Journal of Official Statistics, Vol. 26, No. 1, 2010, pp. 105–125

Adjusting for Nonignorable Sample Attrition Using Survey Substitutes Identified by Propensity Score Matching: An Empirical Investigation Using Labour Market Data

Richard Dorsett[1]

This article assesses the potential for reducing attrition bias by replacing survey dropouts with substitutes drawn from the same population and identified using propensity score matching. By linking register data with survey data, it is possible to observe unemployment outcomes for dropouts and therefore to test models of attrition. Doing so reveals the dropout process for unemployment to be nonignorable in this survey such that the commonly-used method of reweighting non-dropouts on the basis of sample frame information will be ineffective in overcoming attrition bias. The results indicate the effectiveness in theory of using substitutes but suggest that practical applications may only be successful where it is possible to incorporate information additional to that available in the sampling frame. Under such circumstances, it may similarly be possible to address nonignorable attrition by reweighting respondents.

Key words: Attrition; propensity score matching; survey substitutes; unemployment.

1. Introduction

Increasingly, research in economics, sociology and other disciplines makes use of panel data. The appeal of such data (in which the same units are observed at multiple points in time) is that they permit the consideration of more complex relationships than is possible using cross-section data. The drawback is that individuals who participate in the first wave of a panel may drop out in later waves. This may mean that those who continue to respond become increasingly unrepresentative of the original sample. If the tendency to drop out is systematically related to an outcome of interest, estimates of sample moments for this outcome can be biased. In view of this, methods to overcome such attrition bias have an important empirical role.

Most commonly, the approach adopted to account for nonresponse is to reweight the respondent sample using information available at the time the sample was drawn. Whether this is appropriate depends on the underlying process generating nonresponse. In the case

[1] National Institute of Economic and Social Research, 2 Dean Trench Street, London SW1P 3HE, UK. Email: R.Dorsett@niesr.ac.uk

Acknowledgments: The author gratefully acknowledges financial support from the ESRC (ref: R000223901). The article has benefitted from discussion and comments from Peter Dolton, Peter Lynn, Jeffrey Smith, and Michael White as well as participants at seminars at: the Policy Studies Institute; the University of Westminster; the 2003 annual general meeting of the Statistical Society of Canada in Halifax, Nova Scotia; the second workshop on Research Methodology in Amsterdam, June 2003; the Royal Statistical Society one day conference on attrition and nonresponse in social surveys in London, May 2004; and the MOLS 2006 conference at the University of Essex.

of the dropout process being ignorable (Rubin 1976; Little and Rubin 2002), reweighting in this way can be effective. For example, in the UK it is common for response in London to be lower than in other parts of the country. An observable deviation of this type can be addressed through reweighting.

In the nonignorable case, other approaches are needed. Van den Berg et al. (2006) suggest factors that may result in nonignorable nonresponse when dealing with labour market data. It may be that individuals have unobserved characteristics or circumstances that influence both their attitude to surveys and their success in the labour market. For example, those involved in time-consuming job search may have little remaining time (or enthusiasm) to participate in a survey. Furthermore, there may be a causal relationship between the outcome of interest and survey response. This may arise, for instance, by those finding jobs becoming more difficult to contact, perhaps because they have changed location.

In this article, we investigate the potential for using substitutes to replace dropouts from a longitudinal survey of individuals participating in a UK active labour market programme (the New Deal for Young People – NDYP). The survey was carried out in two waves and was characterised by substantial nonignorable attrition. The data were linked to unemployment register data for all individuals, allowing claimant unemployment to be observed on an ongoing basis, regardless of whether individuals responded to the survey at Wave 1, Wave 2 or not at all.

The approach in this article uses propensity score matching to select from the original population substitutes who are similar to the dropouts in terms of observable characteristics. Propensity score matching was introduced by Rosenbaum and Rubin (1983) as a means of constructing a control group with nonexperimental data and has been used extensively in the evaluation of treatment effects (see, for example, Dehejia and Wahba 2002). A key labour market outcome of interest is whether individuals continue to claim benefit. Since this is observed for the full sample, we can compare the level of benefit receipt for non-dropouts to that for dropouts and thereby quantify the extent to which attrition biases estimates of unemployment. Similarly, comparing the level of benefit receipt for dropouts to that for substitutes shows whether replacing dropouts with substitutes can overcome attrition bias.

Clearly, where the outcome of interest is available in register data, the problem of sample attrition is greatly reduced. However, the relevance of the analysis in this article is that it is informative of the more general situation in which the outcome variable of interest is not observable in linked register data. Furthermore, the usual purpose in using panel data is not to consider cross-sectional outcome measures but rather to focus on processes and dynamic relationships. With this in mind, the approach in this article also allows an insight into the extent to which replacing dropouts with substitutes can allow consideration of events prior to the time of substitution.

The approach explored in this article is related to a body of literature concerned with methods to overcome the problem of survey nonresponse. A number of papers consider the role of the propensity score in this regard. Little (1986) considers stratifying the sample on the basis of the estimated propensity score and adjusting for nonresponse through the use of weights calculated as the reciprocal response rate within each stratum (see also Cassel et al. 1983; Little and Rubin 2002). Stratifying according to the propensity score avoids the problem that can arise with direct adjustment (that is, weighting by reciprocal response rates within adjustment cells defined by one or more background variables). Specifically, individuals within an adjustment cell characterised by a very low response rate can receive large weights (Rosenbaum 1987 expands on this issue). Little (1986) also discusses the issue of item nonresponse and how the missing values may be imputed as the mean within strata based on the propensity score. Lavori et al. (1995) build on this approach, using such a stratification for multiple imputation of missing responses among survey dropouts. In their application, each imputed outcome is a random draw within strata defined with reference to the estimated propensity score. Interestingly, they allow the propensity score to be estimated using outcomes observed occurring after the time the original sample was drawn.

