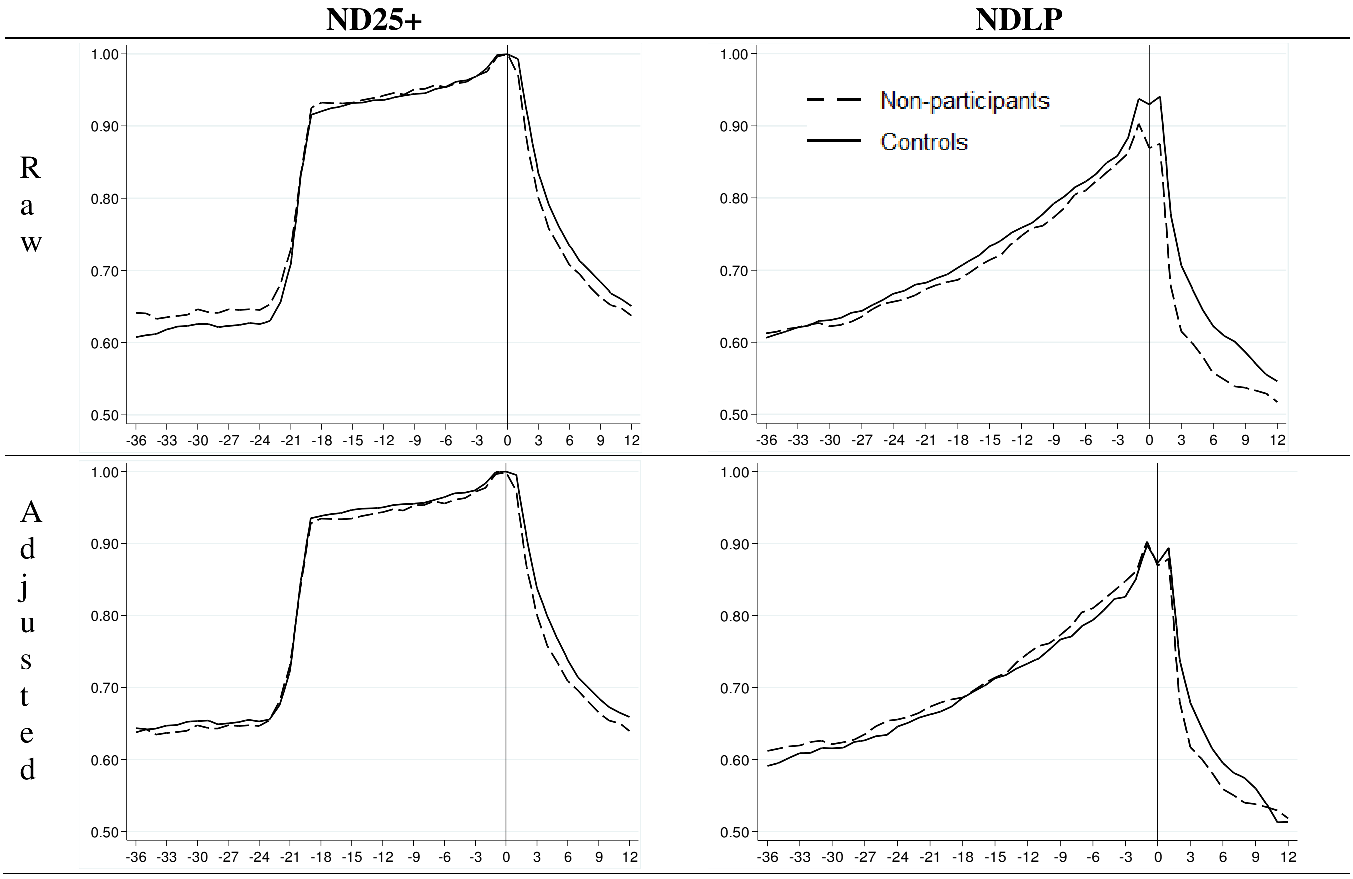
As to the ND25+ group, the first graph shows that the non-participants experience considerably lower employment rates both before and after inflow. (The sharp drop in employment probability at 18 months

Figure A5.1: Non-ERA employment probability by ERA study participation status over time: Raw and adjusted for different ways of constructing labour market histories



Notes: Monthly from 36 months pre- to 12 months post inflow into the New Deal program. For pre-inflow periods, the full sample of participants was used, for post-inflow periods only the controls. All adjusted rates control for all the observables in Table 2, only differing as to how the employment histories were constructed.

Figure A5.2: Benefit receipt probability by ERA study participation status over time (36 months pre- to 12 months post-inflow): Raw and adjusted



Notes: Monthly from 36 months pre- to 12 months post inflow into the New Deal program. For pre-inflow periods, the full sample of participants was used, for post-inflow periods only the controls. Adjusted rates control for all observables in Table 2 and summary employment histories; results were indistinguishable in terms of how histories were constructed.

before inflow is due to the fact that our sample started the ND25+ program, a program which becomes mandatory after 18 months on unemployment benefits). The raw differences in employment probability in the pre-inflow period clearly highlight that sources of selection into the ERA study include employment-relevant characteristics. Three patterns emerge from the remaining graphs showing the adjusted differences. First, controlling for the available observables pretty much eliminates the observed pre-inflow compositional differences. Second, different ways of constructing labour market histories balance the two groups in the pre-inflow period to roughly the same extent; interestingly, monthly dummies are shown to be the best at balancing employment histories, while matching on sequences or on summary measures produces the same amount of (slightly inferior) balancing. Finally, and perhaps most strikingly, no matter how well the two groups have been realigned in terms of the pre-inflow employment rate history, the adjusted post-inflow employment rates remain the same, and indeed extremely similar to the unadjusted rates. In line with the results in Table A5.1, no matter how well balanced the two samples are in terms of pre-inflow characteristics and histories, the raw outcome differences between non-participants and controls remain essentially unchanged after the adjustment.

