# Using Administrative Data to Explore the Effect of Survey Nonresponse in the UK Employment Retention and Advancement Demonstration

Richard Dorsett, Richard Hendra and Philip K. Robins

12 September 2016

Dorsett, Richard. Ph.D. Economics. Independent Researcher and Fellow of the National Institute of Economic and Social Research, 2 Dean Trench Street, Smith Square, London SW1P 3HE United Kingdom (email: R.Dorsett@niesr.ac.uk; tel +44 20 7654 1945; fax +44 20 7654 1900). Corresponding author.

Hendra, Richard. Ph.D. Public Policy. Senior Research Associate. MDRC, New York (email: Richard.Hendra@mdrc.org).

Robins, Philip K. Ph.D. Economics. Professor of Economics, University of Miami, Department of Economics (email: probins@miami.edu).

Acknowledgements: We are grateful to the editors of this special edition, Burt Barnow and David Greenberg, to the editor, Jacob Klerman, and to three anonymous reviewers for their helpful comments. Dorsett acknowledges support from the Economic and Social Research Council (grant number ES/J003581/1). The usual disclaimer applies.

## Abstract

**Background:** Even a well-designed randomized control trial (RCT) study can produce ambiguous results. This paper highlights a case in which full-sample results from a large-scale RCT in the United Kingdom (UK) differ from results for a sub-sample of survey respondents. **Objectives:** Our objective is to ascertain the source of the discrepancy in inferences across data sources and, in doing so, to highlight important threats to the reliability of the causal conclusions derived from even the strongest research designs. **Research design:** The study analyzes administrative data to shed light on the source of the differences between the estimates. We explore the extent to which heterogeneous treatment impacts and survey non-response might explain these differences. We suggest checks which assess the external validity of survey measured impacts, which in turn provides an opportunity to test the effectiveness of different weighting schemes to remove bias. The **Subjects** included 6,787 individuals who participated in a large-scale social policy experiment. **Results:** Our results were not definitive but suggest non-response bias is the main source of the inconsistent findings. **Conclusions.** The results caution against overconfidence in drawing conclusions from RCTs and highlight the need for great care to be taken in data collection and analysis. Particularly, given the modest size of impacts expected in most RCTs, small discrepancies in data sources can alter the results. Survey data remain important as a source of information on outcomes not recorded in administrative data. However, linking survey and administrative data is strongly recommended whenever possible.

**Using Administrative Data to Explore the Effect of Survey Nonresponse in the UK Employment Retention and Advancement Demonstration**

## 1. Introduction

The primary virtue of random assignment as a method of evaluating public policy interventions is that it allows for causal inferences owing to the strong internal validity of the design. Because the treatment and control groups are defined at random, statistical equivalence in average characteristics—both measured and unmeasured—is ensured by the design. As several commentators have pointed out, however, issues in data collection can undermine the ability of a randomized control trial (RCT) to estimate the true impact of an intervention (see Barnow & Greenberg, 2014). Data collection problems are exacerbated by the fact that the impacts an RCT is designed to detect are sometimes small, and relatively small levels of bias introduced by data collection practices can alter the overall conclusions.

There are two reasons why impact estimates based on survey data not be the same as those based on administrative data. The first is that the two data sources may differ in what they measure, either because they capture qualitatively distinct concepts or because they differ in the imperfections with which they measure the same concept. For example, when considering earnings impacts, survey and administrative data may differ in the range of employment types for which earnings are recorded. The second reason is that survey respondents may differ in some way from the full sample. The paper focuses on this second reason. Non-comparability in *characteristics* across samples is a more fundamental issue than non-comparability of *outcomes* recorded by administrative data and survey data in the sense that it raises potential concerns over the causal basis of the RCT. If the RCT can be viewed as supporting causal inference among the survey respondent sub-group for administrative outcomes, it gives more confidence when estimating impacts on outcomes that

differ in their definition from those in the administrative data or, more generally, do not exist in the administrative data.

When considering difference between estimates based on the respondent sample and those based on the full sample we explore two potential explanations. The first is that there are no treatment-control differences in survey response but that respondents are more likely than the full sample to have those characteristics associated with a higher impact. In this case, estimates based on survey respondents can sometimes still be viewed as causal, just for a particular subgroup, so the difference from the full sample estimates is explained by *treatment effect heterogeneity*. The second explanation is that there may be treatment-control differences in survey response due to unobserved characteristics. If these unobserved characteristics are correlated with outcomes, they will influence the estimated impact, which can no longer be regarded as causal. This is the case of *differential selection* into the sample of survey respondents.

The main contribution of this paper is to investigate the potential of administrative data for testing the validity of estimates based on survey data. Where administrative data contain a primary outcome and can be linked to survey data, the estimated impact on the primary outcome for survey respondents can be compared to the full sample estimate of the primary outcome. If these estimates agree, or perhaps can be reconciled through appropriate weighting, we can be more confident that non-response does not undermine the ability of survey data to provide causal impact estimates for other outcomes.

The analysis in this paper is based on the United Kingdom (UK) Employment Retention and Advancement Demonstration (hereafter, ERA). ERA was the largest social trial of its kind in Britain. The last wave of survey data, which was intended to be a main source of data for the final evaluation, produced impact estimates that were substantially larger than those obtained using administrative data for the full sample. The paper provides a

detailed account of the administrative and survey data used in the ERA evaluation and shows how the estimated impacts on earnings for survey respondents are much higher than those for the full sample when earnings measures taken from the administrative data are used to derive the estimated impacts for both samples. We consider the question of whether these higher impacts are due to treatment effect heterogeneity or to differential selection bias and, while not conclusive, suggest that the latter is the more likely explanation. As a broader point, our analysis shows that, had the survey been relied upon as the main source of data, our understanding of the effectiveness of UK ERA would have been substantially overstated, assuming the administrative data fully cover the entire experimental sample and have accurate data. It is hoped that by exposing vulnerabilities in data collection the paper will highlight a number of issues to guard against in collecting data for future RCTs.