Rubin and Zanutto (2002) provide an overview of the empirical use of survey substitutes, drawing mainly on Chapman (1983) and Vehovar (1999). From this it appears that the evidence on the successfulness of substitutes is mixed. A consistent issue that emerges is that the use of substitutes often faces practical difficulties (Chapman 2003; Vehovar 2003), particularly when interviewers have some role in the selection of substitutes. Furthermore, the effort devoted by interviewers to contacting initial sample units may reduce when substitutes are available. Since more time is often allowed for contacting initial sample units than for contacting substitutes, responding substitutes may be more likely to resemble early responders from the initial sample rather than the entire initial sample (Chapman and Roman 1985). Such concerns point to the importance of ensuring that, as far as possible, the distinction between original sample members and potential substitutes is invisible to interviewers, who, ideally, need not even be aware that the sampling strategy includes provision for replacing dropouts.

The analysis in this article makes a number of contributions. First, it suggests how a test of the process generating attrition can be applied in the case where an outcome of interest is observed in linked register data. Second, it explores whether using propensity score matching to identify substitutes for the dropouts can effectively address nonignorable attrition bias. To begin with, the ideal case is considered whereby substitutes are drawn from the pool of individuals who would not respond to the survey. Two more realistic scenarios are then considered. One possibility is to use non-dropouts as the source of substitutes. This can be seen as a form of reweighting. Another possibility is to draw substitutes from a purposive population made up of individuals thought *likely* not to respond. An important feature of the analysis in this article is that outcome information since the time of sampling is incorporated into both the estimation of the propensity scores and the identification of the purposive pool of substitutes (those thought likely not to respond). This is in the spirit of Lavori et al. (1995) and distinguishes the analysis from other related treatments such as that of Chapman (2003), who considers purposive selection of substitutes from a sample on the basis of variables that exist in the sampling frame, or that of Lynn (2004), who argues that the substitute should be a random draw from a stratum defined over one or more frame variables.

The structure of the article is as follows. To fix ideas, we begin in the next section by describing the survey data and the extent of sample attrition. The two dominant models of attrition are outlined and a test is carried out to determine whether the dropout process is ignorable. This in itself is an unusual step since most analyses are based on an untested assumption as to which attrition model is appropriate. However, when survey data can be linked to administrative data in the way described above, such tests can provide a useful guide as to how best to proceed to overcome attrition bias. Section 3 sets out the method of selecting substitutes for the dropouts using propensity score matching and Section 4 describes the strategy for testing this approach. The latter section begins with a consideration of the potential of the approach when an ideal pool of substitutes is available. This is a somewhat theoretical undertaking in that it ignores the fact that these substitutes themselves will also be characterised by nonresponse. To achieve a more realistic example, we then consider a purposive pool of potential substitutes and simulate survey response at the time of the Wave 2 interviews. We also consider how the performance of the approach might improve were it possible to achieve an increase in the response rate. The main results are presented in Section 5, where we examine the extent to which replacing dropouts with survey substitutes can reduce bias in simple estimates of mean unemployment. Section 6 concludes.

2. Attrition in the NDYP Data

NDYP is an active labour market programme introduced in Britain in 1998, mandatory for all those aged 18–24 who have been claiming unemployment benefits for a period of six months or longer. It was evaluated using a survey design that involved two stages of interviews with the same individuals. The sample was drawn from those entering NDYP between 29 August and 27 November 1998. Survey interviews took place at about 6 and 15 months after entering NDYP and the results of these surveys are reported in Bryson et al. (2000) and Bonjour et al. (2001). Only those responding at the first stage of interviewing were approached for a second interview. As noted in Section 1, survey data were linked to register data allowing unemployment outcomes for the full sample to be observed regardless of whether individuals responded to the survey.

There was substantial attrition between the two survey waves. Of the 5,910 individuals who responded at Wave 1, only 3,335 responded at Wave 2. This represents an attrition rate of 44 per cent. Figure 1 shows how unemployment differs over time between the dropouts and those who remained in the sample. The vertical lines correspond to the average dates of the Wave 1 and Wave 2 interviews respectively. The proportions claiming benefit at any point are shown with 95 per cent confidence intervals in order to demonstrate the statistical significance of the differences. In fact, very soon after the point of sampling, significant differences are evident. The size of these differences grows over time, reducing only at the most recent observation periods. Clearly, for most of the period shown in the graph, dropouts were significantly less likely to be unemployed than those who remained in the sample. This accords with the situation of those entering employment being less likely to respond to surveys due to lack of available time, moving to take up work, or some other reason. It should be remembered that the client group for NDYP is young and geographically mobile. The consequence of this is that the probability of not being able to locate individuals increases.

