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

This draft: January, 2012

Barbara Sianesi

Institute for Fiscal Studies

Abstract

The UK Employment Retention and Advancement (ERA) programme has been evaluated by a large-scale randomised experiment. It has however emerged that due to the experimental set-up over one quarter of the eligible population was not represented in the experiment: some eligibles actively refused to be randomly assigned, while some were somehow not even offered the possibility to participate in random assignment and hence in ERA. The fact that ERA was a study and involved random assignment has significantly altered how the intake as a whole was handled, as well as the nature of the adviser/individual interaction in a way that would not have been the case had ERA been normal policy. The pool of participants has been both reduced and altered, which is likely to have led to some randomisation bias or, alternatively, to some loss in external validity in the experimental estimate for the effect on the eligible population. The beauty of the ERA set-up and data is that it offers the rare chance to formally measure the extent of randomisation bias or the loss in external validity. Specifically, the key objective of the paper is to quantify the impact that the full ERA eligible population would have been likely to experience had they been offered the chance to participate in ERA, and to assess how this impact for the full eligible group relates to the experimental impact estimated on the potentially self-selected and advisor-selected subgroup of study participants. We separately consider how to deal with non-participation when follow-up information on the outcomes of the non-participants is available (administrative data) or not available (survey data such as earnings). Non-response to the survey and/or to the earnings question among survey respondents can create additional issues when trying to recover the earnings effect of ERA for the full eligible population.

JEL codes: C21, J18, J38

Acknowledgments: I would like to acknowledge the many helpful suggestions from Costas Meghir, as well as Electra Small for her hard data work and Ingun Borg, Mike Daly, Christine Daniels, Richard Dorsett, Lesley Hoggart, Phil Robins, Jim Riccio and Stacy Sharman for detailed comments on previous drafts. The usual disclaimer applies.

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

1. Introduction

Carefully planned and administered randomised social experiments arguably represent the most reliable method for evaluating whether a programme works, on average, for its participants. Since eligible individuals are allocated randomly between a programme group receiving the services and a control group not receiving them, under reasonable assumptions any systematic difference in later outcomes observed between the two groups can be attributed to the programme.

While experimental studies have played an important role in the design of US welfare and training programmes, they have not been widely used in the UK. A recent exception is the Employment Retention and Advancement (ERA) demonstration, which ran in six Jobcentre Plus districts across the UK between October 2003 and October 2007. The demonstration was set up to test the effectiveness of an innovative package of time-limited support once in work, combining job coaching and advisory services with a new set of financial incentives rewarding sustained full-time work, as well as completing training or education whilst employed. Eligible for this initiative were longer-term unemployed people over the age of 25 in receipt of Jobseeker's Allowance who were mandated to enter the New Deal 25 Plus (ND25+) programme as well as lone parents who volunteered for the New Deal for Lone Parents (NDLP) programme.[1] With over 16,000 individuals being randomly assigned in six districts over one year, the ERA study represented at its inception the largest randomised controlled trial of a social programme in the UK.

Since ERA offered of a package of support once in work, *all* individuals flowing into ND25+ and NDLP in the six evaluation districts during the one-year intake window should automatically have become eligible to be offered the ERA package. It has however emerged that only parts of the target population have entered the evaluation sample: some eligibles actively refused to be randomly assigned and to take part in the experimental evaluation (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 arguably be interested in assessing the average impact of offering ERA services and incentives for all those eligible to receive such an offer, the experimental evaluation can provide unbiased impact estimates only for the ERA study participants – those who reached the randomisation stage and agreed to participate in the demonstration. The concern is that this subgroup may potentially be a selective one, not representative of the full eligible population in the ERA districts who would have been eligible for ERA had it been an official national policy.[2]

Technically, the non-participation problem can be viewed in two ways.

Heckman (1992) and Heckman *et al.* (1999) call "randomization bias" the violation of the indentifying assumption for social experiments ruling out any changes in the programme impact

[1] We focus on the two main ERA target groups, representing 83% of all ERA study participants. The third group – lone parents working part-time and in receipt of Working Tax Credit who have volunteered for ERA – is not considered due to its conceptually different set-up coupled with lack of data.

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

as well as changes in the programme participation process due to the presence of random assignment *per se*. As described in detail in Section 2, the fact that ERA was a study and involved random assignment has significantly altered how the intake as a whole was handled, as well as the nature of the adviser/New Deal entrant interaction in a way that would not have been the case if ERA had been normal policy. Indeed it was the set-up of the experimental evaluation *per se* which gave rise to diverted customers and formal refusers – these eligible customers 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 that would have participated had ERA not been evaluated via random assignment. If the parameter of interest is the impact of offering ERA eligibility on the eligible population, non-participation can thus be seen as potentially introducing randomisation bias in the experimental estimate for the parameter of interest.

Alternatively, if the parameter of interest is the impact of the ERA offer for the sample of participants, non-participation can be viewed as a problem of external validity of the experimental impact estimates, concerning the extent to which the conclusions from the experimental study would hold for – or generalise to – the whole eligible population.

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

The key objective of the paper is thus to recover the causal effect for the full eligible population of making the ERA package available. Specifically, we use non-experimental methods to first quantify the impact that the full ERA eligible New Deal population would have been likely to experience in the year since inflow into the New Deal had they been offered the chance to participate in ERA, and to then assess how this impact for the eligible group relates to the experimental impact estimated on the subgroup of study participants. In most cases, identifying and estimating the average impact on all eligibles requires first identifying and estimating the average ERA impact that the non-participants would have experienced.

After an initial exploratory bounding analysis, we focus on matching and reweighting techniques under the assumption that we observe all outcome-relevant characteristics that drive selection into the ERA study. The rich data we use include individual demographics as well as information on current unemployment spell, detailed labour market histories and local factors. We also suggest simple sensitivity analyses to assess how sensitive the estimates are to straightforward violations of this crucial assumption.

We consider how to deal with non-participation when follow-up information on the outcomes of the non-participants is available (administrative outcome measures) or not available (survey-based outcome measures). Clearly, the latter case will be less informative, and we will have to make more stringent assumptions. Furthermore, non-random non-response to the survey and non-random item non-response among survey respondents potentially create additional issues when trying to recover the effect of ERA on the full eligible population. An in-

teresting feature of our data is that it allows us to test some conditions under which non-response can be safely ignored.
Finally, within this framework we estimate the type of involvement that the non-participants would have had with ERA had they participated in the evaluation study. This allows us to shed some light on the question of whether the non-participants are indeed individuals who even if offered ERA services would not take them up.

The remainder of the paper is organised as follows. We start in Section 2 by outlining how non-participation in the ERA evaluation has come about and summarise the available qualitative and quantitative evidence. Section 3 describes the data, sample definitions and the rich set of variables we have collated from different sources in order to capture key characteristics relating to the individuals themselves, their office and their local area. Our methodological approaches and the type of analyses we perform are presented in Section 4. The results of all the empirical analyses are presented and discussed in Section 5, while Section 6 concludes.

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

2.1 How did non-participation come about

The demonstration was set-up to test the effectiveness of the ERA intervention. ERA is the offer of a package of support. While still unemployed, ERA offers job placement assistance, largely following the same procedures as the regular New Deal programmes. Upon having entered work, customers can avail themselves of a two-year post-employment support of an adviser (ASA) who aims to help them retain their jobs and advance to positions of greater job security and better pay and conditions. Working customers are further eligible to an employment retention bonus of £400 three times a year for up to two years for staying in full-time work for 13 out of every 17 weeks, to training tuition assistance (up to £1,000) and to a bonus (also up to £1,000) for completing training while employed at least part-time, and to access emergency payments to overcome short-term barriers to remain in work.

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

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

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

Taken together, diverted customers and formal refusers make up the group of the “ERA non-participants” – those who whilst being *eligible* for ERA, for some reason or another have not been included in the experimental sample and have thus not participated in the evaluation.

The “ERA study participants” are the group of individuals who were eligible for ERA, were offered the chance to participate in the study *and* agreed to take part in it. These are those making up the evaluation sample, i.e. those who were subsequently randomly assigned either to the programme group, who would receive ERA services and incentives, or to the control group, who would instead only receive the baseline New Deal treatment while unemployed.

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

Qualitative work conducted as part of the ERA evaluation has shed interesting light on the origins and sources of non-participation. In particular, Hall *et al.* (2005) and Walker *et al.* (2006) have looked closely at the assignment and participation process in ERA at selected sites. Based on detailed observations, interviews and discussions with both staff and individuals, the authors have put forward the conjecture that it is quite unlikely for ERA non-participants to be a random subgroup of the two eligible New Deal groups.

Recognising that two parties – the caseworker and the individual – are involved in the decision processes that led to inclusion in the sample of ERA study participants, the discussion of what is known about non-participation from this qualitative work is organized in two parts.

Since the individual can only refuse once having been offered the chance to participate, the individual's decision has direct bearing on the second choice, i.e. the one between participation and formal refusal. On the other hand, the caseworker can affect both types of outcomes: he or she has basically sole decision power as to whom to offer ERA, as well as considerable influence in steering the individual's response to such offer. In an individual case, it might also be lack of understanding of the process from part of the adviser, or even the possibility that the New Deal starting dates (which qualify an individual to be offered ERA) as recorded on the system may not be as precisely perceived by staff.

<u>1. Ensuring that staff randomly assigned all eligible individuals</u>

The six districts could exercise significant discretion in how they organised the ERA recruitment, intake and random assignment processes, so that a number of models ended up being used.[3] Although the expectation in any model was that the intake staff, be it an ERA adviser (ASA) or a New Deal Adviser (PA), would encourage *all* eligible individuals – and encourage all of them *equally hard* – to consent to be randomly assigned and have a chance to participate in ERA, staff could use discretion on two fronts: what individuals to tell about ERA, directly determining the extent of diverted customers, and in what terms to present and market ERA to individuals, thus affecting the likelihood that they would 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.[4] Discretion in their choice of marketing strategy could take various forms, e.g. how 'hard' to

[3] The model closest to the original plan saw ERA intake and random assignment being undertaken by a specifically allocated intake adviser, who had no vested interest in its outcome. In other districts, it was the New Deal Advisers (PAs) who conducted the intake and randomisation, with the ERA advisers (ASAs), being responsible for working with ERA programme group members only after random assignment had taken place. In yet other districts, the ASAs were also responsible, alongside the New Deal PAs, for conducting intake interviews and randomisation. Typically, ASAs in these districts handed over to the New Deal advisers those individuals allocated to the control group and those who had refused to participate in ERA. 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. Furthermore, the intake and randomisation procedures varied over time, in the light of experience and depending on the situation and needs of the district or even single office.

[4] In addition to discretionary choices about how much information to disclose, it also became apparent that probably owing to their greater knowledge of and enthusiasm for ERA, ASAs tended to give clearer explanations of ERA than PAs (Walker *et al.*, 2006, Appendix F).

sell ERA; what features of the programme to mention – in particular whether and in what terms to mention the retention bonus, or whether to selectively emphasise features (e.g. the training bonus) to make ERA more appealing to the particular situation of a given individual; and how far to exploit the misunderstanding that participation in ERA be mandatory.

But why and under what circumstances would caseworkers want to apply such discretion? There could have been situations where the adviser did not deem that the individual would be interested in taking advantage of ERA or would benefit from it. Furthermore, the Jobcentre Plus target structure gave advisers individual-level targets for how many people they moved into work and accordingly rewarded staff 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 who they perceived as clearly likely to move into work quickly. The discussion in Walker et al. (2006) highlights how job entry targets had an asymmetric influence on 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.[5]

It is thus known from this research that ERA non-participants, and especially diverted customers, are not likely to be random subgroups of the eligible population; rather, these are 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*[6], and thus be excluded from the experimental sample of study participants. It is also not fully clear how much individuals actually knew about what they were refusing – according to observations at intake interviews and interviews with the unemployed themselves after those sessions, not much.[7] Had ERA been an official policy, there would have been no need to ask for consent to perform randomisation, nor to severely restrict information on the actual extent of ERA support in order to prevent disappointment among the control group.[8]

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

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

[7] 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 programme group and therefore had been given further details about the services available after random assignment". More generally, "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).

[8] This was relaxed over time, although Walker *et al.* (2006, p.22) conclude that "when invited to participate in ERA, customers would generally have known only that some form of extra help was potentially available if they found work and that they had a 50-50 chance of receiving it".

What is clear from the qualitative work is that recruitment to ERA greatly differed between the two New Deal groups. While lone parents on NDLP were all volunteers to that programme 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 Jobcentre Plus. 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 programme. A further possible reason for refusal is being engaged in benefit fraud.

Overall, the available qualitative evidence on refusals suggests that those who declined to join may, in fact, differ in important respects from those who agreed to participate. Formal refusers, especially those amongst the more problematic ND25+ group, appeared to have weaker job prospects and poorer attitudes than the average New Deal entrant.

In addition, the refusal rate was observed to fall later on during random assignment, likely due to a combination of enhanced adviser experience at selling ERA and the permission to mention the monetary incentives. The refusal process is thus likely to have changed over the intake window, with refusers in later entry cohorts presumably forming quite a selective group.

Finally, as mentioned above, the incentive structure arising from Jobcentre Plus job entry targets had an asymmetric influence on New Deal and on ERA advisers in terms of how hard to sell ERA. Specifically, when New Deal advisers undertook the intake interviews, they could benefit if job-ready individuals refused to participate in ERA and those with bad prospects consented. Conversely, when ERA advisers were leading the intake process, they could benefit if individuals with bad job prospects formally refused to participate, while those with good prospects agreed to participate.