## 2. The ERA Evaluation Design

ERA tested the effectiveness of a new method of improving the labor market prospects of low-income people relying on various government cash transfers. Hendra et al. (2011) provide a full account of ERA and the context within which the evaluation took place. Here, we outline the main features of ERA.

ERA operated in six regions of Britain from 2003 through 2007. For this paper, we focus on one of the three target groups of ERA, namely out of work single parents on welfare[1] who volunteered for the **New Deal for Lone Parents** (NDLP) welfare-to-work program.[2]

Under ERA, individuals received pre-employment welfare-to-work assistance from

[1] At the time of the evaluation, single parents could claim (means-tested) "Income Support" indefinitely without any requirement to look for work.

[2] The other two target groups were single parents working part-time and receiving the Working Tax Credit (WTC, the UK equivalent of the US Earned Income Tax Credit) and long-term unemployed people aged 25 or older. The first of these groups is not considered in this paper since linking administrative data to WTC recipients has only recently become possible. The second group is excluded since long-term survey data were not collected for this group. Hendra et al. (2011) presents impact results for all groups.

Jobcentre Plus, the public employment service in the UK. The design of ERA allowed for a 9-month pre-employment period. Those who found work became eligible for post-employment services. These included a combination of (caseworker-provided) advice and financial incentives to remain employed and advance in work. Participants who entered and remained in full-time work received substantial cash bonuses (covering up to 24 months of employment), help paying for training courses, and cash rewards for completing training while employed. Under ERA, caseworkers had access to a fund to help avert minor financial emergencies that threatened to prevent a participant from continuing to work. All support under ERA lasted a maximum of 33 months after randomization.

Since there was a limited number of available slots, ERA was implemented as an RCT demonstration, meaning that individuals who volunteered for the program were assigned at random — regardless of their background characteristics — to a treatment group that was enrolled in ERA, or to a control group that was not enrolled in ERA. The control group continued to receive the standard NDLP services as well as any other services normally available to them. Individuals were recruited when they came into Jobcentre Plus offices. Caseworkers recorded basic demographic information and informed individuals of the possible advantages of participating in the ERA program. The caseworkers then invited them to enter the demonstration "lottery", told them they had a 50 per cent chance of being selected for the program and asked them to sign an informed consent form.

Enrollment of families into the experiment lasted a little over a year. Using the background information collected just prior to randomization,[3] the characteristics of the treatment and control groups can be compared in order to assess how well randomization worked. The first two columns of Table 1 relate to the full experimental sample (subsequent columns will be discussed later). From these columns it is clear that randomization

[3] This baseline information was collected as part of the randomization process. Since randomization could not take place until this background information was provided, there are no missing observations.

succeeded in creating two groups that, within sampling variability, are observationally equivalent. The assumption then is that they are also likely to be similar with regard to unobserved characteristics, allowing differences between the ERA group and the control group to be viewed as unbiased estimates of the causal impact of ERA eligibility

<Table 1>

As described in Hendra et al. (2011), the evaluation used outcomes taken from both administrative data and survey data. In that report, however, the survey results were deemphasized based on the finding that the impacts for the survey sample were significantly stronger than the results for the same outcomes in the full sample. These divergent findings, reported in Hendra et al. (2011) are what motivate this paper. Because of these divergent findings, it is appropriate to consider how the details of the survey in order to understand selection into the respondent sample.

The Office for National Statistics (ONS) carried out the survey, using administrative records of benefit receipt to help update survey contact information (Ashton & Portanti, 2011). One concern might be that the treatment itself influenced the probability of response to the survey for instance, if greater contact with bonus recipients resulted in their records being more up-to-date. However, it is not clear how often bonus payment information was used in practice to update contact records.[4] Another possibility is that the financial incentives had a "priming effect" in which an increased likelihood to receive payments for meeting one condition (e.g. working stably) makes one more likely to seek an incentive for another behavior (e.g. filling out a survey)[5]. Though speculative, this pattern of higher than expected survey responses among recipients of program related financial incentives has been seen in

[4]Record-keeping of bonus receipt was relatively *ad hoc* and it is unclear how often benefit records were updated through the bonus payment process.

[5] Survey respondents were given a £20 gift voucher.

other studies.[6] A related possibility is that recipients of bonuses felt a sense of obligation to the program which increased their propensity to respond to the survey.

## 3. Methods

The paper uses simple estimation approaches to explore its key questions. These are briefly summarized in this section. In addition, the data are described and their strengths and weaknesses are critically assessed.

### 3.1 Estimation Approach

We ran a series of regression analyses using the administrative data to estimate the extent to which differences in impacts on the primary earnings outcomes for the wave three (60-month) survey respondent sample and a random sample drawn from administrative records (the "fielded" sample, defined below) can be attributed to several possible sources. Impact models were run for both groups. These models had the following specification:

*(1)* $Y_i = \alpha + \beta P_i + \delta X_i + \varepsilon_i$

where: $Y_j$ is the administrative data outcome measure for sample member $i$, $P_i = 1$ for treatment group members and 0 for control group members, $X_i$ is a set of background characteristics for sample member $i$, $\varepsilon_i$ is a random error term for sample member $i$, $\beta$ is the estimate of the impact of the program on the average value of the outcome, $\alpha$ is the intercept of the regression, and $\delta$ is the set of regression coefficients for the background characteristics.[7] Several logistic regressions were also run which tried to predict treatment status or survey response status. These regressions had the following specifications:

[6] For example, in the New York City Work Rewards study, survey respondents in the treatment group were twice as likely to receive a work reward compared to the full treatment group sample (Nunez et al., 2015 forthcoming).