It is not possible to discern from Figure 1 whether or not the evident attrition is ignorable (alternatively, "missing at random" – MAR) and consequently it is unclear how to address the attrition problem. If ignorable, reweighting the respondent sample on the

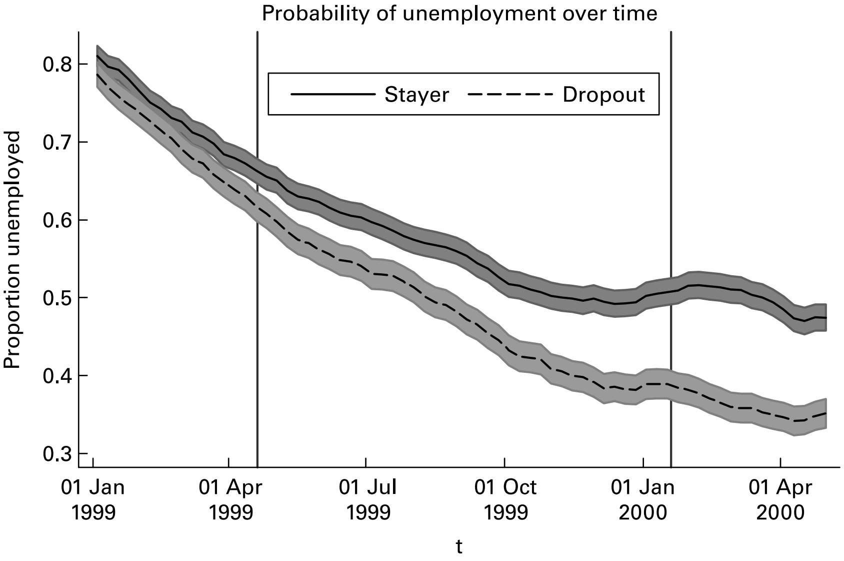

Fig. 1. Difference between stayers and dropouts in the probability of being unemployed (shown with 95% confidence intervals)

basis of variables available in the sampling frame can restore representativeness and therefore overcome attrition bias, or multiple imputation methods can be used (Little and Rubin 1989). Alternatively, for the outcome variable of interest, attrition may be nonignorable; that is, the attrition process may be correlated with the outcome. In a labour market context, this can be exemplified by those changing job being more likely to drop out of the survey (due to lack of time, moving house or some other reason). Hausman and Wise (1979) (HW) developed a two-step procedure for the nonignorable case which is in the spirit of Heckman's selection model (Heckman 1979). This involves estimating the probability of dropping-out and using the result to construct an adjustment term that can be included as a regressor in the outcome equation to correct for the selected nature of the resulting sample. This requires an assumption about the form of the joint distribution of the errors in the selection and outcome equations. Credible implementations also require a suitable instrument (a variable that influences attrition but not outcomes). This can be a major obstacle in practice.

In those empirical analyses that acknowledge the problem of attrition, the approach used to tackle it is typically based on an untestable assumption about the process generating the missing values. For an outcome of interest that is available in register data, it is possible to test whether or not attrition is ignorable. It seems plausible that related outcomes are affected by attrition in similar ways. For example, if attrition is nonignorable in the case of NDYP when considering unemployment, it is likely to also be nonignorable when considering employment or economic inactivity. Consequently, the test may have broader relevance.

Hirano, Imbens, Ridder, and Rubin (2001) (HIRR) consider attrition in the case of panel data covering two periods when a refreshment sample is available. Let Y_{it} be a vector of all time-varying variables for individual i at time t and let X_i be a vector of all fixed variables for individual i. In the first time period, a random sample of size N_P from a fixed population is drawn. This is the *panel*. For each individual in the panel, X_i and Y_{i1} are observed. A subset of size N_{BP} of this sample does not drop out of the second stage survey. This is the *balanced panel* (BP), and for these individuals Y_{i2} is observed. The remaining $N_{IP} = N_P - N_{BP}$ comprise the *incomplete panel* (IP). Those in the IP dropout of the survey in the second stage and Y_{i2} is therefore not observed. In addition to the panel data set, in the second period a new random sample from the original population is drawn. This is the refreshment sample (size N_R).

HIRR do not consider the issue of initial nonresponse. The focus is on those individuals who are interviewed in the first time period. Some will not respond when approached a second time. Let W_i be an indicator of whether an individual is willing to respond at the second wave; $W_i = 1$ represents those who respond at the second wave and $W_i = 0$ represents those who do not. HIRR show that knowing the joint distribution of (Y_1, Y_2, X) or possibly the conditional distribution of (Y_1, Y_2) given X allows the MAR and HW models of attrition to be tested. Using a refreshment sample, these distributions can be identified. This is also possible when register data provide the outcome variable of interest.

With this notation, the MAR model implies $W \coprod Y_2 | Y_1, X$, where $\coprod$ denotes independence. In this scenario, first-period variables can be used to explain dropout and the attrition problem in the sample can be overcome by reweighting the balanced panel by the inverse of the attrition probability or by multiple imputation. The assumption underlying the HW model is $W \coprod Y_1 | Y_2, X$. This implies that the probability of attrition depends on contemporaneous variables but not on first period variables and is known as selection on unobservables (Fitzgerald et al. 1998) since attrition depends partly on variables that are not observed when the individual drops out.

HIRR show that the restrictions implied by the MAR and HW models allow the resulting marginal distribution of the second-period outcome to be tested against that of the refreshment sample. With a binary outcome variable, Y_{it}, and denoting the conditional probability $Pr(Y_{i2} = 1 | Y_{i1} = y, W_i = w, X = x)$ by q_{yw} and the probability $Pr(Y_{i1} = y, W_i = w | X = x)$ by r_{yw}, the marginal distribution of Y_{i2} conditional on x can be calculated as $Pr(Y_{i2} = 1 | X = x) = q_{00}r_{00} + q_{01}r_{01} + q_{10}r_{10} + q_{11}r_{11}$. The terms q_{00} and q_{10} cannot be observed but can be retrieved for the MAR and HW models due to the linear restrictions these models imply. Specifically, MAR implies $q_{00} = q_{01}$ and $q_{10} = q_{11}$. HW implies $q_{00} = \{r_{10}r_{01}(1 - q_{01}) - r_{11}r_{00}(1 - q_{11})\}/\{r_{00}r_{11}q_{11}(1 - q_{01})/q_{01} - r_{11}r_{00}(1 - q_{11})\}$ and $q_{10} = \{q_{00}r_{00}q_{11}r_{11}\}/\{q_{01}r_{01}r_{10}\}$.