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) take the important initial step to thoroughly explore how representative (or policy relevant) the group is for whom one can calculate experimental estimates by understanding both how large and how selective the non-participating groups are. The incidence, composition, determinants and selectivity of non-participation were found to be markedly different between the ND25+ and NDLP intake groups, as well as across districts. As to incidence, non-participation overall was lower amongst the ND25+ group (23% of the eligible group) than amongst NDLP entrants (over 30%). In terms of composition, 9% of all ND25+ eligibles appear to have been diverted and 14% formally refused. By contrast, over one quarter (26.4%) of all eligible NDLP entrants in the six districts appear to have been diverted, while only 4% formally refused. The bulk of non-participation in the ND25+ group was thus due to formal refusals (59%), while in the NDLP group by diverted customers (86%).

There was also marked variation in the incidence 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 them diverted customers. Focusing on the ND25+ group, the performance of Scotland and North West England is particularly remarkable, with *not one single diverted customer*, while North East England stands out with *over one quarter* of eligible ND25+ participants formally refusing to give their consent to being randomly assigned. A very strong and interesting role of Jobcentre Plus office affiliation was also uncovered in determining both ERA offer and consenting choice, though as expected it was stronger in the former. Over time, a fall in the formal refusal rate was observed for both intake groups, likely to reflect increased adviser experience and confidence in selling ERA, as well as the permission to mention ERA financial incentives.

Most of the explained variation in ERA offer, acceptance and participation was accounted for by an individual's district, office affiliation and inflow month, underscoring the key role played by local practices and constraints. Individual employment prospects, as well as attitudes towards and past participation in government programmes were however also found to matter, leaving only a residual role to demographic characteristics.

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

To conclude, the non-participation problem seems to be a relevant one, both in terms of its incidence and of the diversity of the excluded groups. Overall, the NDLP study participants are on average slightly more likely to depend on government benefits than the average lone parent volunteering for NDLP. By contrast, the ND25+study participants are significantly easier to employ than the average ND25+ entrant; ERA advisers are thus working with a group which is considerably more advantaged than the average population, which potentially raises a creaming question for the experiment.

The fact that ERA was a study and involved random assignment has thus significantly altered how the intake as a whole was handled in the context of Jobcentre Plus, as well as the nature of the adviser/New Deal entrant interaction in a way that would not have been the case if ERA had been normal policy. The fact that the pool of participants has been both reduced and altered is likely to have led to some randomisation bias or, alternatively, to some loss in external validity in the experimental estimate for the effect on the eligible population. The analyses in the present paper aim to formally assess and quantify the amount of non-participation bias or the loss in external validity.

3. Data and sample definition

3.1 Data

A number of data files have been put together for the analysis. The administrative data held by the Department for Work and Pensions (DWP) on ND25+ and NDLP entrants provided us with the sampling frame. We extracted files for all cases identified as having entered these New Deal programmes 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 programme participation as well as a number of other relevant individual characteristics.

We have then merged these files with other DWP data on benefit and employment spells – the Work and Pensions Longitudinal Study (WPLS) dataset. This is a relatively recently released, spell-level dataset that contains DWP information about time on benefits and HMRC records about time in employment. These administrative records have been used to construct both detailed labour market histories and outcome measures.

We have further combined the administrative data with data collected specifically for the ERA experimental evaluation in the form of the Basic Information Form (BIF). This file contains all New Deal customers who were approached for recruitment into ERA, including the identifier of those who formally refused to participate. Of this data we mainly use information on customers' decisions as to participation in ERA, as well as the outcome of random assignment (control/programme group) for those who agreed to participate in the study.

We have finally merged in local-area level data (Census, travel-to-work and super-output area data). In section 3.3 we summarise the extensive variables we have selected and derived from all of these sources.

3.2 Sample

To perform our analyses aiming at estimating the impact of ERA for all eligibles, we need to define exactly the criteria determining eligibility and to be able to identify the relevant individuals in the data.[9] We consider as *eligible* for ERA:

1. those who have become 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.

The random (or sample intake) assignment window is actually district- and customer group-specific, since one district started conducting random assignment later than the others and some districts stopped conducting random assignment for some groups earlier. The period when each district was conducting random assignment was as follows:

North West England:	3 January 2004	to 31 January 2005
All other districts:	1 November 2003	to 31 October 2004, with the exception of to 21 August 2004 for NDLP in South East Wales.

The analysis also considers ERA impacts on outcomes (e.g. earnings) collected from the ERA 12-month customer survey. This survey covers the experiences of a sample of the programme

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

group and the control group during the first 12 months following individuals' date of random assignment, with most interviews occurring from December 2004 through November 2005. The intake period for individuals who are eligible to be surveyed is thus 1 December 2003 (3 January 2004 in North West England) to 30 November 2004. When looking at survey outcomes, we thus consider the intersection of the intake window above with this survey sample:

North West England: 3 January 2004 to 30 November 2004
All other districts: 1 December 2003 to 31 October 2004, with the exception of
to 21 August 2004 for NDLP in Wales.

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, separately for the two customer groups. As mentioned already, the incidence of non-participation was substantial: about *one quarter* (26.6%) of all those eligible to take part in the ERA study did not participate. Non-participation was substantially lower amongst the ND25+ group (23% of all eligibles) than amongst NDLP clients (over 30%). We observe survey outcomes for 31% of ND25+ and 35% and NDLP study participants.

Table 1 Sample breakdown by customer group

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

3.3 Outcomes and observable characteristics

ERA impacts are assessed during a 12-month follow-up period in terms of two types of outcome measures: administrative and survey outcomes.

As to the former, data on employment and benefits receipt is available from administrative records for the *full* sample of ERA eligibles in the six evaluation districts, i.e. for both for participants and, most importantly for our purposes, for non-participants too. For these administrative outcomes measures we start counting the 12-month follow-up period from the moment individuals flowed in (i.e. from the moment ND25+ customers started the Gateway, or lone parent customers volunteered for NDLP), and consider the probability of having ever been in employment, the total number of days in employment, and the total number of days on benefits during that period.

Survey outcomes were collected from a first-wave customer survey of a sample of ERA participants during the first 12 months following individuals' date of random assignment. The survey outcomes we consider are total earnings and an indicator for earning above £4273 (the overall median calculated from those with positive earnings).

We have put together an extensive collection of variables aimed at capturing the widest possible range 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. Table 2 groups and summarises the various observable factors we use in our analysis; Section 4.4 contains a more detailed discussion of the content of the data.

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+ programme (Spent <540 days on JSA before entering ND25+)
Labour market history (3 years pre-inflow)	Past participation in basic skills, past participation in voluntary programmes (number of previous spells on: NDLP, New Deal for Musicians, New Deal Innovation Fund, New Deal Disabled People, WBLA or Outreach), past participation in ND25+, active benefit history dummies (JSA and compensation from NDYP, ND25+, Employment Zones and WBLA and Basic Skills), inactive benefit history dummies (Income Support and Incapacity Benefits); employment history dummies
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

4. Methodological approaches

4.1 Analysis framework

We start by setting up the framework and introducing some basic notation. Figure 1 highlights the structure of the problem that needs to be addressed, while Box 1 summarises the notation.

The population of interest are those eligible to be offered ERA services. We implicitly condition on this population throughout. The binary variable Q captures the potential selection into the ERA study, with Q=0 denoting individuals who despite being eligible have not been randomly assigned*,* and Q=1 denoting the study participants. Study participants make up the experimental group which was randomly assigned between a programme group who was offered ERA services (R=1) and a control group who was not (R=0).

The problem we want to address is that because of diversion and of refusal to be randomly assigned, the population under the experimental evaluation (Q=1) does not correspond to the full eligible population, made up by the (Q=1) *and* (Q=0) groups. If selection has taken place into the participating group, the composition of participants will be different from the composition of the eligible population, and impacts estimated on participants will not necessarily be representative of the impacts that the eligibles would have experienced.

Further, let the indicator S denotes 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 will refer to them as "respondents". S=0 by contrast denotes non-surveyed or survey non-respondents or item non-respondents among participants ("non-respondents"). As Figure 1 highlights, it is possible for some selection to have taken place among participants into the responding sample.

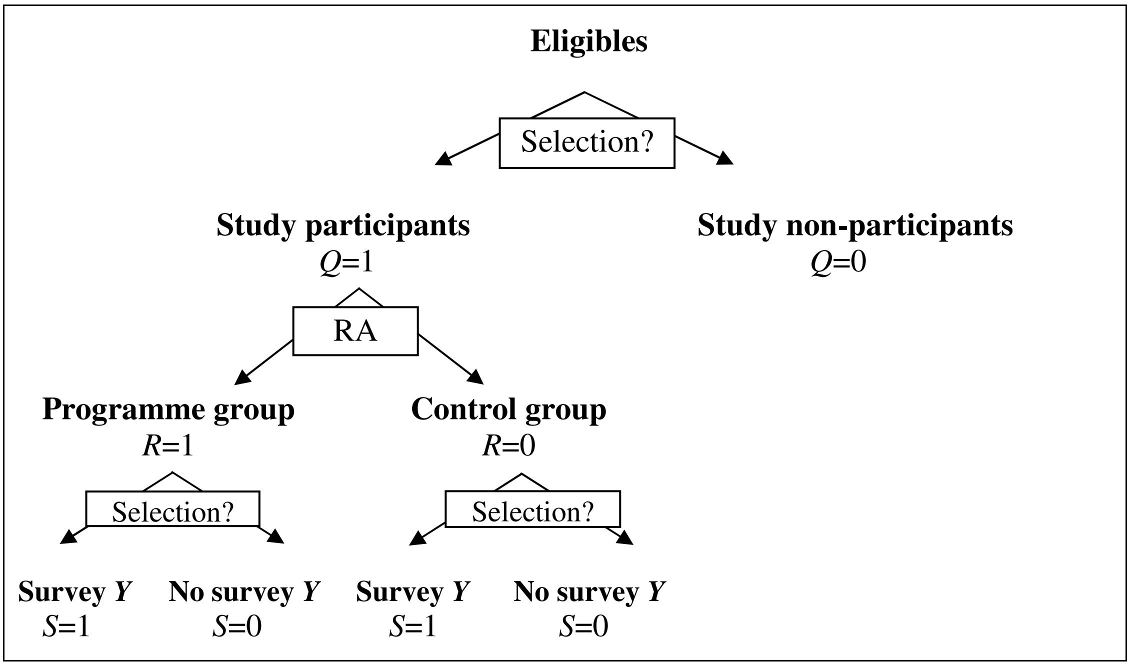
Let $p \equiv P(Q=0)$ be the probability of non-participation among the eligibles. This is directly identified in the data by the proportion of non-participants among the eligibles (see Table 1).

Define the 'propensity score', i.e. the probability that an eligible customer with characteristics $X=x$ does not participate in the ERA study, as: $p(x) \equiv P(Q=0|X=x) = P(Q=0|Q=0 \vee Q=1, X=x)$.

Turning now to outcomes, we follow the potential outcome framework and let Y_1 be the outcome if the individual were offered ERA services (i.e. the treatment outcome) and Y_0 the outcome if the individual were not offered ERA services (i.e. the no-treatment outcome). The observed outcome is denoted by Y. The individual causal effect of ERA is defined as the difference between the two potential outcomes, $Y_1 - Y_0$.

Throughout we need to assume that treatment and no-treatment outcomes among the eligibles are not affected by whether an individual is *offered the chance* to participate in the ERA study or not. In other words, participants and non-participants may be drawn from different parts of the distributions of observed and unobserved characteristics, but the mere fact of being offered the chance to participate in the ERA study does not change the relationship between characteristics on the one hand and treatment and no-treatment outcomes on the other. Formally, this requires the potential outcomes of individual i not to be indexed by Q, i.e.: $Y_{1Qi} = Y_{1i}$ and $Y_{0Qi} = Y_{0i}$ for Q=0, 1.

Figure 1: Simplified structure of the problem



Box 1: Notation

Q=1	ERA study participants (the experimental sample)
Q=0	non-participants
R=1	individuals randomly assigned to the programme group conditional on Q=1
R=0	individuals randomly assigned to the control group conditional on Q=1
S=1	observe survey outcomes conditional on Q=1 ("respondents")
S=0	do not observe survey outcomes conditional on Q=1 ("non-respondents")
X	observed characteristics
p	probability of non-participation among eligibles
$p(x)$	propensity score: $P(Q=0 \mid X=x)$
Y_1	potential outcome if offered ERA services
Y_0	potential outcome if not offered ERA services
Y	observed outcome
ATE	average ERA effect on *all* ERA eligibles (parameter of interest)
ATE_1	average ERA effect on ERA study participants (experimental estimate)
ATE_0	average ERA effect on non-participants
$ATE_{S=1}$	average ERA effect on respondents
$ATE_{S=0}$	average ERA effect on non-respondents
$\Delta_{S=1}$	experimental contrast for respondents

The parameter we are interested in is the average effect of ERA on the *full* ERA eligible population (an Average Treatment Effect): $ATE \equiv E(Y_1 - Y_0)$.