[7] The regression model consisted of a set of exogenous characteristics selected because of their expected correlation with the key employment and earnings outcomes. The covariates were all measured at baseline and

*(2)* $P_i = \alpha + \delta X_i + \upsilon_i$

*(3)* $R_i = \alpha + \beta P_i + \delta X_i + e_i$

In (3), $R_i$ is a dummy variable which indicates survey response status, $R_i = 1$ for respondents and 0 for non-respondents. We also used these regressions to create inverse probability weights discussed later in the paper.

### 3.2 Data sources

The ERA evaluation reported in Hendra et al. (2011) used both administrative records and survey data. Administrative data originated from the UK Department for Work and Pensions' (DWP's) Work and Pensions Longitudinal Study (WPLS) database. This database has grown in importance as a resource for program evaluations (see Dorsett et al., 2013, for an example of another application to the case of a labor market experiment). It provides information on welfare spells (durations and amounts), employment spells and tax year earnings.[8] A key advantage of these data relative to survey data is that they are available for the full experimental sample.

Within the WPLS, administrative data on welfare receipt and amounts are taken from DWP's payment records and are generally regarded as accurate and reliable. Administrative records on employment and earnings in the WPLS originate from the UK tax department (HMRC – Her Majesty's Revenue and Customs) and are derived from three forms:

- P14 – employers submit this form at the end of each tax year, showing earnings and taxes for each employee. The form covers both employees still with the employer and those who left during the tax year.

(because of random assignment) are orthogonal to treatment group status. The covariates included prior work and barriers to working, race, region, education, and a range of demographic variables. Including covariates increases the precision of the estimated treatment effects.

[8] The UK tax year begins on 6 April and ends on 5 April of the following year.

- P45 – This form has multiple parts. Employers are required to submit one part to HMRC when an employee leaves. The form gives details of the leaving date, earnings in the tax year and the amount of income tax deducted from earnings. The departing employee keeps other parts of the form and must give it to his/her next employer.
- P46 – Employees without a P45 (perhaps because they have not had a previous job or because they are starting a second job) are required to submit a P46 to HMRC. The P46 also provides HMRC with the date of starting employment.

The employment and earnings data require quite substantial cleaning before they are suitable for analysis. For instance, precise start or end dates of employment spells are not always available. Where it is known that a job started (or ended) in a given tax year, but not the precise date, this is recorded on the system as 6 (or 5) April, the first (or last) day of the UK tax year. To improve upon this, part of the data processing for the official evaluation (Hendra et al., 2011) was to randomly imputed such dates within the relevant tax year. In fact, the range for the imputed dates was further narrowed by other available data, such as the file date and the dates of benefit spells. Imputation was used extensively since about one-fifth of all employment spells were missing start dates.

In addition, there may be inconsistencies arising from forms not being submitted or being incorrectly completed. When individuals change employer or hold multiple jobs simultaneously, there is scope for disagreement in recorded dates or earnings. Furthermore, submission of forms is only required for employees earnings above the tax threshold. Despite this, some employers will submit forms for all workers, regardless of their pay, perhaps because batch processing of forms for their higher-earning workers means that it is more efficient to treat lower-earning workers the same way. In addition, these forms do not

capture self-employment and self-employed earnings. The same applies of course to informal work.

In addition to the administrative data, ONS carried out a survey approximately 12 months after the individual's date of random assignment, again at their 24-month anniversary, and finally at their 60-month anniversary. The survey was administered by phone or in person to slightly less than half of the sample of those treatment and control group members randomized between December 2003 and November 2004.[9] The key advantage of the survey data over administrative data is that they provide information that was tailored to the case of ERA. They provide much richer data than the administrative records and allow individuals' experiences with ERA to be assessed as well as key outcome information not otherwise observed; wages, hours of work, type of job and so on.

### 3.3 Administrative Records and Surveys, Advantages and Disadvantages

Large, randomized control trials of public policy interventions often rely heavily on administrative records to quantify the difference that a policy or program makes on key outcomes such as earnings, test scores, or public assistance receipt (see Riccio et al., 2013 for a typical example). The strengths of administrative data are well-known and include wide coverage, no recall bias, and low marginal costs of data collection.

A disadvantage of administrative records is that they typically do not cover all jobs or public assistance. For example, in the US context, state records will not have information on what happens outside the state, and employment records will not have information on jobs in the informal sector (Kornfeld & Bloom, 1999). While the conventional wisdom suggests that under-coverage in administrative records should be equivalent across study groups in an

[9] Approximately 44 percent of the full administrative records sample was selected for fielding for both research groups. The same fielded sample was used for all three waves of the survey survey (that is, the 12, 24 and 60 month waves).

RCT, it is easy to imagine cases where under-coverage can interact with intervention strategies to produce bias (Barnow and Greenberg, 2014; Yang & Hendra, 2016).

Surveys are also an important data source for many RCTs because they provide information that administrative records and other data fail to capture. Without survey data, it would be difficult to quantify program dosage, or the extent to which a person actually engaged with a program.[10] These data also provide valuable insight on certain behaviors, beliefs, program experiences, participant or household characteristics, and other issues that may influence outcomes observed in administrative records. In addition, summarized earnings data from administrative records can be better understood with survey data, which provide information about work schedules, rates of pay, and job changes. Finally, in many domains, administrative records are not available and some evaluations have to depend almost completely on surveys (see Lundquist, 2014 or Banerjee et al., 2009 as examples of studies in which only survey data were available for key outcomes).

Unlike administrative records — where data are obtained for the full study sample — it is relatively expensive to collect survey data. Typically, surveys attempt to collect information from a subset of the full sample, often with the expectation that they will represent the full sample. When a survey fails to be representative — through non-response, for example — it is considered biased. Traditionally, the main safeguard against survey bias has been to obtain a high survey response rate. Recent work has shown, however, that obtaining a high response rate is no guarantee of survey data quality and it is not hard to find examples of surveys with high response rates afflicted with survey non-response bias (e.g. Nunez et al., 2015). Several studies have shown non-response bias does not vary substantially with response rates (Groves, 2006, Groves & Peytcheva 2008). Internal research conducted as part of the US Employment Retention and Advancement study of 16

[10] While most programs collect program participation data in management information systems, these data are typically not available for the control group.

surveys found no correspondence between survey response rates and survey response bias. The implication of these findings is that high survey response rates do not guarantee that survey-based results will generalize to the full sample.