To apply the test, Table 1 shows bootstrapped estimates of r_{00}, r_{01}, q_{01}, r_{10}, r_{11} and q_{11} based on a random 50 per cent of the sample of Wave 1 respondents. The estimates of q_{00} and q_{10} implied by the MAR and HW models are also shown. Using these estimates, the probability of Stage 2 unemployment can be estimated. Under the MAR model, this is estimated at 0.492, while the HW estimate is 0.418. The other 50 per cent of the sample of Wave 1 responders is used in place of the HIRR refreshment sample and gives the "true" probability as revealed by the register information: 0.450. Bootstrapping the difference between the true and the model-based estimates results in differences of -0.042 and 0.032 for MAR and HW, respectively. The associated 95 per cent confidence intervals for these differences suggest that the HW model better characterises the data; whereas the MAR

Table 1. Estimated parameters for HIRR attrition model test

	Test sample		
	MAR	HW	Non-test sample $\Pr(y_2 = 1)$
$\hat{r}_{00} =$	0.162	0.162	
	(0.007)	(0.007)	
$\hat{r}_{01} =$	0.182	0.182	
	(0.007)	(0.007)	
$\hat{r}_{10} =$	0.266	0.266	
	(0.008)	(0.008)	
$\hat{r}_{11} =$	0.390	0.390	
	(0.009)	(0.009)	
$\hat{q}_{01} =$	0.256	0.256	
	(0.018)	(0.018)	
$\hat{q}_{11} =$	0.615	0.615	
	(0.014)	(0.014)	
$\hat{q}_{00} =$	0.256	0.132	
	(0.018)	(0.033)	
$\hat{q}_{10} =$	0.615	0.414	
	(0.014)	(0.073)	
$\Pr(y_2 = 1)$	0.492	0.418	0.450
	(0.012)	(0.027)	(0.009)
Difference from true $\Pr(y_2 = 1)$	−0.042	0.032	
	(0.015)	(0.028)	
95% CI	[−0.071, −0.012]	[−0.021, 0.091]	

The parameters reported in this table are described in the text. The variable y_2 is unemployment at the time of the Wave 2 interview, so $\Pr(y_2 = 1)$ is the average level of unemployment at that time. Bootstrapped standard errors in parentheses (1,000 replications). Confidence intervals were calculated using bias-corrected percentiles of the bootstrap distribution. The directly estimable probabilities were estimated using probit models controlling for the following characteristics observed in the administrative data: age; gender; partnership status; disability status; number of JSA claims at New Deal entry; whether living in a rural area; region of residence and type of area (a series of dummy variables combining information on whether in a rural, rural/urban or urban area and whether in an area of high unemployment or an area with a tight labour market).

confidence interval lies entirely below zero, the HW confidence interval spans zero. This is partly explained by the larger variance of the HW test. Note, though, that rejection of MAR alone is sufficient to imply nonignorability. Were both MAR and HW rejected, this would suggest the probability of attrition depended on both Y1 and Y2.

3. The Approach to Choosing Survey Substitutes

With nonignorable attrition, the common practice of reweighting the balanced panel using frame variables may not be appropriate. Assuming a suitable instrument can be found, the HW approach may be appropriate although this must be altered for each specific application and can be cumbersome in practice. Furthermore, estimating more complicated models under HW can be prohibitively difficult.

Using survey substitutes can avoid these complications. As discussed in the introduction, it will often be desirable to find substitutes who are as similar as possible to those they replace. This can be achieved by purposive sampling or by sampling within strata defined according to the characteristics of interest. While the latter option has the advantage that it still remains a probability sample, both approaches may be difficult in practice when there are numerous characteristics on which similarity to dropouts is judged. An alternative approach, explored in this article, is to identify survey substitutes using propensity score matching (see Dolton 2002 for a closely related empirical application).

Propensity score matching originates from Rosenbaum and Rubin (1983). Theorem 1 of Rosenbaum and Rubin (1983) is that the propensity score – the conditional probability of receiving treatment, $pr(d = 1|x)$ – is a balancing score. A balancing score, $b(x)$, is a function of observed covariates, x, such that the conditional distribution of x given $b(x)$ is the same for the treated ($d = 1$) and the nontreated ($d = 0$):

$$x\coprod d|b(x)$$

In the current application, dropping out of the sample is the "treatment" and the fact that the propensity score is a balancing score means that if a group of survey substitutes can be identified who are similar to the dropouts in terms of their propensity score, they should also be similar in terms of their underlying x. The important advantage of this is that the dimensionality of the match can be reduced to one; rather than matching on a vector of characteristics, it is possible to match on just the propensity score.

Operationally, the process of identifying the substitutes is as follows:

1. estimate a model of Wave 2 response
2. generate a propensity score for each individual as the probability of Wave 2 response
3. pool the survey dropouts and the potential substitutes into a single dataset
4. initialise the matching weight of all potential substitutes to 0
5. identify the closest match for each dropout:
    (a) choose one dropout, i
    (b) find the potential substitute, j, with the closest value of the propensity score to that of i
    (c) increment the matching weight of j by 1
    (d) choose the next dropout, $i = i + 1$
    (e) return to (a) until all dropouts have been matched.