What we can however directly identify from the available experimental data is the average effect of ERA for participants in the experiment. This is because the experiment provides the average effect of the programme for individuals who have been randomly assigned, which due to the randomness of R within the Q =1 group is identified by the difference in the mean outcomes of programme and control groups:

$$\begin{aligned} ATE_1 &\equiv E(Y_1 - Y_0 \mid Q{=}1) = E(Y_1 \mid Q{=}1) - E(Y_0 \mid Q{=}1) = E(Y_1 \mid Q{=}1, R{=}1) - E(Y_0 \mid Q{=}1, R{=}0) \\ &= E(Y \mid R{=}1) - E(Y \mid R{=}0) \end{aligned}$$

Denote the average impact of ERA on the excluded eligibles (i.e. on the non-participants) by

$$ATE_0 \equiv E(Y_1 - Y_0 \mid Q{=}0) = E(Y_1 - Y_0 \mid Q{=}0)$$

Using the law of iterated expectations, the parameters ATE and ATE_1 are linked according to:

$$ATE = E(Y_1 - Y_0 \mid Q{=}1)\ P(Q{=}1) + E(Y_1 - Y_0 \mid Q{=}0)\ P(Q{=}0) \equiv (1{-}p){\cdot}ATE_1 + p{\cdot}ATE_0 \qquad (1)$$

Equation (1) simply states that the parameter of interest, i.e. the average impact of ERA on all the eligibles in the six districts, is given by a weighted average of the parameter we can reliably estimate using random assignment, i.e. the impact on participants, and of the impact on non-participants, with weights given by the relative share of participants and non-participants within the eligible pool.

There are two alternative conditions under which the average impact for participants would be the same as the average impact for the full eligible population even in the presence of a non-negligible share of non-participants. The first situation is one of homogeneous ERA impacts, that is $Y_{1i} - Y_{0i} = \beta$ for all eligible individuals i. The second case is one where impacts might be heterogeneous, i.e. $Y_{1i} - Y_{0i} = \beta_i$, the decisions of eligibles or caseworkers on ERA participation are not affected by the realised individual gain from receiving ERA. Formally:

$$\begin{gathered} \text{if } Q \perp (Y_1 - Y_0), \text{ i.e. if } P(Q{=}1 \mid Y_1 - Y_0) = P(Q{=}1) \\ \text{then } E(Y_1 - Y_0 \mid Q{=}1) = E(Y_1 - Y_0 \mid Q{=}0) = E(Y_1 - Y_0). \end{gathered}$$

In either of these cases, the ATE_1 based on experimental data would thus still provide unbiased estimates of the ATE of interest.

We separately consider how to deal with non-participants both when follow-up information on their outcomes is available (administrative outcomes) and when it is not (survey outcomes). The implications of these two situations on equation (1) are as follows.

In case of administrative data, equation (1) becomes:

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

as the observed outcome of the non-participants corresponds to their no-treatment outcome: $E(Y_0 \mid Q{=}0) = E(Y \mid Q{=}0)$.

In case of survey outcomes, both treatment and no-treatment outcomes of the non-participants remain unobserved. Furthermore, in the presence of non-random non-response among ERA study participants, ATE_1 itself will in general remain unobserved:

$$ATE = (1-p)\cdot \boldsymbol{ATE_1} + p\cdot \boldsymbol{E(Y_1 - Y_0 \mid Q{=}0)} \qquad (1b)$$

4.2 Survey outcomes: Survey and item non-response

Survey outcomes Y, in particular earnings, are only observed for a subsample of participants, i.e. those survey respondents who answered the earnings question.

Define the respondents (S=1) as those ERA study participants (Q=1) with non-missing survey outcome information, as non-respondents those ERA study participants (Q=1, S=0) with missing survey outcome information – whatever the reason (not randomly selected for the survey, not contactable, refused to be interviewed, were interviewed but did not fill in the earnings question). Note thus that 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.

In addition to the loss in precision resulting in a reduction of the study's statistical power to detect effects, non-response raises two important validity issues for the evaluation of earnings impacts:

1. *Internal validity*: if the programme and control group experience systematically different non-response, the responding programme and control groups are no longer comparable to one other. In this case the benefits of the original random assignment are lost, and a comparison of the responding programme group members and the responding control group members no longer provides unbiased impact estimates (for the respondents).

2. *External validity*: even if the responding programme and control group members have maintained comparability to one another so that the experimental contrast recovers the average impact for respondents, how do they relate to the original sample? If the responding sample differs substantially from the original one, the results might not generalize to the original target population.

Define $\Delta_{S=1}$ as the experimental contrast calculated on those participants who responded to the earnings question:

$$\Delta_{S=1} \equiv E(Y \mid Q{=}1, S{=}1, R{=}1) - E(Y \mid Q{=}1, S{=}1, R{=}0)$$

$\Delta_{S=1}$ is identified in our data, but we are interested in ATE_1 as one of the two components needed to recover the ATE for the full group of eligibles. The question thus that naturally arises is under what conditions the experimental contrast for respondents recovers the ATE for the full group of participants, i.e. $\Delta_{S=1} = ATE_1$.

Although this condition can indeed be tested on administrative outcomes, which are available for the full group of participants (indeed, for the full group of eligibles), whether it resulted to be met or not would not be easy to interpret. In answering this question it is instead useful to separately consider the following two 'causal-inference' issues related to the internal and external validity issues above.

(a) Internal validity: Under what conditions does $\Delta_{S=1}$ recover the *ATE* for respondents, $ATE_{S=1} \equiv E(Y_1 - Y_0 \mid Q=1, S=1) \equiv E(Y_1 - Y_0 \mid S=1)$?

Since the average ERA impact for respondents is not identified without additional assumptions, to exploit random assignment one has to assume that randomisation keeps holding within the responding sample, i.e. that *R* is still random (possibly given *X*) among respondents:

(I-V) $E(Y_1 \mid S=1, R=1) = E(Y_1 \mid S=1, R=0) = E(Y_1 \mid S=1)$
$E(Y_0 \mid S=1, R=1) = E(Y_0 \mid S=1, R=0) = E(Y_0 \mid S=1)$

Under the internal-validity condition (I-V) that even restricting attention to the subgroup of respondents, randomisation still holds, the *ATE* for respondents, $E(Y_1 - Y_0 \mid S=1)$, can be estimated using the experimental contrast, $E(Y \mid S=1, R=1) - E(Y \mid S=1, R=0)$:

$ATE_{S=1} \equiv E(Y_1 - Y_0 \mid S=1) \equiv E(Y_1 \mid S=1) - E(Y_0 \mid S=1)$
$= \text{(I-V)} = E(Y_1 \mid S=1, R=1) - E(Y_0 \mid S=1, R=0) = E(Y \mid S=1, R=1) - E(Y \mid S=1, R=0) \equiv \Delta_{S=1}$

Condition (I-V) cannot be directly tested; supporting evidence can however be obtained by assessing whether randomization still holds between the two responding subsamples in terms of their *observed* characteristics.

(b) External validity: Under what conditions can the subsample of respondents be assumed to be a representative subsample of the ERA study participants, in the sense that the ATE among respondents is the same as the ATE for the full group of participants, i.e. $ATE_{S=1} = ATE_1$?

The average ERA impact is the same for the full sample of participants and for those participants who responded to the survey if participants do not select into responding based on ERA impacts. Formally:

(E-V) $E(Y_1 - Y_0 \mid Q=1) = E(Y_1 - Y_0 \mid Q=1, S=1)$

Since the impact for respondents is not identified *a priori*, to 'test' condition (E-V) one has first to assume that condition (I-V) holds. Under (I-V), condition (E-V) can be tested on administrative data as:
$E(Y_1 \mid Q=1, R=1) - E(Y_0 \mid Q=1, R=0) = E(Y_1 \mid Q=1, R=1, S=1) - E(Y_0 \mid Q=1, R=0, S=1)$

Note that under (I-V), condition (E-V) is implied by the stronger set of conditions:
(E-V') (a) $E(Y_1 \mid Q=1, R=1) = E(Y_1 \mid Q=1, R=1, S=1) = E(Y_1 \mid Q=1, R=1, S=0)$
(b) $E(Y_0 \mid Q=1, R=0) = E(Y_0 \mid Q=1, R=0, S=1) = E(Y_0 \mid Q=1, R=0, S=0)$

Conditional on random assignment status, non-response is unrelated to potential outcomes, i.e. programme group members do not select into responding based on treatment outcomes, nor do control group members select into responding based on no-treatment outcomes. Put differently, programme and control group members who respond are not selected on outcome-relevant variables. Assumption (E-V') thus rules out selection on outcome-relevant unobservables into responding to the earnings question conditional on random assignment status.

Like assumption (E-V), assumption (E-V') can be tested on administrative outcomes. This is accomplished by testing whether (possibly controlling for observables X), the administrative outcomes of those programme (control) group members who responded to the survey are statistically different from the outcomes of those programme (control) group members for whom we do not observe the survey outcomes.

To conclude, the experimental contrast for respondents, $\Delta_{S=1}$, which is readily obtained from the data, would recover the ATE for the full group of participants, ATE_1, under (I-V) and either (E-V') and/or (E-V). In this case, non-response can be ignored in calculating the average effect on earnings for participants.

It would be hard to believe, though possible, that $\Delta_{S=1}$ just happens to coincide with ATE_1 on administrative outcomes – or that condition (E-V') is met –, even without the need to give a causal interpretation to $\Delta_{S=1}$ (via (I-V)). If there is good support for (I-V), though, the evidence is likely to be more robust.

4.3 Bounds without assumptions on the selection process

For this type of analysis, outcomes need to be bounded. To fix ideas, suppose in the following that the outcome Y (e.g. employment probability) is bounded between 0 and 1.

When follow-up data on the non-participants are available, equation (1a) shows that bounds for the parameter of interest can be constructed as $ATE \in [\underline{ATE}, \overline{ATE}]$, where $\underline{ATE}$ = (1–p)·ATE_1 – p·$E(Y \mid Q=0)$ is the worst-case scenario (non-participants would all be non-employed had they received ERA, i.e. $E(Y_1 \mid Q=0) = 0$) and $\overline{ATE}$ = (1–p)·ATE_1 + p·(1–$E(Y \mid Q=0)$) is the best-case scenario (non-participants would all be employed had they received ERA, i.e. $E(Y_1 \mid Q=0) = 1$). The width of the bounds for the ATE is given by p, the proportion of non-participants among the eligibles. (If there were none, the upper and lower bounds would trivially collapse on the point estimate $ATE_1 = ATE$).

We can further explore how sensitive the estimate of the ATE is to assumptions about the selection process into the group of study participants, as reflected by assumptions on the relative magnitude of $E(Y_1 \mid Q=0)$ and $E(Y_1 \mid Q=1)$. Specifically, we can thus calculate the ATE as a function of θ, ATE_θ, for various values of θ (θ=0.5, …, 1.50) assuming that:

$$E(Y_1 \mid Q=0) = \theta E(Y_1 \mid Q=1) \; (= \theta E(Y \mid Q=1, R=1))$$

i.e. that the average ERA-treatment outcome that the non-participants would have experienced had they participated in the study is θ times the average treatment outcome experienced by the participants, where the latter is identified by the actual outcomes of the randomised programme group subset of the participants.

By varying the values of θ, we can depict different types of selection processes. In particular, θ=1 models a situation where the decisions to participate in the ERA study are unrelated to treatment outcomes; θ<1 models negative selection into the non-participants sample (non-participants would have experienced on average lower treatment outcomes than what the participants experience), while θ>1 represents positive selection.

From equation (1a), the *ATE* as a function of θ, is

$$ATE_{\theta} = (1-p)\cdot ATE_1 + p\cdot\{\,\theta E(Y \mid R=1) - E(Y \mid Q=0)\}$$

Thus, ATE_{θ} increases, and linearly, with θ.

The minimum allowable θ for our outcomes is 0, the maximum allowable $\theta = 1/E(Y_1 \mid \text{Q}=1, X)$ for the binary outcome Y we consider.

If no follow-up information is available on the non-participants, we have to construct bounds based on (1b). It follows that $\underline{ATE} = (1-p)\cdot ATE_1 - p$ and $\overline{ATE} = (1-p)\cdot ATE_1 + p$. The width of the bounds for the *ATE* is now $2\cdot p$, double as large as when we did observe the outcomes of the non-participants. In case non-response cannot be ignored, the bounds will necessarily – and trivially – be the widest possible ones, and unrelated to data content: $ATE \in [-1, 1]$.

4.4 Point estimate under selection on observables

The approaches outlined in this section allow point identification of the average ERA impact for the non-participants (and hence for all eligibles) which can only take into account *observed* differences between non-participants and ERA study participants. To the extent that unobserved differences between the two groups are important determinants of subsequent labour market outcomes, these will erroneously show up as part of the ERA impact estimates. The reliability of such estimates thus crucially depends on the range and quality of characteristics observed. Section 3.3 has summarised the data at our disposal; here we provide a brief discussion of its content in relation to the estimation problem we face.

All our outcomes of interest – employment probabilities and durations, reliance on benefits and earnings – are related to labour market performance. As listed in Table 1, we rely on an extensive collection of individual, office and local area characteristics that are most likely to affect individuals' labour market performance, and that might potentially have affected participation into the ERA study.

In addition to a number of individual demographic characteristics contained in the administrative data (gender, age, ethnicity, partner and children, disability and illness), we have summarised information on a customers' current unemployment spell, including in particular indicators of a very recent/current employment spell, how long it took them to start the Gateway or volunteer for NDLP once having become mandatory for it or being told about it, and of whether ND25+ entrants volunteered for the Gateway ahead of time.