A non-representative survey sample presents an issue of external validity. With an RCT, if non-response affects the treatment and control groups equally, the resulting estimates can still be regarded as causal. The issue in that case is that the results apply to the selected sample of respondents and may not hold for the full population. A more serious issue arises when there is a difference in the response behavior of the treatment and control groups. This differential response behavior results in respondents having different characteristics from the control group respondents so that treatment-control differences in outcomes can no longer confidently be attributed to the program. In other words, differential non-response has the potential to undermine the internal validity of a randomized trial.

## 4. Results

### 4.1 Estimated impacts using administrative data

As described earlier, the main focus of this paper is on the extent to which estimated impacts on outcomes *from the same data source* differ between the subsample of survey respondents and the full experimental sample. To examine the effects of using different samples, we must consider outcomes from the administrative data because this is the only source available for both survey respondents and non-respondents.

Table 2 shows impacts on earnings as recorded in the administrative data for the 2007/8 tax year and the 2008/9 tax year. The "fielded sample" – that is, those individuals for whom a survey was attempted – shows an impact of £343 which is not statistically significant. The fielded sample was drawn from those randomized between December 2003 and November 2004, while the intake period for the experiment ran from October 2003 to

December 2004. Also, the sampling fraction varied by region, resulting in the fielded sample having a different geographic distribution from the full sample. For these reasons, we would not expect the fielded sample results to necessarily agree with the full sample results (also reported in Table 2).

The impact for the respondent sample is much larger (£623 for wave 3 respondents) and is statistically significant. To address the question of whether the estimate for the respondent sample differs significantly from that of non-respondents, we augmented the regression for the fielded sample to include one dummy variable indicating response at wave 3 and another dummy constructed as the interaction of the response dummy with the ERA dummy. The interaction term had a p-value of 0.097, indicating that the estimated impact for respondents differs from the impact for non-respondents at the 10 percent significance level. For 2008/9 earnings, the impact (£320) is again higher than the impact estimated for the fielded sample (£40) but is not statistically significant. Furthermore, estimating the augmented regression again indicates that the difference in impacts between the respondent and non-respondent sample is not statistically significant at the 10 per cent level (p-value of 0.134)

<Table 2>

**4.2 The nature of non-response in the survey data collected for the evaluation**

The main potential problems with survey data are that non-response can harm external validity and treatment-control differences in non-response can harm internal validity. Table 3 shows that there were 6,787 individuals in the full sample. Of those, 2,995 were selected to be in the fielded sample. Interviews were achieved for 87, 77, and 62 per cent of the fielded sample in waves 1, 2 and 3 respectively (that is, 1, 2 and 5 years post-randomization). We

note that response rates are higher among the ERA group than the control group and return to this point below.

<Table 3>

As already noted, overall sample non-response can harm external validity. In other words, the achieved survey sample may not be representative of the population from which it is drawn. Unless we make the assumption that impacts do not vary across individuals, survey non-response means that ERA may affect respondents differently from the full sample. Table 1 provides an indication of the extent to which the background characteristics (observed at the time of randomization) of the fielded sample differ from those of the full sample. For the reasons already given above, we see that there are differences in the geographic distribution of individuals and also in the distribution of randomization timings. In other regards, the fielded sample looks rather similar to the full sample. Table 1 also shows the background characteristics of those individuals who responded to the wave 3 survey (including both treatment and control group members). In the absence of systematic differences in response, wave 3 respondents should resemble the fielded sample.

Table 4 highlights the differences between survey-respondents and non-respondents. Since some of the characteristics seemingly influencing response may be correlated — for example, education and weekly earnings — logistic regression is used to determine which differ across respondents and non-respondents while taking other characteristics into account. Table 4 shows the results of regressing an indicator of response status on the characteristics shown in Table 1, as well as an indicator of research group, in order to better understand the process affecting response. The 'Odds Ratio' column captures the effect of each characteristic on the probability of responding to the survey; asterisks denote the significance level of these relationships.

Survey respondents differ from non-respondents in several characteristics. Those who, at baseline, were from Wales, those who were unmarried and living alone, and those with no qualifications were less likely to respond. These differences suggest that the survey sample may not be representative of the fielded sample. Selective response due to a nonrepresentative survey sample can result in different impacts if the sample that is more likely to respond has a different pattern of impacts compared to the full sample. Impacts from the selected sample may still be internally valid and therefore provide valid causal estimates but, with treatment effect heterogeneity, these impacts will not generalize to the full fielded sample.

More worrisome is the possibility that attrition results in estimated impacts that are no longer internally valid. Attrition bias arises when treatment group respondents differ from control group respondents with regard to unobserved characteristics correlated with outcomes. This includes the possibility that the program itself may influence survey response (for reasons described earlier in the paper). Table 4 indeed shows that those in the treatment group are more likely to respond than those in the control group.[11] While not a sufficient condition for internal validity to be undermined — it is straightforward to see that a randomly lower response rate among the control group will not bias impact estimates — it raises a note of caution. Thus, if nothing else, differential non-response serves to reduce the credibility, or face validity, of the experimental design.