This procedure results in the identification of individuals who can be used as substitutes for the dropouts. If retrospective data can be collected, this means that, for some outcomes, subsequent analysis can be based on a complete dataset. It is clear from the algorithm that a single individual may substitute for multiple dropouts. This must be accounted for when carrying out further analysis. The extent to which this process can overcome attrition bias can be assessed by comparing the mean outcome of the dropouts with the (weighted) mean outcome of their substitutes. We also consider a variant on this approach which prevents any one individual substituting for more than one dropout.

4. The Empirical Strategy for Testing the Approach

Since no pools of potential substitutes were available with the NDYP data, they were artificially constructed. The starting point is the original population drawn from administrative records and including all NDYP entrants, regardless of whether they responded to the first or second interview. This was randomly divided into two equally-sized subsamples. The basic idea is to address the attrition among Wave 1 respondents in one of these subsamples (referred to as the "test" sample in the remainder of this article), using the other sample (the "nontest" sample) as the source of potential substitutes. This is related to the "supplemental samples" discussed in Kish (1965).

The test sample was used to estimate the probability of Wave 2 response based on those variables available in the sampling frame. From the estimated coefficients, the propensity score – the probability of Wave 2 response – was generated for individuals in either the test sample or the nontest sample. The propensity scores were then used to match test sample dropouts to individuals in the nontest sample, thereby identifying the survey substitutes.

As discussed in the Introduction, we examine the effectiveness of drawing substitutes from a number of specifically defined subgroups within the overall nontest sample. To demonstrate whether the approach could work under the most favourable conditions imaginable, a pool of potential substitutes made up of only those in the nontest sample who responded at Wave 1 but not at Wave 2 was considered first. Such individuals should, in principle, be statistically equivalent to the dropouts. They are referred to in the remainder of this article as the "ideal substitute pool" (ISP). They can be regarded as providing a theoretical benchmark. The fact that individuals in the ISP did not respond at Wave 2 means that they are of no practical relevance.

The other sources of potential substitutes that are considered attempt to mimic the kind of substitutes that may be available in practice. We begin by considering the BP. Taking substitutes from the BP amounts to a reweighting of the BP and therefore is suited to the case where attrition is MAR. As seen from the HIRR test, it appears that attrition in the NDYP sample is nonignorable, so reweighting in this way is unlikely to address the problem unless additional information can be introduced when constructing the weights. We return to this point in the next section.

The third source of potential substitutes is designed to resemble the dropouts more closely. To achieve this requires an understanding of the process generating nonresponse. Attrition has been shown to be nonignorable when controlling for background characteristics available in the administrative data defined at the time of sampling. Table 2 provides strong evidence in support of circumstances at the time of the interview being a key "unobservable" determining survey response. The first column of results relates to a probit model of Wave 1 response, estimated on the full sample. The key point to note is the highly significant coefficient for contemporaneous unemployment; being unemployed substantially increases the probability of responding at Wave 1. The other columns show the results of using administrative data to estimate a probit model of Wave 2 response among those who responded at Wave 1. Note that estimating Wave 1 and Wave 2 response simultaneously gave a correlation term between the two equations that was not statistically significant. In view of this, Wave 2 response is estimated for Wave 1 respondents only.

Table 2. Modelling survey response

	(1) Response at Wave 1	(2) Response at Wave 2, excl. unemployment transitions	(3) Response at Wave 2, incl. unemployment transitions
Age: 18–19	0.206	0.212	0.179
	(5.10)**	(3.71)**	(3.11)**
Age: 20–21	0.175	0.094	0.089
	(4.31)**	(1.64)	(1.52)
Age: 22–23	0.072	0.136	0.139
	(1.74)	(2.28)*	(2.31)*
Female	0.098	0.064	0.108
	(3.53)**	(1.71)	(2.83)**
Disabled	0.110	0.160	0.147
	(2.95)**	(3.23)**	(2.95)**
Number of JSA claims since January 1995	−0.039	−0.019	−0.046
	(7.25)**	(2.59)**	(5.83)**
Rural area	0.231	0.235	0.257
	(2.41)*	(1.96)*	(2.13)*
Region: northern	−0.125	−0.145	−0.162
	(2.06)*	(1.83)	(2.03)*
Region: north-west	−0.134	−0.156	−0.142
	(2.45)*	(2.17)*	(1.96)*
Region: Yorkshire & Humberside	−0.072	−0.178	−0.175
	(1.24)	(2.35)*	(2.31)*
Region: Wales	−0.154	−0.480	−0.471
	(2.10)*	(5.03)**	(4.90)**
Region: west Midlands	−0.093	−0.162	−0.144
	(1.45)	(1.91)	(1.69)
Region: east Midlands and eastern	−0.149	−0.177	−0.181
	(2.47)*	(2.25)*	(2.29)*
Region: south-west	−0.033	−0.317	−0.325
	(0.32)	(2.37)*	(2.41)*
Region: London and south-east	−0.426	−0.539	−0.536
	(8.48)**	(8.02)**	(7.92)**
Area: rural, high unemployment	0.353	0.569	0.577
	(5.94)**	(7.14)**	(7.19)**
Area: rural/urban, tight labour market	0.197	0.175	0.216
	(4.10)**	(2.63)**	(3.21)**
Area: rural/urban, high unemployment	0.295	0.308	0.366
	(5.84)**	(4.57)**	(5.37)**
Area: urban, tight labour market	0.128	0.121	0.149
	(3.00)**	(2.05)*	(2.51)*
Area: urban, high unemployment	0.174	0.167	0.207
	(4.61)**	(3.21)**	(3.95)**
Unemployed Wave 1	0.337		
	(12.14)**		

Table 2. Continued

	(1) Response at Wave 1	(2) Response at Wave 2, excl. unemployment transitions	(3) Response at Wave 2, incl. unemployment transitions
Transition: unemployed Wave 1, not unemployed Wave 2			−0.406 (10.32)**
Transition: not unemployed Wave 1, unemployed Wave 2			−0.008 (0.10)
Transition: not unemployed Wave 1, not unemployed Wave 2			−0.454 (9.10)**
Constant	−0.106 (1.64)	0.167 (1.91)	0.479 (5.18)**
Observations	11,059	5,915	5,915
Log likelihood	−7361.5183 $\chi^2_{21} = 555.31$	−3922.8855 $\chi^2_{20} = 255.98$	−3851.6421 $\chi^2_{23} = 398.47$

Absolute value of z-statistics in parentheses. *significant at 5% level; **significant at 1% level. Base for categorical variables: age – 24–25; region – Scotland; area– inner city, high unemployment; transition: unemployed Wave 1, unemployed Wave 2.