We have further constructed three years' worth of labour market history, with variables summarising the proportion of time employed and the proportion spent on benefits, separately on active benefits (JSA and compensation whilst on a labour market programme) and inactive benefits (Income Support and Incapacity Benefits). We have also created variables capturing the extent of past participation in voluntary employment programmes (as a crude indicator of willingness to improve one's circumstances), in the ND25+ (a mandatory programme) and in Basic Skills (a programme designed to address basic literacy, numeracy and IT skills).

The Census has provided us with information on local labour market conditions (specifically, travel-to-work area unemployment rates), as well as on the deprivation of the area the customer lives in (index of local deprivation). Additionally, we have 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 participating in the ERA study as well as on subsequent labour market outcomes.

Despite offering such rich and detailed information, none of the available administrative data contain reliable information on education – which thus remains an unobservable in our data, together with "innate ability", discipline or work commitment. The previous literature has however indicated the potential for detailed labour market histories (like those we have constructed) to help proxy such unobserved traits and thus to eliminate much of the bias due to unobservables (see for example, Dolton *et al.*, 2008, Heckman and Smith, 1999, Heckman *et al.*, 1998, and Heckman *et al.*, 1999).[10]

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

To obtain a point estimate of the *ATE*, equation (1a) shows that we need to identify $E(Y_1|Q=0)$, the treatment outcome of the non-participants.

This problem is akin to getting the average treatment effect on the non-treated using matching methods, where invoking the "selection-on-observables" assumption, $E(Y_1 \mid Q=0)$ is estimated based on the (observed) treatment outcome of the participants, $E(Y_1 \mid Q=1) = E(Y \mid Q=1, R=1)$.

In this case, we allow the effect (or treatment outcome) to depend on observable characteristics X in an arbitrary way, as well as for eligible individuals to decide to participate in the experiment based on these Xs.

To clarify the assumptions required, specialise the model as follows (note that additive separability is not required for matching).

$$Y_{1i} = m_1(X_i) + u_i + b_i$$
$$Y_{0i} = m_0(X_i) + u_i$$

where $Y_{1i} - Y_{0i} \equiv \beta_i = [m_1(X_i) - m_0(X_i)] + b_i \equiv b(X_i) + b_i$.

In this set-up, β_i, the individual impact from receiving ERA services, is allowed to be heterogeneous across individuals in both observable and unobservable dimensions: $b(X_i)$ represents the impact for individuals with characteristics X_i and thus captures observable heterogeneity in effects; b_i represents the individual-specific unobserved impact conditional on X_i. The unobserved component u_i represents some unobservable individual trait, such as ability or motivation, that affects the outcome irrespective of treatment receipt.

[10] For their main analysis of the NDLP programme, Dolton *et al.* (2008) rely on the same administrative data we use. When using a subset of their sample for whom detailed additional survey information (including a variety of attitudinal measures) is available, they find that such variables in fact add little to the analysis once the lagged outcomes available in the main administrative data are controlled for. They interpret this finding as indicative of the fact that outcome histories capture these otherwise unobserved factors and supporting of their approach based on the selection-on-observables assumption.

Assume that for the eligibles, selection into Q is not based on the unobserved, person-specific component of the impact of ERA b, nor on unobserved 'ability' u for given observable characteristics X:

$$Q_i \perp (b_i, u_i) \mid X_i$$

This ensures that the "selection-on-observables" assumption (A1) is met:

$$\text{(A1)} \qquad E(Y_1 \mid Q{=}0, X) = E(Y_1 \mid Q{=}1, X)$$

To give (A1) empirical content, we also need to assume the existence of common support (i.e. overlap in the distribution of observed characteristics X between participants and non-participants:

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

Specifically, the experimental evaluation cannot provide estimates of the impact of ERA for individuals with observed characteristics $\tilde{X}$ if no participant displays those values. In other words, although there may be *eligibles* with characteristics $\tilde{X}$, if the selection into the ERA experiment is such that nobody with characteristics $\tilde{X}$ is offered ERA or consents to take part so that $P(Q{=}1 \mid \tilde{X}) = 0$, we cannot identify the effect for this subset of eligibles (unless under some arbitrary functional form assumption that allows us to extrapolate).

We can then predict $E(Y_1 \mid Q{=}0)$ as:

$$\begin{aligned} E(Y_1 \mid Q{=}0) &= E_X[E(Y_1 \mid Q{=}0, X) \mid Q{=}0] = \text{(A1)} = E_X[E(Y_1 \mid Q{=}1, X) \mid Q{=}0] \\ &= \text{(RA)} = E_X[E(Y_1 \mid R{=}1, X) \mid Q{=}0] = E_X[E(Y \mid R{=}1, X) \mid Q{=}0] \end{aligned}$$

As for implementation, we can match to each non-participant one or more similar programme group member(s) based on the propensity score $p(x) \equiv P(Q{=}0 \mid X) = P(Q{=}0 \mid Q{=}0 \vee Q{=}1, X)$.

To increase matching quality, it might be worth using only the programme group R=1 (a random hence representative subset of the Q=1 group) rather than the full Q=1 group (i.e. both the programme and control groups) to estimate the propensity score; that is, estimate $p(x)$ based on $P(Q{=}0 \mid Q{=}0 \vee R{=}1, X)$.

Sensitivity analysis

As with the bounds, we can explore how sensitive the estimate of the *ATE* is to straightforward violations of assumption (A1). In particular, replace (A1) by:

$$\text{(A1')} \qquad E(Y_1 \mid Q{=}0, X) = \theta\, E(Y_1 \mid Q{=}1, X)$$

i.e. thus allowing participants and non-participants with the same observed characteristics X to differ in terms of some unobservable, which translates into a proportional difference of θ.

Under (A1'), $E(Y_1 \mid Q{=}0) = \theta\, E[E(Y \mid R{=}1, X) \mid Q{=}0]$, which simply involves rescaling the matched outcome by θ.

Again, ATE_θ increases (linearly) with θ.

The sensitivity analysis can be easily expanded by allowing θ to depend on X via the propensity score $p(X) \equiv P(Q=0 \mid X)$:

(A1") $E(Y_1 \mid Q=0, X=x) = \theta(x)\, E(Y_1 \mid Q=1, X=x)$ where $\theta(x) = \theta(p(x))$

Among customers with the same *a priori* study participation probability p, those who do not participate would have experienced an average treatment outcome which is a fraction $\theta(p)$ of the one of the participants.

Under (A1"), $E(Y_1 \mid Q=0) = \theta\, E_p[\theta(p)\, E(Y \mid R=1, p) \mid Q=0]$

This is most easily performed by stratification matching.

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

This problem is akin to attrition and involves reweighing the outcomes of the ERA study participants (programme and control groups) on the basis of the characteristics X of the full eligible group (i.e. ERA programme group, ERA control group and ERA non-participants) to make them representative – in terms of X – of the full eligible population.

Assume that, once conditioning on observables X, participants and non-participants on average experience the same treatment and no-treatment outcomes, or just that (A2) holds in terms of impacts:

(A2) $E(Y_1 - Y_0 \mid Q=1, X) = E(Y_1 - Y_0 \mid Q=0, X)$ hence $= E(Y_1 - Y_0 \mid X)$

To estimate the ATE of interest, write it as:

$$ATE \equiv E(Y_1 - Y_0) = E_X[\, E(Y_1 - Y_0 \mid X)\,] = \text{(A2)} = E_X[\, E(Y_1 - Y_0 \mid Q=1, X)\,]$$
$$= \text{(RA)} = E_X[E(Y_1 \mid R=1, X)\,] - E_X[E(Y_0 \mid R=0, X)\,] = E_X[E(Y \mid R=1, X)\,] - E_X[E(Y \mid R=0, X) \quad (2)$$

The empirical counterpart can be derived in several ways; we consider in particular reweighing and matching estimators, both ignoring and allowing for selective non-response to the survey and/or to the earnings question.

1) **Reweighing**

The following are the estimators for the ATE derived in Appendices 1a and 1b.

Ignoring survey and item non-response

$$\hat{ATE} = \left[\frac{(1-p)p_R}{\#(R=1)} \sum_{i \in \{R=1\}} \frac{y_i}{(1-p(x_i))\, p_R(x_i)} \right] - \left[\frac{(1-p)(1-p_R)}{\#(R=0)} \sum_{i \in \{R=0\}} \frac{y_i}{(1-p(x_i))(1-p_R(x_i))} \right]$$

Allowing for survey and item non-response

$$A\hat{T}E = \left[\frac{1}{\#(R=1,S=1)} \sum_{i\in\{R=1,S=1\}} \frac{(1-p)p_{RS1}}{(1-p(x_i))\,p_{RS1}(x_i)} y_i \right] - \left[\frac{1}{\#(R=0,S=1)} \sum_{i\in\{R=0,S=1\}} \frac{(1-p)p_{RS0}}{(1-p(x_i))\,p_{RS0}(x_i)} y_i \right]$$

2) **Matching**

An alternative to the method of directly weighting the outcomes of the (responding) participant group so as to reflect the distribution of observables in the original eligible population is to construct the weights by performing matching. The latter offers the advantages that the exact specifications of the propensity score and of the response probabilities are not needed and that one can assess the extent of the actual comparability of groups.

This matching-based idea can be implemented in two ways; either to separately recover the missing ATE_0 and then combining it with the experimental ATE_1 to get the ATE, or to recover the ATE directly. Again, we have considered both a situation where non-response is ignored and one where it is not. Appendices 2a and 2b provide full details of the matching protocols.

3) Analysis of take-up

This section outlines a simple yet informative analysis which aims at estimating the type of involvement that the non-participants would have had with ERA and more generally with Jobcentre Plus had they participated in the evaluation study – either as part of the programme group or of the control group. Specifically, this type of analysis aims to answer:

1. Are the non-participants individuals who even if offered ERA services would not take them up?
2. What kind of involvement would non-participants have had with Jobcentre Plus had they participated in the ERA study and been assigned to the control group?

One can get a handle on these questions by looking at measures of take-up of services and of contact with Jobcentre Plus staff, such as whether the individual has had any type of contact with Jobcentre Plus staff, has received help or advice from Jobcentre Plus staff when not working, has had an education or training course arranged by Jobcentre Plus staff, or, if assigned to the programme group, has heard of the employment and of the training bonuses.

To perform this analysis, the selection-on-observables assumption is again invoked, which requires that, once conditioning on the rich set of observables X, ERA study participants and non-participants would have taken up the same amount of ERA services on average. In other words, this assumption rules out selection into the ERA study based on unobserved characteristics that also affect take-up of ERA services once in the programme group.

The trick is then to simply view such take-up/involvement measures as outcomes, and assess them in essentially the same way as done for employment and earnings outcomes. Specifically, let Y be (a measure of) take-up of ERA services.

To answer question (1), we need to estimate $E(Y_1|Q{=}0)$. Under assumption (A2.a):

$$\begin{aligned} E(Y_1 \mid Q{=}0) &= E_X[\, E(Y_1 \mid Q{=}0, X) \mid Q{=}0] = \text{(A2.a)} = E_X[\, E(Y_1 \mid Q{=}1, X) \mid Q{=}0] \\ &= \text{(RA)} = E_X[\, E(Y \mid R{=}1, X) \mid Q{=}0] \end{aligned}$$

To implement this estimator, one can match to each non-participant one or more ‘similar’ programme group members and take the latter’s reweighted outcomes.

A similar type of analysis can be performed on the non-participants and the control group to answer question (2). It requires that, once conditioning on the observables, ERA study participants and non-participants would on average have had the same involvement with Jobcentre Plus if assigned to the control group.

As a final note, although such take-up/involvement measures are obtained from the 12-month follow-up survey, non-response to these questions is truly negligible (less than 1%), so that it can be safely ignored when performing both types of exercise.

5. Implications of non-participation for the experimental impact estimates

5.1 Experimental findings

This section presents the benchmark experimental findings concerning the average impact of ERA for the participants on a series of outcomes in the first follow-up year.[11] Table 2 displays both the raw experimental contrast and the adjusted impact estimated by linear regression controlling for the observables in Table 1. Although randomisation has worked very well so that the ERA programme and control groups are well-balanced in terms of such characteristics, controlling for them can increase the precision of the experimental impact estimate by reducing the residual variance of the outcome. This seems to be largely the case in this application, where most standard errors decrease following the regression adjustment. Furthermore, the adjustment allows one to control for differences in observables between the programme and the control group that have occurred by chance. This also seems to matter in our application, as impact estimates are often found to change once conditioning on observables.

No impact could be detected on the probability of being employed in the follow-up year. Employment durations have similarly remained unaffected for the NDLP group, while a small positive overall effect of ERA (plus 5 days) has been uncovered for the ND25+ group.

Time spent on benefits appears to have been slightly reduced by the offer of ERA for the NDLP group; once chance imbalances in the observables are controlled for, though, the effect drops into non-significance.

By contrast, the experimental impact of ERA on average earnings in the first follow-up year is estimated to be quite substantial and highly statistically significant for the NDLP group (+£730). For the ND25+ group, the experimental contrast highlights a much smaller impact, which is significant only at the 10% level (+£393). ND25+ customers were also not affected in their probability of earning above the median, while NDLP customers saw a marginal increase of almost 4 percentage points.

[11] These findings do not always correspond to those reported in Dorsett *et al.* (2007). The reasons for any discrepancy are the latter’s use of survey-based rather than administrative outcomes, focus on the survey rather than the full sample, adjustment for survey rather than administrative characteristics and use of a different weighting scheme.