As to the NDLP group, raw pre-inflow raw differences between participants and non-participants become notable shortly before inflow into NDLP, with non-participants experiencing higher employment rates than participants. Such differences are seen to persist for up to 9 months post inflow. Matching on observables and histories in the form of monthly employment dummies does an impressive job in balancing pre-inflow employment rates; capturing histories using summary indicators performs worse, and using detailed sequences the worst. However, as was the case for the ND25+ group, having balanced the pre-inflow employment rates between the two groups does not help in balancing their post-inflow non-ERA employment rates (indeed, non-participants are now found to perform worse than observationally equivalent controls).

Figure A5.2 shows the same type of charts in terms of benefit dependency. (Time in employment and time on benefits are not mutually exclusive, as individuals can be employed at the same time as claiming a benefit such as income support; this is particularly the case with the WPLS data, which contains no information on the amount of hours worked). For the ND25+ group, larger raw differences between participants and non-participants in the probability of being on benefits are present during the third year before inflow (with non-participants more likely to depend on benefits) and re-emerge after inflow (with non-participants less likely to depend on benefits). Adjusting the groups based on different ways of creating histories balances the pre-inflow benefit incidence equally well, and indeed very well; however the post-inflow differences remain unchanged by the adjustment. In the case of the NDLP group, non-participants are less likely to be on benefits before starting the NDLP program and even more so after inflow. Again, the adjustment works well in balancing benefit receipt history but is found to only decrease post-inflow bias.

Selection into the ERA study and how it is captured in the available data

Summarising the discussion in Sections 2.2 and 5.4, there appeared to be two main sets of factors driving selection:

1) predicted short-term non-ERA job entry probability, driving diversion incentives for caseworkers (either positively or negatively depending on whether they were New Deal or ERA advisers), as well as refusal decisions of individuals (who were more likely to refuse participation in the study if they had been unemployed for a long time and were thus finding it difficult to envisage what might happen after they obtained a job or else if they were feeling close to getting a job and did not want to stay in touch with the employment office); and
2) specific traits and attitudes driving individuals' refusal decisions such as a strong antipathy to government and feeling alienated from systems of support, being resistant to change or taking risks, enjoying being able to refuse to do something in the context of a mandatory program, possibly being engaged in benefit fraud.

Our prior beliefs were that past labour market histories in terms of employment, benefit receipt and program participation over the previous three years would indeed capture both of these sets of selection drivers. First, (non-ERA) employment histories should closely reflect and hence capture short-term (non-ERA) employment outcomes. Secondly, past histories are often convincingly viewed as encapsulating important traits of individuals such as tastes for leisure, discipline and work commitment, accumulated experience, tenure and on-the-job training, family commitments, health status and reliance on government passive and active support. Indeed whether ND25+ entrants volunteered for the Gateway ahead of time as well as variables capturing the extent of past participation in voluntary employment programs should reflect an individual's willingness to improve one's circumstances and general attitudes to systems of labour market support.

It was thus a great surprise that such an extensive set of variables aimed at capturing the underlying factors driving selection did such a poor job in addressing it.

Comparison to Dolton and Smith (2011)

While the non-experimental analysis of the NDLP program by Dolton and Smith (2011), DS in the following, is the closest and most comparable to the one in this paper, a few notable differences ought to be highlighted.

First, the selection problem is different, both in terms of the treatment into which selection takes place, the population subject to selection and the selection decision unit. DS address self-selection into the voluntary NDLP program by eligible lone parents; here we address selection (mainly) by advisers into the ERA study among those lone parents who had already volunteered for the NDLP program.

This paper thus considers selection into an additional treatment that was offered to the sample that had self-selected in the DS analysis (though DS analyze data from the very beginning of the NDLP implementation in 2000-01; selection into that program may have been different at that point than it was in 2003-04 during ERA). As to the selection decision unit in this paper, most (87%) of the selection into ERA was driven by the advisers performing the intake, rather than the individual lone parents themselves (see Table 3).

Second, the available conditioning variables are different, as DS did not have access to employment or earnings measures either from surveys or administrative data but did have access, for much of their sample, to survey measures of attitudes to work. The history variables constructed by DS have a finer temporal coarseness but only encompass history of benefit receipt over the previous 14 months, while the ones used in this paper consider time on benefits, time in employment and time spent neither in employment nor on benefits over the previous 36 months.

Finally, it has to be stressed that DS do not claim to have removed all of the selection bias in their estimates, only to have moved them towards more plausible values compared to the official evaluation. They find that flexibly conditioning on histories rather on summary indicators is important in this respect. By contrast, Table A5.1 clearly highlights how in our case the way histories are modelled does not make any difference whatsoever in decreasing selection bias (a bias which the availability of the experimental control group allows us to measure).