Two parameterisations of the model were estimated. These differ in that the second specification includes variables capturing unemployment transitions between the times of Wave 1 and Wave 2 fieldwork, while these variables are absent from the first specification. The results show those characteristics most strongly associated with responding to the Wave 2 interview. The unemployment variables are highly significant, their omission from the specification having the test statistic $\chi^2_3 = 142.49$. The probability of responding to the survey was substantially lower where the individual was not unemployed at the time for the interview. While speculative, this finding might be explained by employed individuals having less time to participate, being more difficult to contact or perceiving the survey as being of less relevance to them than to unemployed individuals.

In view of this finding, the "purposive substitute pool" (PSP) was designed to disproportionately represent those in work at the time for the Wave 2 interview. In reality, this cannot be observed in the register data until some months after the time for the Wave 2 interview. To simulate a practical application, the PSP was defined as being made up of those individuals in the nontest sample whom the administrative data showed not to be unemployed in the week beginning 16 August 1999 but who were predicted by the model shown in Table 2 to respond at Wave 2. The date of 16 August 1999 was chosen as being about six months before the Wave 2 interviews took place and as therefore allowing for the delay before unemployment status can be observed in the administrative data (about four months) and a shorter period (1–2 months) between drawing the sample and beginning the fieldwork. In other words, were identified substitutes to be interviewed at Wave 2, the latest available administrative unemployment information at the time of drawing the sample would roughly relate to mid-August 1999.

In addition, we consider how performance might alter were it possible to improve response rates among the PSP. There are numerous ways in which such an increase could be achieved. As noted in Lepkowski and Couper (2002), response to second (and later) stages of panel surveys requires that individuals be located, that contact can be made with located individuals and that contacted individuals cooperate. The probability of locating individuals can be increased by devoting more effort to tracing address changes. The probability of achieving contact can be increased by making more call-backs or by varying the times at which interviewers call. Finally, the probability of cooperation can be increased by changing the mode of interview or by introducing financial incentives to participate (Singer 2002). Consistent with the findings in this analysis, Lynn et al. (2002) show that hard-to-get respondents in the UK Family Resources Survey (i.e., those in the case of whom six or more visits are needed to achieve an interview or those who are persuaded to cooperate after initially refusing) are more likely to be in work than other respondents.

We illustrate the potential effect of this by expanding the PSP to include an additional (randomly selected) 5 per cent of the PSP nonrespondents. We refer to this as the "boosted PSP" (BPSP). In a similar spirit, we examine the effect of increasing the Wave 2 response rate among the Wave 1 responders. This is illustrated by constructing a pool of potential substitutes made up of the BP plus 5 per cent of the dropouts. We refer to this as the "boosted BP" (BBP) – note that simulating an increase in the Wave 2 response rate among the Wave 1 responders both increases the size of the pool of potential substitutes and reduces the number of dropouts. Finally, we consider pooling each of the ISP, PSP, and BPSP with the BP to see whether this can improve performance.

5. Results

The key results in this section are based on bootstrapped estimates. Bootstrapping was necessary in order to avoid the possibility of the results being specific to the particular test sample and nontest sample. That is, since the definition of these samples contains a random element (as noted in the previous section), it may be that this drives the results. To address this, the entire process set out in Section 3 was bootstrapped.

Within each bootstrap, the results of the probit estimation were used to generate propensity scores for all individuals as the fitted probability of Wave 2 nonresponse. These propensity scores were then used for matching by pooling the dropouts from the test sample (the IP) with a particular pool of potential substitutes and, as outlined in Section 3, matching each dropout to that potential substitute with the most similar propensity score. Eight different pools of potential substitutes were tried: ISP; BP; PSP; BPSP; ISP and BP combined; PSP and BP combined; BPSP and BP combined; and BBP.

Table 3 shows the results when matching with replacement. In other words, the results in Table 3 allow individuals in the pool of potential substitutes to be associated with more than one dropout. The number of dropouts matched to a single substitute is shown by the matching weight. Ideally, all matching weights would be 1 since this minimises the variance of any subsequent estimates. However, this is unlikely to be achieved in practice since particular types of individual may be rare among the potential substitutes but relatively common among the dropouts so that a single substitute represents the closest

Table 3. Differences in unemployment status between dropouts and matched substitutes, matching with replacement