Table 2 Experimental findings

	Raw		*Adjusted*		*N*
	Effect	*Std.Err.*	*Effect*	*Std.Err.*	
ND25+					
Ever employed	0.014	(0.012)	0.017	(0.011)	6,006
Days employed	4.0	(2.7)	4.6*	(2.4)	6,006
Days on benefits	-3.0	(3.2)	-3.0	(3.0)	6,006
High earnings	0.029	(0.020)	0.026	(0.019)	1,840
Earnings	378.6*	(228.6)	393.2*	(222.7)	1,840
NDLP					
Ever employed	0.003	(0.014)	-0.006	(0.013)	5,052
Days employed	-0.1	(4.0)	-2.2	(3.5)	5,052
Days on benefits	-8.2**	(4.0)	-5.1	(3.7)	5,052
High earnings	0.054**	(0.022)	0.039*	(0.021)	1,745
Earnings	885.2***	(230.3)	730.2***	(225.5)	1,745

Note: adjusted for the observables constructed from administrative data for the full sample.
Robust standard errors for ever employed and for high earnings; *** significant at 1%, ** at 5%, * at 10%.

Testing for survey and item non-response using administrative outcomes

The raw and adjusted experimental contrasts in terms of average earnings in the first follow-up year in Table 2 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 customers could however be located, accepted to participate, or could be interviewed. Response rates remained high though: 87% among the NDLP and 75% among the ND25+ fielded samples. Of these respondents, 10% have however missing information on yearly earnings. Thus, for only one third of all ERA study participants do we observe earnings (31% in the ND25+ and 35% in the NDLP group). It thus follows that earnings information is available for one quarter of the ERA eligibles (23.6% of the ND25+ and 24.1% of the NDLP eligibles).

The survey sample was randomly chosen, and while there is good evidence (Dorsett *et al.*, 2007, Appendix G) that the respondents to the survey 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.

In this context, this section 'tests' a number of conditions (discussed in section 4.2) which help us assess whether comparing the average earnings of those with non-missing earnings information among the programme group with their counterparts among the control group would recover the ERA effect on earnings for the full group of participants (ATE_1).

We start by providing supporting evidence for the assumption that randomisation still holds within the group of respondents (the internal-validity condition). If this is the case, the experimental contrast within the subgroup of respondents will still provide an unbiased estimate of the average effect for respondents. Indeed, the rich set of observables has very little power in predicting whether a respondent is a programme or a control group member; their joint significance is rejected at any level.[12] These findings thus provide very strong evidence that the programme and control respondents subgroups are still balanced in terms of observed characteristics, which spells well for unobservables (and hence for potential outcomes) to be balanced too. In the following empirical analyses we thus consider the internal-validity condition to be met, and interpret the experimental contrast taken over the respondents as an estimate of the average effect of ERA for the respondents.

Since employment and benefit outcomes from the administrative data are available for *all* participants (respondents or non-respondents), we can use them to test whether the average impact on such outcomes for the responding participants is the same as the average impact for the full group of participants, i.e. whether the external-validity condition holds. For both customer groups, Table 3 shows that differences in impacts for all three administrative outcomes are very small and nowhere near statistical significance, both unconditionally and once controlling for observables.

Given the supporting evidence we have found for the internal-validity assumption and the fact that external-validity condition was found to hold in the administrative data, we can safely ignore non-response in calculating the average effect on earnings for participants; in other words, we can take the experimental contrast for respondents, which is readily obtained from the data, as an unbiased estimate of the ERA impact for the full group of participants.

For completeness, Table 4 presents the results of testing the stronger set of external-validity conditions. Again, we rely on the administrative data and test whether (possibly controlling for observables), the administrative outcomes of those programme (control) group members who responded to the survey are statistically different from the outcomes of those programme (control) group members for whom we do not observe the survey outcomes for whatever reason – either because they were not selected for the survey, or because they did not respond to the survey, or because they did not respond to the earnings question. This is an (unnecessarily) stricter test, as this external-validity condition for levels implies the external-validity condition for impacts, but not vice versa, and all we need is external validity in impacts. We do nonetheless report these results as they are informative in themselves.

For the ND25+ group, there is evidence that non-responding programme and especially control group members spend significantly fewer days on benefits (13.5 and 8.5) during the follow-up year than do responding programme and control group members. Controlling for our extensive set of background characteristics does not eliminate such differences; in fact, selective differences in employment probability arise, with non-responding members of both the programme and control groups exhibiting 3-4 percentage points higher employment probability than their responding counterparts with the same characteristics. Similarly, while outcome differences between responding and non-responding NDLP programme/control group members could only be weakly detected to affect the control group in terms of benefits, once we

[12] The pseudo-R squared from a Probit regression of random assignment status on the observables for the respondents' subsample is only 2% for both the ND25+ and NDLP groups, with the *p*-value of the likelihood ratio test of the null that the observables are jointly insignificant in predicting random assignment status being 0.175 for the former and 0.495 for the latter.

condition on observables, selective differences appear quite marked for both programme and control groups and in terms of most outcomes. Non-responding NDLP customers experience significantly better employment and benefit outcomes than their responding counterparts, with very similar differences within the programme and the control group.[13]

Table 3 Testing equality of impacts for responding and non-responding participants

	Ever employed		Days employed		Days on benefits	
	diff	*p*-value	*diff*	*p*-value	*diff*	*p*-value
Unconditional on *X*						
ND25+	0.022	0.218	6.3	0.131	3.4	0.457
NDLP	-0.015	0.413	-0.4	0.944	2.6	0.636
Conditional on *X*						
ND25+	0.016	0.326	5.0	0.187	4.5	0.310
NDLP	-0.009	0.614	3.0	0.515	1.7	0.749

Notes: *diff* is the difference in the average ERA impact for participants compared to the experimental contrast for responding participants; *p*-value based on bootstrapped significance (500 reps); *** significant at 1%, ** at 5%, * at 10%.
Sample sizes: 5,724 for ND25 Plus and 4,770 for NDLP.

Table 4 Testing equality of outcomes between non-responding and responding programme (1) and control (0) group members

	$P_{S=0\|R=1}$	$P_{S=0\|R=0}$	Unconditional on *X*		Conditional on *X*	
			diff(1)	*diff*(0)	*diff*(1)	*diff*(0)
ND25+						
Ever employed			0.028	-0.004	0.045**	0.038**
Days employed	0.678	0.680	3.138	-6.125	3.958	1.932
Days on benefits			-8.551*	-13.527***	-6.386	-19.953***
NDLP						
Ever employed			0.009	0.033	0.048**	0.050**
Days employed	0.626	0.642	4.597	5.053	15.271***	7.900
Days on benefits			-6.380	-10.207*	-17.154***	-17.819***

Notes: *p*-values based on heteroskedasticity-robust standard error; *** significant at 1%, ** at 5%, * at 10%.
$P_{S=0|R=1}$ is the proportion of non-respondents among the programme group, $P_{S=0|R=0}$ among the control group.
diff(.) is the difference in average outcomes of non-respondents compared to respondents within the programme group (*diff*(1)) or within the control group (*diff*(0)).
Sample sizes: 5,724 for ND25 Plus and 4,770 for NDLP.

[13] This pattern is consistent with selection into survey/item response within experimental group depending partly on unobservables; in such a situation, conditioning on observed characteristics may accentuate, rather than eliminate, outcome differences between responding and non-responding individuals within the two experimental groups.

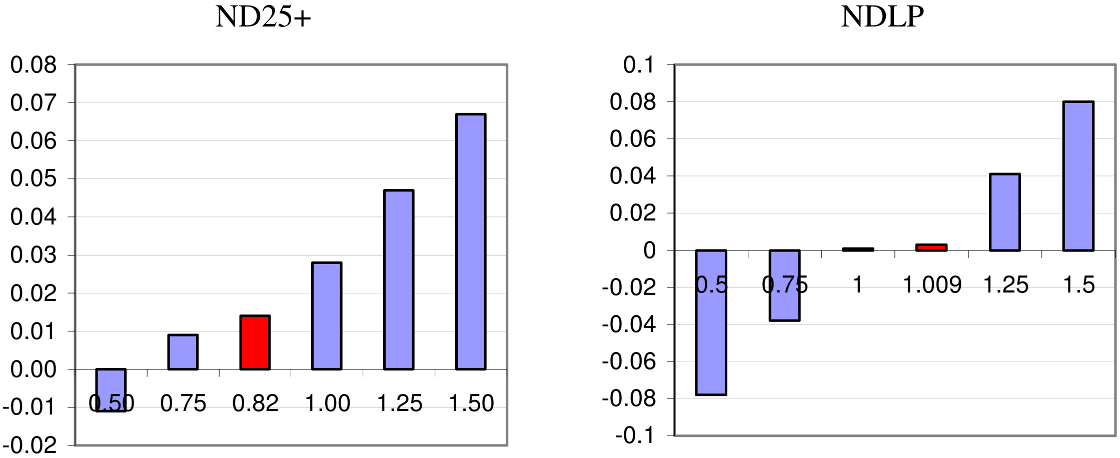
5.2 Bounds

By construction, where the share of non-participants is sizeable and the experimental impact negligible, the bounds are very wide. This is indeed the case for the two customer groups overall, for whom the zero impact on employment probability is bounded between -5 and 18 percentage points (ND25+) and -16 and 15 percentage points (NDLP). Bounds for survey outcomes were so wide as to be totally uninformative.

The sensitivity analysis for administrative outcomes can however be at times quite informative. Indeed, for the ND25+ group overall, the average effect remains positive and small under the most plausible assumptions, in contrast to the NDLP group overall, for whom the *ATE* could be negative, positive or zero depending on the type of selection mechanism underlying participation in the ERA study (Figure 3).

Another interesting finding from the sensitivity analysis is that the type of assumption (i.e. value of θ) required for the experimental impact to be an unbiased estimate of the average effect for the full eligible population is different for the two customer groups. In particular, in order to ignore non-participation in the NDLP group, one would need to assume a more favourable selection into the ERA study than in the case of ND25+.[14]

Figure 2: Sensitivity analysis: ATE_θ for ever employed, θ from 0.5 to 1.5



In red: experimental impact estimate and corresponding θ.

[14] For the ND25+ customer groups, to take the experimental impact as representative of the impact on the eligibles, one would need to assume that non-participants among the ND25+ eligibles would have experienced a 20% *lower* employment probability had they been offered ERA services than what actual participants receiving ERA are observed to experience. For the NDLP customer group by contrast, the experimental estimate would recover the average effect under the assumption that the non-participants did not select into the ERA study based on treatment outcomes.

5.3 Results for administrative outcomes

We now turn to our impact estimates under the assumption that we observe all outcome-relevant characteristics that drive selection into the ERA study.

Table 5 presents the matching results for ND25+ and NDLP overall. An overarching comment which applies to all the results based on selection on observables is that, provided the identifying assumption is met, the estimates can be viewed as very reliable, since the matching exercise has performed extremely well in balancing the observable characteristics (see Appendix Table 1).

Table 5 Matching estimates on administrative outcomes

	p	ATE_1	ATE_0	ATE	$ATE_1 \neq ATE$
ND25+					
Ever employed	0.230	0.017	0.056***	0.026**	***
Days employed		4.560**	9.984***	5.805***	*
Days on benefits		-2.966	8.862**	-0.250	***
NDLP					
Ever employed	0.304	-0.006	0.015	0.000	
Days employed		-2.208	-1.957	-2.132	
Days on benefits		-5.078	8.881**	-0.831	***

Notes: 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%.
Sample sizes: 4,831 for ND25 Plus and 4,768 for NDLP.

Starting with the results for the ND25+ customer group, once we correct for differences in observed characteristics between participants and non-participants in estimating the effect of ERA on non-participants and on the full eligible population, we find that non-participants would have experienced a *worse* ERA impact on benefit dependency than participants. In particular, had they been offered ERA services, the group of non-participants would have spent almost 9 days longer on benefits (significant at the 5% level) in the follow-up year than if they had not been offered ERA. By contrast, participants are found to spend a statistically insignificant 3 days less on benefits thanks to ERA. The ERA impact on eligibles at around 0 is statistically different from the one on the participants.

In terms of employment outcomes, by contrast, ERA impacts for the non-participants in the ND25+ group would have been *consistently better* than those experimentally estimated for the subgroup of participants. Specifically, non-participants overall would have enjoyed a highly significant, 5.6 percentage point increase in their follow-up employment probability due to ERA, compared to an insignificant 1.7 increase for participants. The *ATE* for the full group of eligibles would correspondingly have been a significant increase of almost 3 percentage points. Similarly, non-participants would have enjoyed more than double an increase in days employed (10) than do participants (4.6), resulting in an overall average impact of 6 days, all effects being highly statistically significant.

These findings point to the possibility that the ND25+ non-participants might in fact be easier to help back into the labour market than the average ND25+ entrant and that for these more labour-market detached ND25+ entrants some extra help in the form of advice and financial incentives might be particularly helpful in improving their labour market situation.

For the ND25+ customer group, the experimental impact estimate of ERA thus *underestimates* the contribution that the programme can give to all eligibles in terms of improving their employment outcomes, though, as we have seen, the opposite is true when considering benefit dependency. The most likely implication of the finding that ERA would have increased both employment durations and time on benefits for the ND25+ non-participants is that ERA would have reduced the time these customers spend in "uncompensated" non-employment, i.e. outside the labour market as well as the government support system.[15]

Moving now to the NDLP customer group, the findings are somehow less compelling, as it is more difficult to reach statistical significance. The employment effect in terms of either employment probability or employment duration would have been the same – and statistically indistinguishable from zero – for the non-participants as for the experimental group. As was mostly the case for the ND25+ customer group, NDLP non-participants would have experienced a worse ERA impact on benefit dependency than the experimental group; while participants remained unaffected, non-participants would have seen their time on benefits increase by a significant duration of 9 days. Overall, for the NDLP customer group, the experimental estimate of no ERA impact on employment outcomes is thus largely representative of the average effect for all eligibles; the experimental finding however overestimates the impact ERA would have had on all eligibles in terms of reducing their benefit dependency.