<Table 4>

A common practice when assessing whether survey non-response may have introduced bias is to compare baseline characteristics of respondents in the treatment and control groups. Just as differential non-response is not a sufficient condition for bias, neither is balanced response a sufficient condition for unbiasedness (Barnow and Greenberg, 2015). In other words, having

[11] This result can be inferred from Table 4 in a number of ways, including the observation that ERA group members constitute a larger fraction of the respondent sample compared to the non-respondent sample.

equivalent response rates by research group does not preclude the possibility of compositional differences under the surface. Again, Table 1 is informative and suggests, at first glance, that treatment and control group respondents are similar. To explore treatment-control comparability, we estimated a logistic regression to determine the extent to which baseline characteristics could predict whether a respondent was a member of the treatment group (among wave 3 respondents only). Table 5 shows that none of the baseline characteristics is statistically significant as a predictor. In other words, among the survey-respondent sample, there are no differences between treatment and control group respondents in these background characteristics.

<Table 5>

### 4.3 Attempts to reconcile earnings impacts estimated on the respondent sample with those estimated on the fielded sample

Tests of response bias often focus on baseline data. However, when suitable administrative data are available, it is also informative to investigate differences in outcomes subsequent to baseline. One pattern of note in ERA was that treatment group members who worked stably were disproportionately likely to respond to the survey compared to stably employed control group members. This differential survey response is consistent with the fact that respondents tended to have higher earnings than non-respondents. The extent of this differential survey response varied across treatment status; in the control group, mean income among non-respondents was 84 per cent of the mean income for respondents while, in the treatment group, this fell to 72 per cent (these percentages were stable across both 2007/8 and 2008/9 earnings). Furthermore, even when administrative data are not available for both research groups, it may be possible to explore how response to the survey is correlated with program take-up among the control group. With ERA, treatment group respondents were over 7 percentage points more likely to receive the work retention bonus compared to treatment group non-respondents.

Reflecting these findings about differences in employment stability and bonus receipt outcomes across samples, we explored several reweighting strategies intended to bring the survey respondent sample earnings impact estimate into alignment with the fielded sample impact estimate. The results of these efforts are summarized in Table 6. The first two rows show again, for convenience, the estimated impacts for the fielded and respondent survey samples. The next row gives the results of a conventional weighting strategy using weights based on the inverse of the probability of responding conditional on a set of background characteristics. This conventional weighting strategy had little effect on aligning earnings impacts across the samples (the estimated impact is £546). It is not surprising that this approach was ineffective given the finding (discussed earlier) that there was no observable bias based on background characteristics. Nonetheless, it is still noteworthy that the common approach of using weights defined on the basis of background characteristics does little to bring the respondent sample impact estimates closer to the fielded sample impact estimates.

We also attempted non-experimental weighting strategies that control for post-randomization outcomes. Analyses of RCTs are usually careful to condition only on pre-randomization treatment-control differences since randomization itself ensures that unobserved characteristics balance post-randomization. Controlling for post-randomization outcomes runs counter to pre-randomization control and represents a significant departure from standard practice. However, as illustrated in this paper, non-response can undermine the statistical properties of an RCT to the extent that the basis for causal interpretation of estimated impacts is eroded. In such a scenario, it may be appropriate to consider weighting individuals according to their outcomes. In the case of ERA, such an approach is successful in reconciling the impacts estimated for respondents with those for the fielded sample.

As discussed above, survey respondents were disproportionately more likely to receive the employment retention bonus. Weighting based on a combination of baseline

characteristics and bonus receipt rates brought the survey respondent sample impact estimate into approximate alignment with the fielded sample impact estimate (£394 compared to £343). It was also noted above that treatment group members who worked stably were more likely to respond to the survey compared to control group members who worked stably. Weighting based on employment stability also brought the impact estimate for the survey respondent sample much closer to the impact for the fielded sample estimate (£334 compared to £343).[12]

<Table 6>

## 5. Conclusion

To summarize the evidence presented in this paper, UK ERA was a well-executed social experiment and there is every indication that two statistically equivalent groups were obtained from the random assignment procedure. A survey carried out five years post-randomization achieved a response rate of 62 per cent; 64 per cent for the treatment group and 60 per cent for the control group. As documented in Hendra et al. (2011), the estimated earnings impact using survey data for the survey-respondent sample was greater than the estimated earnings impact for the full sample using administrative data, raising the question of how to interpret this difference. Access to administrative data allows an assessment of the degree and nature of non-response bias that would not otherwise be possible and we have explored this in this paper. Had the survey been the only source of earnings data, the estimated impacts would have overstated the effectiveness of the program, assuming the administrative data represent the truth about the sample.[13]

Also important is that the non-response appears to bias the estimated earnings impact despite the treatment and control groups in the respondent sample being similar with regard

[12] Another approach to weighting these results would be to adjust earnings impacts by the ratio of non-respondent to respondent earnings separately for the treatment and control groups. Such an approach was used in the National Job Corps study (Schochet, McConnell, & Burghardt, 2003)

[13] Some authors question whether administrative data, even if covering the entire experimental sample, represent the truth. See for example, Kapteyn and Ypma (2007) and Abowd and Stinson (2013).

to observed characteristics. Demonstrating such similarity is often used as part of the evidence to argue that impacts estimated on respondent subsamples retain their causal interpretation. However, the results in this paper demonstrate that this is not a sufficient condition, and raises the possibility that post-random assignment factors including differential access to the research groups or even aspects of the intervention can cause bias that would not be evident by examining characteristics at baseline.

It might still be the case that the impacts estimated for the respondent sample are causal but that impacts are heterogeneous in the population and are different for respondents compared to non-respondents. In line with this, the results confirm that respondents and non-respondents have different characteristics. If impact heterogeneity were the sole reason for the difference in estimated impacts, one would expect that re-weighting the respondent sample to resemble the full sample *using background characteristics* would bring estimated impacts closer. Attempts to do this were unsuccessful, so we conclude that respondents differ from non-respondents in some unobserved way.[14] It is still conceivable that the impacts estimated for the respondent sample are causal. This would rely on there being an unobserved characteristic that was positively correlated with both survey response and impacts. However, an alternative possibility is that the estimated impact for the respondent sample is no longer causal due to unobserved treatment-control differences. We have no obvious way of distinguishing between these two scenarios. However, the fact that we have no strong theory to suggest the former is the case leads us to regard the latter as being the more likely explanation.