			Covariate imbalance			Unemployed, 26 Jun 2000			Unemployed, 22 Feb 1999		
Source of substitutes:	Conc. (%)	Max wt (mean)	Before	After	(b − a)/b	Mean	S.E.	95% C.I.	Mean	S.E.	95% C.I.
Excluding Wave 1 → Wave 2 unemployment transitions from propensity score model											
ISP	24.6	16.6	3.2	3.0	1.5	0.0	2.7	[−4.4, 4.5]	−0.3	2.4	[−4.2, 3.7]
BP	27.3	18.3	7.6	3.5	53.4	−11.0	3.1	[−15.9, −6.0]	−2.6	2.8	[−6.8, 2.0]
PSP	44.0	218.0	18.3	13.3	27.8	−9.9	9.7	[−23.8, 4.7]	13.9	12.7	[−3.8, 32.8]
BPSP	34.5	35.4	16.5	6.1	63.0	0.4	3.9	[−6.0, 6.6]	17.0	5.3	[8.9, 26.5]
ISP & BP	24.8	16.1	4.8	2.3	51.0	−4.6	2.6	[−8.7, −0.6]	−1.4	2.4	[−5.3, 2.6]
PSP & BP	26.4	18.2	11.5	3.5	69.7	−8.5	2.9	[−12.7, −3.7]	2.2	3.4	[−3.6, 7.8]
BPSP & BP	26.0	17.6	10.9	3.4	68.3	−6.7	2.9	[−11.2, −1.9]	3.9	3.2	[−1.7, 8.9]
BBP	25.8	17.1	7.3	3.5	52.3	−10.3	3.2	[−15.3, −5.3]	−2.5	2.8	[−7.0, 2.1]
Including Wave 1 → Wave 2 unemployment transitions in propensity score model											
ISP	22.3	11.3	3.2	3.3	−6.7	0.1	2.1	[−3.4, 3.6]	−0.7	1.7	[−3.5, 2.3]
BP	23.2	15.3	7.6	3.6	52.2	−2.7	2.4	[−6.8, 1.3]	−1.1	1.9	[−4.1, 2.3]
PSP	55.8	469.0	18.3	30.5	−66.8	8.8	15.8	[−24.4, 20.7]	18.8	20.4	[−6.6, 44.3]
BPSP	39.0	34.2	16.5	8.2	50.2	11.0	3.3	[5.2, 15.9]	18.8	4.6	[11.7, 26.3]
ISP & BP	22.3	10.3	4.8	2.7	43.8	−1.0	1.9	[−4.2, 2.0]	−1.2	1.3	[−3.3, 0.9]
PSP & BP	25.8	15.3	11.5	3.9	65.8	1.5	2.3	[−2.3, 5.3]	3.4	2.1	[−0.6, 6.7]
BPSP & BP	24.0	13.8	10.9	3.9	64.0	1.9	2.2	[−1.7, 5.4]	4.3	2.1	[0.8, 8.0]
BBP	23.8	14.0	7.3	3.7	49.5	−2.6	2.5	[−6.7, 1.4]	−1.2	2.0	[−4.3, 2.0]

Note: All results based on 500 bootstrap replications. The column headed "Conc." gives the concentration ratio – the percentage of the dropouts matched to those substitutes in the highest decile of matching weights. The measures of covariate imbalance show the mean standardised difference between dropouts and their substitutes for the variables included in the estimation of the propensity score (other than the variables showing unemployment transitions, which are excluded to allow comparability across the top and bottom panels of the table). As such, these measures show the degree of balance across the two groups – the differences are expressed as a percentage of standard error. For each variable, the absolute difference in means across the two groups is divided by the square root of the average of the two associated variances and multiplied by 100. Averaging across all variables yields the entry in each cell. Results are shown before matching (that is, the comparison between the dropouts and all those in the given pool of potential substitutes) and after matching (the comparison between the dropouts and their matched substitutes). The percentage improvement in the balance is shown in the third column.

match for a number of dropouts. We later consider matching without replacement (which prevents matching weights greater than 1).

We begin by considering the upper panel of Table 3, which corresponds to the case where the model used to estimate the propensity score does not include transitions in unemployment status (see Column 2 of Table 2). The first two columns summarise the matching weights for the identified substitute samples. The column labelled "Conc." gives the concentration ratio; the number of dropouts accounted for by the decile of substitutes with the largest weights. A low figure here is preferable, indicating that there is not excessive reliance on a small number of substitutes. The ISP provides something of a benchmark against which to judge the degree of reliance. In that case, roughly a quarter of the dropouts are matched to those substitutes in the highest decile of matching weights. This is not dissimilar to the case for BP and therefore for the combination of BP with any of the other sources of substitutes. The PSP especially but the BPSP also performs relatively poorly, with concentration ratios of 44.0 and 34.5, respectively. This impression is reinforced when considering the largest weight. The average for the PSP greatly exceeds that for any other source of substitutes, but the BPSP average is also high relative to the ISP. The next three columns provide further diagnostic evidence on the performance of the match by summarising the degree of covariate imbalance between the dropouts and the substitutes (see footnote to Table 3 for a definition of these measures). This provides an insight into how similar the selected substitutes are to the dropouts, at least in terms of observable characteristics. The "after" column shows the extent to which the identified substitutes resemble the dropouts in terms of their observed characteristics. On this basis too, the PSP and BPSP perform less well than the other sources of potential substitutes. Overall, these diagnostics caution against use of the PSP and BPSP.

The remainder of the results in the upper panel show the extent to which matched substitutes can overcome the attrition bias in the data. This is done by comparing the mean outcome of the dropouts with the mean outcome of their substitutes. The size and significance of the difference indicates how successfully the survey substitutes capture the outcomes of the dropouts they replace. A significant difference would suggest that the outcomes of the dropouts have not been captured by the substitutes. The outcomes considered are unemployment status at two points in time: the week commencing 26 June 2000 (roughly four months after the Wave 2 interviews were carried out) and the week commencing 22 February 1999 (roughly the time of the Wave 1 interview), taken from administrative records. A negative value in Table 3 indicates that unemployment is lower among the dropouts than it is among those selected to replace them.