Sensitivity analysis

This sensitivity analysis relaxes the selection-on-observables assumption (A1) by allowing participants and non-participants with the same observed characteristics to still differ in terms of some unobserved dimension – summarised by θ – that affects their treatment outcome. For favourable outcomes such as employment probability or days employed, θ>1 implies positive selection into the non-participants sample (i.e. non-participants would have enjoyed better employment outcomes under ERA than observably similar participants) and θ<1 negative selection. For unfavourable outcomes such as days on benefits, the opposite holds. For θ=1, we obviously obtain the matching estimates discussed above.

In line with the bounds analysis in Section 5.2, the sensitivity analysis in Table 6 is quite informative for the ND25+ group and paints a rather favourable picture for the impact that ERA would have had on all eligibles. In particular, the employment effect of ERA for the eligibles would have been positive, albeit rather small in size (except than under the most extreme selection scenario of θ much larger than one). Similarly, the impact on benefit outcomes for the eligibles would appear to be quite favourable under most selection scenarios.

[15] Some individuals could still be in work even in the absence of employment records in the available administrative data (the WPLS). Note in any case that time in employment and time on benefits are not mutually exclusive (individuals can be employed at the same time as claiming a benefit such as income support); this is particularly the case with the WPLS, which contains no information on the amount of hours worked.

By contrast, in terms of both employment and benefit outcomes, relaxing assumption (A1) under a number of plausible values for θ does not allow one to say much for the NDLP group, for whom the average impact for all eligibles would range from substantial and negative to substantial and positive.

As to the value of θ required for the experimental impact on employment outcomes to be an unbiased estimate of the average effect for the full eligible population, for the ND25+ group we again find that it is always well below 1, and mostly below the corresponding value for NDLP customers. Thus, in order to take the experimental impact as representative of the impact on the eligibles, one would need to assume that the non-participants among the ND25+ eligibles would have experienced much lower employment probabilities and fewer days in employment had they been offered ERA services than what actual participants receiving ERA are observed to experience. In terms of benefit outcomes, though, the θ corresponding to the experimental estimate would imply a favourable selection into the non-participation sample. Given this marked divergence in the direction of selection required for the experimental estimate to recover the average effect for employment as opposed to benefit outcomes, such a set of assumptions would seem questionable.

By contrast, for the NDLP group there seems to be more consistency in the requirements imposed on θ for the two types of outcomes, as in order to ignore non-participation one needs to assume no selection into the ERA study in terms of employment outcomes, and a slightly unfavourable selection in terms of benefit outcomes (in particular, had they received ERA, non-participants would have spent on benefits 93% of the time that participants spend on benefits).

Table 6 Sensitivity analysis: ATE_θ, θ from 0.5 to 1.5

ND25+

Ever employed		Days employed		Days on benefits	
θ	ATE_θ	θ	ATE_θ	θ	ATE_θ
0.50	-0.011	0.50	-0.783	0.50	-30.424
0.75	0.007	0.75	2.511	0.75	-15.337
0.88	**0.017**	**0.91**	**4.560**	**0.96**	**-2.966**
1.00	0.026	1.00	5.805	1.00	-0.250
1.25	0.044	1.25	9.099	1.25	14.836
1.50	0.062	1.50	12.393	1.50	.

NDLP

Ever employed		Days employed		Days on benefits	
θ	ATE_θ	θ	ATE_θ	θ	ATE_θ
0.50	-0.081	0.50	-20.027	0.50	-32.977
0.75	-0.040	0.75	-11.079	0.75	-16.904
0.96	**-0.006**	**0.99**	**-2.208**	**0.93**	**-5.078**
1.00	0.000	1.00	-2.132	1.00	-0.831
1.25	0.041	1.25	6.816	1.25	15.242
1.50	0.082	1.50	15.763	1.50	31.315

In bold: experimental impact estimate and corresponding θ.
Missing ATE_θ denotes an inadmissible θ value.
Sample sizes: 4,831 for ND25 Plus and 4,768 for NDLP.

5.4 Results for survey outcomes

Table 7 presents our weighting and matching results for survey-based earnings outcomes, where both methods account for non-response.

Table 7 Weighting and matching estimates of the average ERA impact on earnings for all eligibles accounting for non-response

	ND25+		NDLP	
	Weighting	*Matching*	*Weighting*	*Matching*
ATE	559.9**	580.2***	644.7**	718.2***
$E(Y_1)$	2772.3	2779.6	3557.9	3509.2
$E(Y_0)$	2212.3	2199.4	2913.2	2791.1
Δ	393.2*		730.2***	
N	7,399		6,809	

Notes:
ATE is the average ERA impact for all eligibles;
$E(Y_1)$ are average earnings of all eligibles under ERA treatment; $E(Y_0)$ are average earnings for all eligibles without ERA treatment;
Δ is the experimental estimate ignoring potential non-response bias;
Matching estimator: kernel matching with epanechnikov kernel (bandwidth of 0.06), common support imposed separately for each term.
Statistical significance based on bootstrapped bias-corrected confidence intervals (1000 replications for the weighting estimator, 500 for the matching estimator): *** significant at 1%, ** at 5%, * at 10%.
In this table, Δ is never statistically significantly different from the *ATE* according to bootstrap-based statistical significance of the difference.

Because of (survey and/or item) non-response, in the following discussion the estimated *ATE* for the full eligible population, $ATE \equiv E(Y_1) - E(Y_0)$, has to be compared to the experimental contrast calculated on the responding participants, Δ.

First of all, the evidence emerging from both the weighting and matching estimators tells a pretty consistent story, despite the former estimator's more pronounced sensitivity and difficulty in achieving statistical significance. The point estimates are also quite close.

Although more formal bootstrap-based tests of the difference between the experimental contrast on respondents and the estimated *ATE* fail to uncover any statistically significant difference, the evidence in terms of both point estimates and their statistical significance tells a consistent story: the ERA impact on earnings estimated on the responding experimental group underestimates the average impact of the programme on the full eligible population for the ND25+ group while being a representative estimate of the full impact for the NDLP group. Specifically, once non-response and non-participation are taken into account, point estimates increase for ND25+ and remain largely stable for NDLP customers.

Focusing on the matching estimates, the experimental estimator for respondents of an increase in earnings of £393 (significant only at 10%) is contrasted to a highly significant estimated increase for all eligibles of £580 for the ND25+ group.

As to the NDLP group, as already mentioned both the point estimates and their significance remain largely stable. The highly significant overall experimental estimate of £730 is in line with a similarly significant estimate for all eligibles of £718.

In terms of the underlying matching quality, which can only be assessed for the matching (as opposed to the weighing) estimator, the indicators are extremely encouraging (see Appendix Table 1).

We have also derived and estimated the matching estimates when non-response can be ignored. For convenience of comparison, in Table 8 we report again the matching estimates just discussed which allow for non-response.

Table 8 Matching estimates of the average ERA impact on earnings for all eligibles

		ND25+	NDLP
Δ		393.2*	730.2***
ATE	allowing for non-response, separate CS	580.2***	718.2***
	ignoring non-response, separate CS	442.8*	662.8***
	ignoring non-response, joint CS	443.5*	660.4**
% lost to joint CS		0.8	1.0
N		7,399	6,809

Notes: Statistical significance based on bootstrapped bias-corrected confidence intervals (500 repetitions): *** significant at 1%, ** at 5%, * at 10%.
In this table, *Δ* (the experimental estimate ignoring potential non-response bias) is never statistically significantly different from the *ATE* according to bootstrap-based statistical significance of the difference.
Kernel matching with epanechnikov kernel (bandwidth of 0.06).
Separate CS: common support imposed on the non-participants separately for each term; Joint CS: estimates pertain to those non-participants satisfying both support conditions.
When ignoring non-response, *Δ* is assumed to be equal to ATE_1.

In the main, the results for the *ATE* ignoring non-response are much closer to the experimental estimates than those allowing for it (our preferred estimates).

For the ND25+ group, taking account of non-participation but ignoring non-response still raises the positive impact estimates on earnings estimated on the responding experimental sample, but does so by a smaller magnitude than when allowing for non-response (though this only concerns the point estimates; neither of the estimates of the *ATE* are statistically significantly different from the experimental one at conventional levels).

For the NDLP group, the estimates ignoring non-response line up very closely to the experimental ones. Compared to those allowing for non-response, there is a slightly larger though still minor fall in the point estimate.

In the case where non-response is not taken into account, the two different ways of imposing the common support were found to produce strikingly close point estimates and statistical significance, despite the at times large differences in the proportions of the sample being excluded from the analysis.

5.5 Analysis of take-up

Although an analysis of the effect of ERA *eligibility* would need to include the non-participants irrespective of their potential take-up of the programme, it is still very interesting to know the type of involvement they would have had with ERA – and more generally with Jobcentre Plus – had they participated in the evaluation study, either as part of the programme group or of the control group.

Table 9 presents the results of these analyses in terms of a number of measures of take-up of services and of contact with Jobcentre Plus staff. Specifically, we consider

- measures of presence, type and intensity of contact with Jobcentre Plus staff (any contact, customer has initiated face-to-face visits, very intense contact in the form of 10 or more face-to-face meetings);
- measures of help or advice received from Jobcentre Plus staff when the customer was not working (staff offered any help/advice, performed a better-off calculation, suggested customer attend a Jobclub/Programme Centre, arranged an education or training course, offered advice without being requested);
- measures of the customer's assessment of the advice received; and
- for the programme group analysis only, measures directly linked to knowledge of ERA features (whether the customer has heard of the employment and of the training bonuses).

Recall from Section 4.4 that all results hinge on the assumption that there is no selection into the ERA study based on *unobserved* characteristics that also affect take-up of ERA services or involvement with Jobcentre Plus if participating in the study. Subject to this proviso, the findings provide interesting evidence on the two sets of questions we consider.

First we focus on the take-up that the non-participants would have exhibited had they been assigned to the programme group. Are the non-participants individuals who even if offered ERA services would not take them up? And could this be the underlying reason for Jobcentre Plus caseworkers not offering them the chance to participate in the randomisation in the first place, or, for those who were offered such a chance, the reason driving their own refusal to participate in the demonstration? If this is the case, one might argue that even if ERA became an official policy, they would not be interested in effectively taking up the support and incentives it offers.[16]

For the ND25+ group, there are statistically significant differences between the non-participants and the programme group in two measures of involvement with Jobcentre Plus staff and in terms of awareness of the ERA bonuses, but such differences are not striking. Specifically, while 85% of the programme group has received help or advice from Jobcentre Plus staff while not working, our model predicts that 82.5% of the non-participants would have received such help had they been assigned to the programme group. Similarly, the non-participants would have a 2 percentage point lower likelihood than the programme group of being offered help by staff without being requested. Non-participants would also have been less aware of the bonuses than the actual programme group is (72.9% rather than 75.4% for the employment bonus and 40.1% rather than 43% for the training bonus).

[16] Again note that if some eligibles are not fully informed about ERA or do not otherwise avail themselves of its services, they will dilute the effect of ERA eligibility on the eligibles.

Overall, had they been randomised into the programme, the ND25+ non-participants would have been quite heavily involved with ERA and Jobcentre Plus. And although we find that they would have been statistically significantly less aware of ERA features and would have experienced slightly less contact than the actual programme group, such differences are arguably small from a substantive point of view.

The conjecture that if the programme became official, non-participants would be mostly uninterested in taking up its support and incentives finds no strong support for the NDLP group either. In fact, had they become eligible to ERA services and incentives, the non-participants would have been over 3 percentage points more likely than the programme group to be involved in training and education activities arranged by Jobcentre Plus, as well as more likely to be directed to a Jobclub or Programme Centre. The two groups are not found to differ significantly in any other measure of awareness and involvement, with the notable exception of the likelihood to receive help or advice from Jobcentre Plus when not working. As was the case for ND25+ customers, it is again the programme group who is 2.4 percentage points more likely to receive such help than the non-participants. As many as 75% of the latter are however still predicted to receive such support when out of work.

The second question we have looked at concerns the kind of involvement that non-participants would have had with Jobcentre Plus had they participated in the ERA study and been assigned to the control group. Among the reasons that the qualitative research has highlighted for ND25+ customers to formally refuse to participate, there was a feeling of being close to get a job in the near future and not wanting to stay in touch with Jobcentre Plus, or a strong antipathy to government and systems of support and governance. The question thus arises of whether the ND25+ non-participant group is made up of individuals who would shun involvement with Jobcentre Plus at all cost. This supposition is not borne out in the data: had they been assigned to the control group, the involvement that the ND25+ non-participants would have had with Jobcentre Plus would not have been statistically different from the one displayed by the actual control group in any of the dimensions considered.

As opposed to ND25+ customers, NDLP customers were easy to recruit to the ERA study once having been offered the chance to participate in it. In fact, most (87%) of the non-participants amongst the NDLP group were diverted customers. One might thus conjecture that had they been offered the chance to participate, the NDLP non-participants would in fact have been quite involved with Jobcentre Plus even if assigned to the control group. According to the results in Table 14, this seems to be the case. Indeed, it is estimated that compared to the control group, NDLP non-participants would have had the same type and intensity of involvement with Jobcentre Plus staff, while being 4 percentage points more likely to rate their advice as very helpful.