Our findings highlight the usefulness of administrative data for exploring the reliability of impacts estimated using survey data. Of course, for outcomes available in administrative data, there may be no need to rely on the survey-respondent subgroup.

[14] Although weights based on outcomes brought the samples into alignment on earnings impacts, these weights were not used in the impact report because they result in non-experimental estimates.

However, the real value of such an exploration derives from its implications for the analysis of outcomes not present in administrative data. An important reason for carrying out surveys is to collect information that is not available elsewhere. If respondent sample results can be satisfactorily reconciled with full sample results of an outcome available in the administrative data, we can be more confident that impacts on outcomes only available in the survey data can be credibly estimated.[15] As an aside, we note that the issues discussed here are not unique to experiments. It is however true that experiments make the potential problems more visible.

Recent methodological developments provide some hope of dealing with possibly biasing non-response. A simulation study by Puma et al. (2009) shows that implementing multiple imputation to fill in missing survey data using information from administrative records can substantially address missing survey data problems. In addition, development of estimators suited to the case of non-random subsamples remains a live issue in econometric theory (d'Haultfoeuille, 2010; Ramalho and Smith, 2013). Furthermore, improvements in survey data weighting, notably through the use of so-called survey "paradata", have shown some progress in improving alignment across the data sources.[16]

While these approaches might help, they inevitably complicate the estimation of impacts and reduce the transparency that is an attractive feature of RCTs. Indeed, some approaches rely on assumptions that imply it is no longer appropriate to regard the resulting estimates as truly experimental. While such approaches hold promise, a pragmatic approach is to employ an ensemble strategy, using multiple data sources to estimate impacts and

[15] This holds if there is no significant item nonresponse bias in the survey.

[16] Paradata are survey administrative data which capture the effort required to reach a respondent (Krueter et al, 2010; Heffetz & Rabin, 2013). The essential idea, consistent with the Lin and Schaeffer (1995) model of a "continuum of resistance" in survey response, is that late survey responders, who take more effort to reach (measured, typically, by the number of attempts to reach them) are more similar to nonrespondents compared to those who respond earlier. Therefore, up-weighting late respondents can be an effective means of addressing nonresponse bias.

attempt to achieve a clear understanding of the uncertainties inherent in the ability of any particular data source to capture the true impact.

## References


Abowd, John M. and Martha Stinson. 2013. "Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data, *The Review of Economics and Statistics*, 95(5): 1,451-1,467.

Ashton, K. & Portanti, M. (2011) "Tracking Procedures on the Employment, Retention and Advancement Survey" Office for National Statistics Survey Methodology Bulletin No. 68: 12-22.

Banerjee, A. V., Duflo, E., Glennerster, R., & Kinnan, C. (2013). The miracle of microfinance? Evidence from a randomized evaluation.

Barnow, B. And Greenberg, D. (2014) "Do Estimated Impacts on Earnings Depend on the Source of the Data Used to Measure Them? Evidence from Previous Social Experiments". *Evaluation Review*.

Bloom, D., Hendra, R., & Page, J. (2006). The Employment Retention and Advancement Project: Results from the Chicago ERA Site. *Report to the US Department of Health and Human Services. New York: MDRC*.

d'Haultfoeuille, Xavier. "A new instrumental method for dealing with endogenous selection." *Journal of Econometrics* 154.1 (2010): 1-15.

Dorsett, R., Smeaton, D. and Speckesser, S. (2013) "The effect of making a voluntary labour market programme compulsory: evidence from a UK experiment". *Fiscal Studies* 34(4): 467-489.

Groves, Robert M. and Emilia Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias." *Public Opinion Quarterly* 72, 2: 167-189.

Groves, Robert M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70, 5: 646-675.

Heffetz, O., & Rabin, M. (2013). Conclusions regarding cross-group differences in happiness depend on difficulty of reaching respondents. *American Economic Review*, *103*(7), 3001-3021.

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

Kapteyn, Arie and Jelmer Y. Ypma. 2007. "Measurement Error and Misclassification: A Comparison of Survey and Administrative Data." *Journal of Labor Economics*, 25(3): 513-551.

Lundquist, E, Hsueh, J., Lowenstein, A., Fauceetta, K. Gubits, D. Michalopoulos, C., and Knox, V. (2014). "A Family-Strengthening Program for Low-Income Families: Final Impacts from the Supporting Healthy Marriage Evaluation." New York: MDRC.

Kornfeld and Bloom. 1999. "Measuring Program Impacts on Earnings and Employment: Do Unemployment Insurance Wage Reports from Employers Agree with Surveys of Individuals?" *Journal of Labor Economics 17, 1.*

Kreuter, F., Couper, M., & Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. In *Proceedings of the Joint Statistical Meetings, American Statistical Association* (pp. 282-296).

Lin, I-Fen and Nora Cate Schaeffer. 1995. "Using Survey Participants to Estimate the Impact of Nonparticipation." *Public Opinion Quarterly* 59, 2: 236-258.

Maguire, Shelia, Joshua Freely, Carol Clymer, Maureen Conway, and Deena Schwartz. 2010. *Tuning in to Local Labor Markets: Findings from the Sectoral Employment Study.* Philadelphia, PA: Public/Private Ventures.

Nunez, S. Yang, E. and Verma N. (2015, forthcoming) "Interim Impacts of the Work Rewards Demonstration.". New York, NY: MDRC.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). What to Do when Data Are Missing in Group Randomized Controlled Trials. NCEE 2009-0049. *National Center for Education Evaluation and Regional Assistance*.

Ramalho, E. and Smith, R. (2013) "Discrete choice non-response." *The Review of Economic Studies* 80(1):343–364.