Overall, the share of the eligible population that has been excluded (i.e. the diverted customers) or has formally refused to take part in the ERA study displays observed characteristics that make them quite likely to be involved with Jobcentre Plus generally, both with and without ERA.

Table 9 Take-up and involvement with Jobcentre Plus predicted for the non-participants both under ERA and without ERA

	ND25+				NDLP			
	ERA outcome		Non-ERA outcome		ERA outcome		Non-ERA outcome	
	Programme group	Non-participants	Control group	Non-participants	Programme group	Non-participants	Control group	Non-participants
Has had contact with JCP staff	84.8	83.7	78.2	78.2	85.3	86.4	71.9	74.6
Has ever initiated face to face visits	55.4	54.5	50.4	49.7	62.0	61.3	55.5	56.5
Had face to face contact with JCP staff ≥10 times	43.0	43.5	41.0	42.1	14.2	15.5	9.8	9.1
Received help/advice from JCP staff when not working	85.0	82.5***	84.9	85.8	77.2	74.8*	73.7	71.2
JCP staff did better-off calculation when not working	41.6	41.0	38.6	39.4	63.8	63.2	64.2	64.7
JCP staff suggested attend a Jobclub/Programme Centre	32.7	34.3	32.9	35.2	5.3	6.6*	6.2	7.1
JCP staff arranged education/training	30.4	31.3	31.5	31.4	14.6	17.8***	12.3	14.0
JCP staff offered help/advice without being requested	18.4	16.2**	7.8	7.9	26.3	27.6	9.4	9.9
Found advice from JCP staff overall very helpful	33.1	31.2	23.6	22.8	42.6	43.2	31.1	35.1**
Found advice from JCP staff overall not at all helpful	4.7	5.0	5.8	5.2	3.4	2.5	4.1	3.7
Has heard of employment bonus	75.4	72.9**	–	–	72.8	71.0	–	–
Has heard of training bonus	43.0	40.1**	–	–	50.8	52.9	–	–
N	1,014	1,675	996	1,675	1,014	2,039	994	2,039

Note: Programme group and control group columns report the observed rates; non-participants columns report the predicted rate for participants under ERA and without ERA.

Statistical significance of the difference in rates between non-participants and programme (or control group) is based on bootstrapped bias-corrected confidence intervals (500 replications): *** significant at 1%, ** at 5%, * at 10%.

Note: the sample sizes shown for the programme and control groups refer to those with non-missing information on “has had contact with JCP staff”.

6. Summary and conclusions

In our descriptive examination of the non-participation problem (Goodman and Sianesi, 2007), we speculated that it would be hard for the non-participants to give rise to an estimate for all eligibles that tells a different 'story' from the one arising from the experimental estimate (where the 'story' could be one among: ERA is harmful, it has basically no effect, it has a 'relatively small' effect or it has a 'relatively large' effect – whatever one may mean with 'relatively large' or 'relatively small').

Indeed, we have found that the story does not radically change – in statistical as well as qualitative terms.

For the NDLP group, in fact, the story remains unchanged. Specifically, the bottom-line in the first-year follow-up is that ERA has had no effect on employment and benefit outcomes, while it has significantly and substantially increased their yearly earnings. We find that what the programme has done for the participants, it would have done also for the non-participants and hence for the whole eligible population.

For the ND25+ group, by contrast, the story changes somewhat in the direction of a slightly more effective ERA treatment: positive impacts surface, become larger in size or stronger in statistical significance. Specifically, while no significant impact has been detected on the employment probability of participants, statistically significant effects emerge for all eligibles, the treatment effect on employment durations for all eligibles is 29% higher than the one obtained using the experimental sample, while the ERA effect on earnings would have been 48% higher for all eligibles than it is for the study participants.

We thus *do* find evidence of randomization bias (or of some loss in external validity) in the data for the ND25+ group. When we adequately account for non-participation, we find that the employment and earnings impact estimates that rely on experimental data alone *underestimate* the true impact of ERA on all ND25+ entrants. Of course, there is always the issue of how different the estimates for the eligibles and for the experimental sample need to be for us to view the issue as a particularly important one. Randomised experiments are however conceptually designed to provide with accuracy the 'true' answer to the evaluation question. Finding an effect for the eligibles which is 30 or 50% larger (or 15% smaller) than the experimental estimate can be viewed as a finding of substance.

In this paper we have not only extensively assessed the external validity of the short-term ERA findings, but we have set the foundation work and developed a sound and thorough methodological framework for the analysis of non-participation. Given that in many evaluation settings the problem of non-participation is an empirically relevant one (see e.g. Kamionka and Lacroix, 2005), the framework we have developed can be applied to assessing this issue in any study which can exploit the three critical features of 1) being interested in assessing the impact of offering a new treatment, (2) eligible for this offer under an official policy would be a well-defined population, (3) for whom background (and ideally, outcome) information is recorded in the available data.

References

Blundell, R., Dearden, L. and Sianesi, B. (2005), “Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey”, *Journal of the Royal Statistical Society: Series A*, 168, 473-512.

Dolton, P., Smith, J. and Azevedo, J.P. (2008), “The Impact of the UK New Deal for Lone Parents on Benefit Receipt”, mimeo, March.

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, February.

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

Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998), “Characterising Selection Bias Using Experimental Data.” *Econometrica* 66, 1017-1098.

Heckman, J., LaLonde, R. and Smith, J. (1999). “The Economics and Econometrics of Active Labor Market Programs”, in Orley Ashenfelter and David Card (eds.), *Handbook of Labor Economics, Volume 3A*. 1865-2097.

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

Horowitz, J.L. and Manski, C.F. (2000), “Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data”, *Journal of the American Statistical Association*, 95, 77-84.

Goodman, A. and Sianesi, B. (2007), “Non-participation in the Employment Retention and Advancement Study: A Quantitative Descriptive Analysis”, Department for Work and Pensions Technical Working Paper No.39.

Hall, N., Hoggart, L., Marsh, A., Phillips, J., Ray, K. and Vegeris, S. (2005), “The Employment Retention and Advancement Scheme: Report on the Implementation of ERA during the First Months. Summary and Conclusions”, Department for Work and Pensions Research Report No 265, August.

Kamionka, T. and Lacroix, G. (2005), “Assessing the External Validity of an Experimental Wage Subsidy”, IZA Discussion Paper No. 1508.

Imbens, G.W. and Manski, C.F. (2004), “Confidence Intervals for Partially Identified Parameters”, *Econometrica*, 72, 1845-1857.

Jawitz, J. (2004), “Moments of truncated continuous univariate distributions”, *Advances in Water Resources*, 27, 269-281.

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.

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, March.

Appendices

Appendix 1a Reweighting: Ignoring survey and item non-response

As to the first term of equation (2): $E_X[E(Y \mid R{=}1, X)]$

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

$$\equiv \int E(Y \mid R=1,x)\omega_1(x) f(x \mid R=1)dx = \int E(\omega_1(x)Y \mid R=1,x) f(x \mid R=1)dx$$

Using first Bayes' rule and the law of iterated expectations and noting that $P(R{=}1 \mid Q{=}0) = 0$, we get:

$$\omega_1(x) \equiv \frac{f(x)}{f(x \mid R=1)} = \frac{P(R=1)}{P(R=1 \mid x)} = \frac{P(Q=1)P(R=1 \mid Q=1)}{P(Q=1 \mid x)P(R=1 \mid Q=1,x)} \equiv \frac{(1-p)p_R}{(1-p(x))p_R(x)} \overset{RA}{=} \frac{1-p}{1-p(x)}$$

where

- $p_R \equiv P(R{=}1 \mid Q{=}1)$ is the probability of being randomly assigned to the programme group conditional on participating in the ERA study (Q=1), and
- $p_R(x) \equiv P(R{=}1 \mid Q{=}1, x)$ is the corresponding conditional probability.

Under randomisation, $p_R{=}p_R(x)$.

The first term: $E_X[E(Y \mid R{=}1, X)] = E_X[E(\omega_1(x) \cdot Y \mid R{=}1, X) \mid R{=}1]$ can hence be estimated by reweighing the outcomes of the programme group by $\omega_1(x)$ and averaging them:

$$\frac{1}{\#(R=1)} \sum_{i \in \{R=1\}} \omega_1(x_j)y_j = \frac{(1-p)p_R}{\#(R=1)} \sum_{i \in \{R=1\}} \frac{y_i}{\left(1-p(x_i)\right) p_R(x_i)}$$

Under randomisation, $p_R{=}p_R(x)$ so that:

$$\frac{1}{\#(R=1)} \sum_{i \in \{R=1\}} \omega_1(x_j)y_j = \frac{1-p}{\#(R=1)} \sum_{i \in \{R=1\}} \frac{y_i}{1-p(x_i)}$$

Similarly, the second term of (2) can be rewritten as: $E_X[E(Y \mid R{=}0, X)]$

$$= \int E(\omega_0(x)Y \mid R=0,x) f(x \mid R=0)dx = E_X[E(\omega_0(x) \cdot Y \mid R{=}0, X) \mid R{=}0]$$

where (noting that due to randomisation, the weight ω is the same):

$$\omega_0(x) \equiv \frac{f(x)}{f(x \mid R=0)} = \frac{P(Q=1)P(R=0 \mid Q=1)}{P(Q=1 \mid x)P(R=0 \mid Q=1,x)} \equiv \frac{(1-p)(1-p_R)}{(1-p(x))(1-p_R(x))} \overset{RA}{=} \frac{1-p}{1-p(x)} = \omega_1(x)$$

which can be estimated by reweighing the outcomes of the control group and averaging them:

$$\frac{1}{\#(R=0)} \sum_{i \in \{R=0\}} \omega_0(x_j)y_j = \frac{(1-p)(1-p_R)}{\#(R=0)} \sum_{i \in \{R=0\}} \frac{y_i}{\left(1-p(x_i)\right)\left(1-p_R(x_i)\right)}$$

Under randomisation, $p_R{=}p_R(x)$, hence:

$$\frac{1}{\#(R=0)} \sum_{i \in \{R=0\}} \omega_0(x_j)y_j = \frac{1-p}{\#(R=0)} \sum_{i \in \{R=0\}} \frac{y_i}{1-p(x_i)}$$

We can thus estimate the ATE in (2) by reweighing and averaging the outcomes of the full group of participants (Q=1):

$$A\hat{T}E = \left[\frac{(1-p)p_R}{\#(R=1)} \sum_{i \in \{R=1\}} \frac{y_i}{\left(1-p(x_i)\right) p_R(x_i)}\right] - \left[\frac{(1-p)(1-p_R)}{\#(R=0)} \sum_{i \in \{R=0\}} \frac{y_i}{\left(1-p(x_i)\right)\left(1-p_R(x_i)\right)}\right]$$

Taking full advantage of the randomisation and noting that #(R=1) = #(R=0) due to the 50-50 random allocation:

$$A\hat{T}E = \frac{1-p}{\#(R=1)} \sum_{i\in\{Q=1\}} \frac{R_i y_i - (1-R_i) y_i}{1-p(x_i)}$$

However, although randomisation worked very well, especially when conditioning on X there might be residual imbalances due to pure chance. Even more crucially, this analysis can only be performed for the survey subgroup, and indeed for that subgroup of survey respondents who responded to the earnings question. For this reason, in implementing this estimator we allow for the more general case.

Appendix 1b Reweighting: Allowing for survey and item non-response

Here we allow for selective non-response, provided such selection into the responding sample happens only in terms of observable characteristics. We thus relax the assumptions from Section 4.2 by invoking them conditional on X:

(E-V'.X) (a) $E(Y_1 \mid R{=}1, S{=}1, X) = E(Y_1 \mid R{=}1, S{=}0, X)$ and

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

$$\begin{aligned}
ATE &\equiv E(Y_1 - Y_0) = E_X[\, E(Y_1 - Y_0 \mid X)] = \text{(A2)} = E_X[\, E(Y_1 - Y_0 \mid Q{=}1, X)] \\
&= \text{(RA)} = E_X[E(Y_1 \mid R{=}1, X)] - E_X[E(Y_0 \mid R{=}0, X)] \\
&= \text{(E-V'.}X\text{)} = E_X[E(Y_1 \mid R{=}1, S{=}1, X)] - E_X[E(Y_0 \mid R{=}0, S{=}1, X)] \\
&= E_X[E(Y \mid R{=}1, S{=}1, X)] - E_X[E(Y \mid R{=}0, S{=}1, X)] \qquad (3)
\end{aligned}$$

ATE is thus identified in the data and can be empirically estimated as follows.
As to the first term of expression (3):
$E_X[E(Y \mid R{=}1, S{=}1, X)]$

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

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

with

$$\omega_1(x) \equiv \frac{f(x)}{f(x \mid R=1, S=1)} = \frac{P(R=1, S=1)}{P(R=1, S=1 \mid x)} = \frac{P(Q=1)P(R=1, S=1 \mid Q=1)}{P(Q=1 \mid x)P(R=1, S=1 \mid Q=1, x)} \equiv \frac{(1-p)p_{RS1}}{(1-p(x))p_{RS1}(x)}$$

where $p_{RS1} \equiv P(R{=}1, S{=}1 \mid Q{=}1)$ is the probability among participants of being randomly assigned to the programme group *and* of responding to the earnings question, and $p_{RS1}(x) \equiv P(R{=}1, S{=}1 \mid Q{=}1, x)$ is the corresponding conditional probability.

$E_X[E(Y \mid R{=}1, S{=}1, X)] = E_X[E(\omega_1(x)\cdot Y \mid R{=}1, S{=}1, X) \mid R{=}1, S{=}1]$
can hence be estimated by reweighing by $\omega_1(x)$ the outcomes of the programme group members who responded to the earnings question and averaging them over this subgroup:

$$\frac{1}{\#(R=1, S=1)} \sum_{i\in\{R=1,S=1\}} \omega_1(x_j) y_j = \frac{(1-p)p_{RS1}}{\#(R=1, S=1)} \sum_{i\in\{R=1\}} \frac{y_i}{\left(1-p(x_i)\right) p_{RS1}(x_i)}$$

Similarly, the second term of expression (3) can be rewritten as:
$E_X[E(Y \mid R{=}0, S{=}1, X)] = E_X[E(\omega_0(x)\cdot Y \mid R{=}0, S{=}1, X) \mid R{=}0, S{=}1]$
with

$$\omega_0(x) \equiv \frac{f(x)}{f(x \mid R=0, S=1)} = \frac{P(Q=1)P(R=0, S=1 \mid Q=1)}{P(Q=1 \mid x)P(R=0, S=1 \mid Q=1, x)} \equiv \frac{(1-p)p_{RS0}}{(1-p(x))p_{RS0}(x)}$$

where p_{RS0} is the probability among participants of being randomly assigned to the control group *and* of responding to the earnings question. (Note that p_{RS0} is not equal to $1-p_{RS1}$).

This term can be estimated by reweighing the outcomes of the control group who responded to the earnings question and averaging them over this subgroup:

$$\frac{1}{\#(R=0,S=1)}\sum_{i\in\{R=0,S=1\}}\omega_0(x_j)y_j=\frac{(1-p)p_{RS0}}{\#(R=0,S=1)}\sum_{i\in\{R=0,S=1\}}\frac{y_i}{\left(1-p(x_i)\right)p_{RS0}(x_i)}$$

Hence we can estimate the *ATE* in equation (3) by reweighing and averaging the outcomes of all those participants who responded to the survey (Q=1 and S=1):

$$A\hat{T}E=\left[\frac{1}{\#(R=1,S=1)}\sum_{i\in\{R=1,S=1\}}\frac{(1-p)p_{RS1}}{\left(1-p(x_i)\right)p_{RS1}(x_i)}y_i\right]-\left[\frac{1}{\#(R=0,S=1)}\sum_{i\in\{R=0,S=1\}}\frac{(1-p)p_{RS0}}{\left(1-p(x_i)\right)p_{RS0}(x_i)}y_i\right]$$

Appendix 2a Matching Protocol: Ignoring survey and item non-response

Assume that, once conditioning on observables X, ERA study participants and non-participants on average experience the same treatment and no-treatment outcomes:

(A2) (a) $E(Y_1 \mid Q=1, X) = E(Y_1 \mid Q=0, X)$
(b) $E(Y_0 \mid Q=1, X) = E(Y_0 \mid Q=0, X)$

Ignoring non-response allows one to treat the responding participants as representative of the full group of participants. We make the following assumptions as to non-response:

(E-V) $E(Y_1 - Y_0 \mid Q=1) = E(Y_1 - Y_0 \mid Q=1, S=1)$
(E-V') (a) $E(Y_1 \mid R=1) = E(Y_1 \mid R=1, S=1) = E(Y_1 \mid R=1, S=0)$
(b) $E(Y_0 \mid R=0) = E(Y_0 \mid R=0, S=1) = E(Y_0 \mid R=0, S=0)$
(I-V) $E(Y_1 \mid S=1, R=1) = E(Y_1 \mid S=1, R=0) = E(Y_1 \mid S=1)$
$E(Y_0 \mid S=1, R=1) = E(Y_0 \mid S=1, R=0) = E(Y_0 \mid S=1)$

(A) Obtaining the *ATE* after having first obtained the *ATE*₀

Starting from $ATE = (1-p)\cdot ATE_1 + p\cdot \boldsymbol{ATE_0}$ (1b)

- p is observed
- $ATE_1 \equiv E(Y_1 - Y_0 \mid Q=1)$ = (E-V) = $E(Y_1 - Y_0 \mid Q=1, S=1)$ = (I-V)
= $E(Y \mid S=1, R=1) - E(Y \mid S=1, R=0)$.
Note that we control for X in deriving this estimate.

To recover the ATE_0, we need to estimate $E(Y_1 \mid Q=0)$ and, given the absence of survey outcomes for non-participants, $E(Y_0 \mid Q=0)$ as well.

- $E(Y_1 \mid Q=0)$ = (A2) = $E_X[E(Y_1 \mid Q=1, X) \mid Q=0]$ = (RA) = $E_X[E(Y_1 \mid R=1, X) \mid Q=0]$
= (E-V'.X) = $E_X[E(Y \mid S=1, R=1, X) \mid Q=0]$
Match to each non-participant in the Q=0 group one or more 'similar' individuals from the pool of responding programme group members (S=1, R=1) and take the latter's reweighted outcomes.
- $E(Y_0 \mid Q=0)$ = (A2), (RA), (E-V'.X) = $E_X[E(Y \mid S=1, R=0, X) \mid Q=0]$
Match to each non-participant in the Q=0 group one or more 'similar' individuals from the pool of responding control group members (S=1, R=0) and take the latter's reweighted outcomes.

With the ATE_0 in hand, we can then use the experimental ATE_1 to get the *ATE* via (1b).

(B) Obtaining the *ATE* directly

We need to estimate $E(Y_1)$ and $E(Y_0)$.

- $E(Y_1) \equiv E(Y_1 \mid Q=1 \vee Q=0)$ = (A2) = $E_X[E(Y_1 \mid Q=1, X) \mid Q=1 \vee Q=0]$
= (RA) = $E_X[E(Y \mid R=1, X) \mid Q=1 \vee Q=0]$ = (E-V'.X) =
= $E_X[E(Y \mid R=1, S=1, X) \mid Q=1 \vee Q=0]$ or $E_X[E(Y \mid R=1, S=1, X) \mid (R=1, S=1) \vee Q=0]$
Match each individual in the group made up by the (Q=0 and Q=1) or the (Q=0 and (R=1, S=1)) groups to individuals in the responding programme group sample (R=1, S=1) and calculate the weight that gets assigned to each individual in the latter group (this weight will be >1). Reweigh

the outcomes in this (R=1, S=1) group using these weights and take their average over the (R=1, S=1) group, i.e. use the matched outcome to estimate $E(Y_1)$.

One can match on the basis of this propensity score: $P(Q=0 \mid Q=0 \vee (R=1, S=1), X)$.

- $E(Y_0) \equiv E(Y_0 \mid Q=1 \vee Q=0)$ = (A2) = $E_X[E(Y_0 \mid Q=1, X) \mid Q=1 \vee Q=0]$
  = (RA) = $E_X[E(Y \mid R=0, X) \mid Q=1 \vee Q=0]$ = (E-V'.X) =
  = $E_X[E(Y \mid R=0, S=1, X) \mid Q=1 \vee Q=0]$ or $E_X[E(Y \mid R=0, S=1, X) \mid (R=0, S=1) \vee Q=0]$

  Match each individual in the group made up by the (Q=0 and Q=1) or the (Q=0 and (R=0, S=1)) groups to individuals in the responding control group sample (R=0, S=1) and calculate the weight that gets assigned to each individual in the latter group (this weight will be >1). Reweigh the outcomes in the (R=0, S=1) group using these weights and take their average over the (R=0, S=1) group, i.e. use the matched outcome to estimate $E(Y_0)$.

  One can match on the basis of this propensity score $P(Q=0 \mid Q=0 \vee (R=0, S=1), X)$.

Because of random assignment, the two propensity scores above should be the same and should coincide with $p(x)$.

Appendix 2b Matching Protocol: Allowing for survey and item non-response

In this case we weight the outcomes of the respondents amongst the participants (S=1) so as to reflect the distribution of observables in the full original eligible population.

The first procedure outlined in (A2a) above can correct the ATE_0 for non-response, but would need to be repeated to get a non-response corrected ATE_1 as well:

(A) Obtaining the *ATE* after having first obtained the ATE_0

As in case 2a), to recover the ATE_0, we need to estimate $E(Y_1 \mid Q=0)$ and, given the absence of survey outcomes for non-participants, $E(Y_0 \mid Q=0)$ as well.

Under (A2) and (E-V'.X):

- $E(Y_1 \mid Q=0) = E_X[E(Y \mid R=1, S=1, X) \mid Q=0]$
  This term can be estimated by the matched outcome from matching to each non-participant in the Q=0 group, one or more 'similar' participants from the (R=1 & S=1) group.
- $E(Y_0 \mid Q=0) = E_X[E(Y \mid R=0, S=1, X) \mid Q=0]$
  This term can be estimated by the matched outcome from matching to each non-participant in the Q=0 group, one or more 'similar' participants from the (R=0 & S=1) group.

However, the experimental contrast obtained as $E(Y \mid R=1) - E(Y \mid R=0)$ does not take into account non-response.

One could obtain the correct ATE_1 again by reweighing. Under (RA) and (E-V'.X):

- $E(Y_1 \mid Q=1) = E_X[E(Y \mid R=1, S=1, X) \mid Q=1]$
  This term can be estimated by the matched outcome from matching to each participant in the full Q=1 group, one or more 'similar' programme group members from the respondents, i.e. the (R=1 & S=1) group.
- $E(Y_0 \mid Q=1) = E_X[E(Y \mid R=0, S=1, X) \mid Q=1]$
  This term can be estimated by the matched outcome from matching to each participant in the full Q=1 group, one or more 'similar' control group members from the respondents, i.e. the (R=0 & S=1) group.

To allow for non-response it is thus more convenient to follow option (B) and recover the *ATE* directly:

(B) Obtaining the *ATE* directly

To recover the *ATE*, we need to estimate $E(Y_1)$ and $E(Y_0)$.

Under (A2), (RA) and (E-V'.X):

- $E(Y_1) \equiv E(Y_1 \mid Q=1 \vee Q=0)$ = (A2) = $E_X[E(Y_1 \mid Q=1, X) \mid Q=1 \vee Q=0]$ = (RA) =
  = $E_X[E(Y \mid R=1, X) \mid Q=1 \vee Q=0]$ = (E-V'.X) = $E_X[E(Y \mid R=1, S=1, X) \mid Q=1 \vee Q=0]$

Match each individual in the eligible group, i.e. the Q=0 and Q=1 groups, to individuals in the subgroup of programme group members who responded to the earnings question (R=1 & S=1) and calculate the weight that gets assigned to each individual in the latter subgroup (this weight will be >1). Reweigh the outcomes in the latter subgroup using these weights and take their average over this subgroup.

That is, use the matched outcome to estimate $E(Y_1)$.

One can match on the basis of this propensity score $P(R{=}1 \;\&\; S{=}1 \mid Q{=}0 \vee Q{=}1, X)$.

- $E(Y_0) \equiv E(Y_0 \mid Q{=}1 \vee Q{=}0) = \text{(A2)} = E_X[E(Y_0 \mid Q{=}1, X) \mid Q{=}1 \vee Q{=}0] = \text{(RA)} =$
  $= E_X[E(Y \mid R{=}0, X) \mid Q{=}1 \vee Q{=}0] = \text{(E-V'.}X\text{)} = E_X[E(Y \mid R{=}0, S{=}1, X) \mid Q{=}1 \vee Q{=}0]$

  Match each individual in the eligible group, i.e. the Q=0 and Q=1 groups, to individuals in the subgroup of control group members who responded to the earnings question (R=0 & S=1) group and calculate the weight that gets assigned to each individual in the latter subgroup (this weight will be >1). Reweigh the outcomes in the latter subgroup using these weights and take their average over this group.

  That is, use the matched outcome to estimate $E(Y_0)$.

  One can match on the basis of this propensity score $P(R{=}0 \;\&\; S{=}1 \mid Q{=}0 \vee Q{=}1, X)$.

Take the difference in the two matched outcomes to obtain the *ATE*.

Table A1: Covariate balancing indicators before and after matching

	Prob>chi		Pseudo R2		Median bias		% lost CS
	Before	After	Before	After	Before	After	
Administrative outcomes							
ND25	0.000	1.000	0.069	0.001	4.2	0.6	0.2
NDLP	0.000	1.000	0.121	0.001	3.8	0.8	0.2
Survey outcomes							
a) allowing for non-response, separate common support							
- eligibles vs responding programme group							
ND25	0.000	0.000	0.030	0.005	4.2	1.3	0.3
NDLP	0.000	0.000	0.036	0.006	2.9	1.1	0.1
- eligibles vs responding control group							
ND25	0.000	0.000	0.033	0.006	3.9	1.4	1.2
NDLP	0.000	0.000	0.042	0.008	3.4	1.1	0.6
b) not allowing for non-response, joint common support							
- non-participants vs responding programme group							
ND25	0.000	1.000	0.094	0.003	4.5	1.2	0.8
NDLP	0.000	0.997	0.182	0.005	3.5	1.2	1.0
- non-participants vs responding control group							
ND25	0.000	1.000	0.098	0.004	5.3	1.4	0.8
NDLP	0.000	0.740	0.193	0.008	4.7	2.3	1.0

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.

% lost to CS: Share of the group of non-participants falling outside of the common support.