Riccio, J., Dechausay, N., Miller, C., Nunez, S., Verma, N., & Yang, E. (2013). Conditional Cash Transfers in New York City: The Continuing Story of the Opportunity NYC-Family Rewards Demonstration. *MDRC.*

Schochet, P., Burghardt, J., & McConnell, S. (2006). *National job corps study and longer-term follow-up study: Impact and benefit-cost findings using survey and summary earnings records data.* Princeton, NJ: Mathematica Policy Research Inc.

Yang.,E & Hendra, R., (2016) The Importance of Gathering Data From Both Administrative Records and Surveys: Lessons from Two Conditional Cash Transfer Programs in New York City. Mimeo.

**Table 1 Descriptive statistics for full sample and wave 3 fielded and respondent samples**

| | Full sample | | Fielded sample, | | Respondents sample | |
|---|---|---|---|---|---|---|
| | Control | ERA | Control | ERA | Control | ERA |
| District | | | | | | |
| East Midlands | 0.24 | 0.24 | 0.17 | 0.17 | 0.18 | 0.17 |
| London | 0.23 | 0.22 | 0.18 | 0.17 | 0.17 | 0.16 |
| North East England | 0.19 | 0.19 | 0.17 | 0.18 | 0.21 | 0.20 |
| North West England | 0.15 | 0.15 | 0.17 | 0.17 | 0.17 | 0.18 |
| Scotland | 0.09 | 0.09 | 0.16 | 0.16 | 0.16 | 0.16 |
| Wales | 0.10 | 0.10 | 0.16 | 0.15 | 0.12 | 0.13 |
| Date of random assignment (RA) | | | | | | |
| October 2003 - December 2003 | 0.10 | 0.10 | 0.04 | 0.04 | 0.03 | 0.03 |
| January 2004 - March 2004 | 0.29 | 0.30 | 0.35 | 0.34 | 0.34 | 0.33 |
| April 2004 - June 2004 | 0.21 | 0.21 | 0.24 | 0.25 | 0.25 | 0.25 |
| July 2004 - September 2004 | 0.24 | 0.24 | 0.26 | 0.26 | 0.25 | 0.27 |
| October 2004 - December 2004 | 0.14 | 0.14 | 0.12 | 0.12 | 0.13 | 0.12 |
| January 2005 - April 2005 | 0.02 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 |
| Female | 0.95 | 0.95 | 0.95 | 0.94 | 0.96 | 0.94 |
| Single | 0.71 | 0.72 | 0.72 | 0.73 | 0.71 | 0.71 |
| Number of children | | | | | | |
| None | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 |
| One | 0.51 | 0.50 | 0.53 | 0.52 | 0.53 | 0.50 |
| More than one | 0.43 | 0.45 | 0.42 | 0.44 | 0.43 | 0.45 |
| Education | | | | | | |
| O-level | 0.48 | 0.48 | 0.47 | 0.48 | 0.47 | 0.51 |
| A-level or above | 0.21 | 0.22 | 0.22 | 0.22 | 0.23 | 0.22 |
| Other | 0.08 | 0.07 | 0.08 | 0.07 | 0.08 | 0.06 |
| None | 0.23 | 0.23 | 0.24 | 0.23 | 0.21 | 0.21 |
| Months worked in three years prior to RA | | | | | | |
| 12 or fewer | 0.73 | 0.72 | 0.72 | 0.72 | 0.72 | 0.72 |
| 13-24 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.12 |
| More than 24 | 0.14 | 0.15 | 0.15 | 0.16 | 0.15 | 0.16 |
| Worked in the past year | 0.28 | 0.30 | 0.28 | 0.30 | 0.28 | 0.31 |
| Weekly earnings in the past year for current/most recent job (£) | 26.12 | 28.49 | 26.34 | 28.67 | 26.00 | 27.88 |
| Average number of months on benefits in the two years prior to RA | 17.44 | 17.16 | 17.32 | 17.02 | 17.38 | 17.11 |
| N | 3,422 | 3.365 | 1,511 | 1,481 | 903 | 951 |

**Table 2: The estimated impact of ERA for different samples, using administrative records on earnings**

| | Earnings 2007/8 | Earnings 2008/9 |
|---|---|---|
| Full sample | 124 | -25 |
| | (0.77) | (0.14) |
| N | 6787 | 6787 |
| | | |
| Fielded sample | 343 | 40 |
| | (1.46) | (0.16) |
| N | 2992 | 2992 |
| | | |
| Respondents to survey, wave 3 | 623** | 320 |
| | (2.00) | (0.92) |
| N | 1854 | 1854 |

Note: t-statistics in parentheses. Asterisks indicate statistical significance of the estimates: * significant at the 90% level, ** significant at the 95% level, ** * significant at the 99% level. Estimates control for region, cohort, sex, age, qualifications, number of months employed in the three years before randomization, number of months on welfare in the two years before randomization and whether their youngest child is under the age of five at randomization.

**Table 3 Survey response rates**

| | Total | ERA group | Control group |
|---|---|---|---|
| Full sample size | 6,787 | 3,365 | 3,422 |
| Wave 1 (12-month survey) | | | |
| Fielded sample size | 2,995 | 1,482 | 1,513 |
| Respondent sample size | 2,604 | 1,317 | 1,287 |
| Non-respondent sample size | 391 | 165 | 226 |
| Response rate (%) | 86.9 | 88.9 | 85.1 |
| Wave 2 (24-month survey) | | | |
| Fielded sample size | 2,995 | 1,482 | 1,513 |
| Respondent sample size | 2,297 | 1,188 | 1,109 |
| Non-respondent sample size | 698 | 294 | 404 |
| Response rate (%) | 76.7 | 80.2 | 73.3 |
| Wave 3 (60-month survey) | | | |
| Fielded sample size | 2,992 | 1,481 | 1,511 |
| Respondent sample size | 1,854 | 951 | 903 |
| Non-respondent sample size | 1,138 | 530 | 608 |
| Response rate (%) | 62.0 | 64.2 | 59.8 |

**Table 4 Descriptive statistics for wave 3 (60-month) respondent and non-respondent samples, and odds ratios from a logistic regression of survey response**

| | | Respondents | Non-respondents | Odds Ratio | S.E. | |
|---|---|---|---|---|---|---|
| ERA | | 51.3 | 46.6 | 1.206 | 0.077 | ** |
| District (%) | | | | | | |
| | East Midlands | 17.7 | 15.3 | 1.387 | 0.131 | ** |
| | London | 16.3 | 19.2 | | | |
| | North East England | 20.1 | 12.8 | 1.935 | 0.134 | *** |
| | North West England | 17.3 | 15.9 | 1.380 | 0.132 | ** |
| | Scotland | 16.1 | 16.3 | 1.227 | 0.132 | |
| | Wales | 12.5 | 20.5 | 0.756 | 0.132 | ** |
| Date of random assignment (RA) (%) | | | | | | |
| | Oct 03 – Dec 03 | 3.2 | 4.6 | | | |
| | Jan 04 – Mar 04 | 33.5 | 35.8 | 1.388 | 0.206 | |
| | Apr 04 – Jun 04 | 24.8 | 23.9 | 1.514 | 0.210 | ** |
| | Jul 04 – Sep 04 | 26.3 | 24.9 | 1.529 | 0.209 | ** |
| | Oct 04 – Dec 04 | 12.1 | 10.9 | 1.450 | 0.227 | |
| | Jan 05 – Mar 05 | 0.0 | 0.0 | n/a | n/a | |
| Female (%) | | 95.0 | 93.7 | 1.278 | 0.169 | |
| Single (%) | | 71.4 | 75.4 | 0.792 | 0.091 | ** |
| Number of children (%) | | | | | | |
| | None | 1.4 | 0.7 | | | |
| | One | 53.1 | 55.9 | 0.666 | 0.328 | |
| | More than one | 45.5 | 43.4 | 0.722 | 0.330 | |
| Education (%) | | | | | | |
| | O-level | 49.1 | 44.2 | 1.514 | 0.098 | *** |
| | A-level or above | 22.6 | 21.3 | 1.481 | 0.116 | *** |
| | Other | 7.3 | 6.9 | 1.538 | 0.164 | *** |
| | None | 21.0 | 27.6 | | | |
| Number of months worked in three years prior to RA (%) | | | | | | |
| | 12 or fewer | 72.2 | 71.5 | 0.954 | 0.131 | |
| | 13-24 | 12.2 | 14.2 | 0.795 | 0.147 | |
| | More than 24 | 15.6 | 14.3 | | | |
| Worked in the past year (%) | | 29.5 | 28.7 | 1.246 | 0.137 | |
| Weekly earnings most recent job pre-RA (£) | | 27.0 | 28.4 | 0.998 | 0.001 | |
| Avg months on benefits in the 2 years pre-RA | | 17.2 | 17.0 | 1.005 | 0.005 | |
| N | | 1,854 | 1,138 | 2,992 | | |

Note: Asterisks indicate statistical significance of the estimates: * significant at the 90% level, ** significant at the 95% level, ** * significant at the 99% level.

**Table 5 Baseline characteristics as a predictor of treatment status, among wave 3 survey respondents**

| | Odds ratio | (Standard error) |
|---|---|---|
| District (%) | | |
| East Midlands | 0.997 | 0.161 |
| North East England | 0.997 | 0.157 |
| North West England | 1.169 | 0.164 |
| Scotland | 1.034 | 0.165 |
| Wales | 1.158 | 0.178 |
| Date of random assignment (RA) (%) | | |
| January 2004 - March 2004 | 0.872 | 0.275 |
| April 2004 - June 2004 | 0.937 | 0.278 |
| July 2004 - September 2004 | 1.019 | 0.277 |
| October 2004 - December 2004 | 0.841 | 0.295 |
| January 2005 - April 2005 | *n/a* | *n/a* |
| Female (%) | 0.736 | 0.221 |
| Single (%) | 1.033 | 0.108 |
| Number of children (%) | | |
| One | 0.713 | 0.371 |
| More than one | 0.829 | 0.372 |
| Education (%) | | |
| O-level | 1.124 | 0.125 |
| A-level or above | 0.977 | 0.146 |
| Other | 0.755 | 0.202 |
| Number of months worked in three years prior to RA (%) | | |
| 12 or fewer | 1.089 | 0.155 |
| 13-24 | 0.941 | 0.182 |
| Worked in the past year (%) | 1.262 | 0.164 |
| Weekly earnings in the past year for current/most recent job (£) | 0.999 | 0.001 |
| Number of months on benefits in the two years prior to RA | 0.995 | 0.007 |
| Sample size | 1,854 | |

Note: Asterisks indicate statistical significance of the estimates: * significant at the 90% level, ** significant at the 95% level, ** * significant at the 99% level.

**Table 6 Exploring weighting approaches to reconcile 2007/8 earnings impacts across fielded and respondentWave 3 (60-month) samples**

| | ERA | Control | Impact | P-value |
|---|---|---|---|---|
| Fielded sample (unweighted) | 5,504 | 5,161 | 343 | 0.143 |
| Respondent sample (unweighted) | 5,987 | 5,364 | 623** | 0.045 |
| Respondent sample, weighted according to: | | | | |
| - baseline characteristics only | 6,589 | 6,043 | 546* | 0.080 |
| - baseline characteristics and bonus Receipt indicators | 6,352 | 5,957 | 394 | 0.204 |
| - baseline characteristics and employment stability | 6,078 | 5,743 | 334 | 0.277 |

Note: Asterisks indicate statistical significance of the estimates: * significant at the 90% level, ** significant at the 95% level, ** * significant at the 99% level. Estimates control for region, cohort, sex, age, qualifications, number of months employed in the three years before randomization, number of months on welfare in the two years before randomization and whether their youngest child is under the age of five at randomization.