As noted in the introduction, since the outcome is taken from administrative records, it is observed for dropouts and non-dropouts alike so the question of dropout biasing estimates of the proportion unemployed does not arise directly in this specific application. However, the analysis is informative because it simulates a situation where comparable administrative data are not available and unemployment outcomes are measured by a survey. This is a common situation and gives the current investigation broader relevance.

The results show that taking substitutes from the ISP gives good estimates of the unemployment status of dropouts at both points in time. While of little practical relevance, this finding shows the potential for the approach in a theoretical application. Turning to a more realistic case, substitutes taken from the BP substantially overestimate unemployment at the later time point. This is an unsurprising result in view of the rejection of the MAR assumption presented earlier – identifying substitutes from the BP is, after all, equivalent to reweighting the BP which relies on the MAR assumption. Earlier unemployment is better captured, possibly reflecting the fact that contemporaneous unemployment affects Wave 1 response similarly for the dropouts and the BP (all respond at Wave 1). The estimated bias for the PSP is marginally smaller than for the BP when it comes to the later outcome but the standard error is much larger, reflecting the high PSP matching weights. It is surprising that those in the PSP are found to be more likely than the dropouts to be unemployed (the same direction of bias as found for the BP). However, the poor performance of the match for the PSP (and BPSP) indicates that one should not read too much into this finding. For the Wave 1 outcome, the PSP performs poorly. Interestingly, the BPSP performs rather well for the later outcome. However, the Wave 1 outcome is very badly predicted by the substitutes identified from the BPSP.

Combining the ISP with the BP gives results that are worse than those when using the ISP alone but better than when using the BP alone. This suggests that, even when the ideal pool of potential substitutes is available, pooling this with the BP as a source of potential substitutes may have negative consequences. On the other hand, when the potential substitutes available are less than ideal, combining them with the BP may bring benefits. The pooled PSP/BP results are better than for either PSP or BP alone. Combining the BPSP with the BP results in increased bias for the later outcome than was achieved under BPSP alone, but reduced bias at Wave 1. Finally, the BBP results are slight improvements on the BP results.

It is worth highlighting at this stage that the approach used so far to identify matched substitutes does not use any information other than that available at the time the original sample was drawn. Since the survey data are linked to register data, there is clearly the opportunity to include outcome information after the time of sampling in the variable set used to construct the propensity score. Such a model is shown in Column 3 of Table 2. Three variables indicating the unemployment transition between the time of the Wave 1 and Wave 2 surveys are included. The purpose of doing this is to balance individual transitions of dropouts with those of their substitutes so that it becomes more legitimate to consider trends over time. These results show that, relative to those individuals who were unemployed at both Wave 1 and Wave 2, those who moved out of unemployment between the two waves and those who were not unemployed at either wave were significantly less likely to respond at Wave 2.

The bottom panel of Table 3 corresponds to the matched substitutes identified using this version of the propensity score. The diagnostics of the match are qualitatively similar to those already discussed, highlighting the relatively poor performance of the PSP and BPSP. In terms of how well the matched substitutes capture the unemployment status of dropouts, we see again the strong performance of the ISP and the weak performance of the PSP and BPSP. More interesting are the differences from the results in the top panel of Table 3. Including unemployment transitions in the propensity score estimation dramatically improves the performance of the BP; for both the later outcome and the Wave 1 outcome, estimated bias does not differ significantly from zero. Again, the BBP performs marginally better still. Pooling the BP with the PSP achieves even better results for the later outcome. The (absolute) difference in unemployment levels between the dropouts and their matched substitutes is smaller, more precisely estimated and not statistically significant. The Wave 1 outcome is not as well captured as it is with the BP alone, but still is not statistically significant. The levels of (absolute) bias achieved when pooling the BP with the BPSP are higher than when using the PSP/BP sample, but are still not statistically significant.

Given the large weights attached to substitutes drawn from the PSP and the BPSP, it is of interest to consider performance when we do not allow substitutes to be matched to multiple dropouts. Generally, we would expect this to have three consequences. First, where there are fewer potential substitutes than there are dropouts, not all dropouts will be matched to a substitute. Second, matching without replacement is more demanding of the data in that, with each successive match, the stock of remaining potential matches is reduced. Hence, match quality is likely to deteriorate. Third, since matching weights greater than one are avoided, resulting standard errors tend to be smaller than when matching with replacement.

Table 4 shows the results of matching without replacement. The first column shows the percentage of the dropouts for whom substitutes can be identified. Regardless of whether the propensity score includes the unemployment transition variables (i.e., in both the upper and the lower panels), in most cases the majority of dropouts are matched. The PSP and (to a lesser extent) the BPSP perform noticeably less well than the other sources of potential substitutes, providing matches for 81.5 and 87.3 per cent of dropouts, respectively. The covariate imbalance diagnostics also point to an inferior performance for the PSP and BPSP and again suggest caution when using these as the sole source of potential substitutes.

The results in the bottom panel are mostly more positive than those in the upper panel, so we concentrate on the lower panel here. There are two main points of interest. First, compared to the results when matching with replacement is performed (Table 3), the results using the BP are less encouraging. In particular, the matched substitutes from the BP are significantly more likely than the dropouts to be unemployed. Second, compared to matching with replacement, the results when pooling the PSP (or the BPSP) with the BP are better for the later outcome but worse for the Wave 1 outcome.

Finally, Table 5 shows how well unemployment in the week commencing 26 June 2000 among the dropouts is captured by their matched substitutes once sampling weights addressing Wave 1 nonresponse are taken into account. These are inverse probability weights generated by the model presented in the first column of Table 2. While this is somewhat distinct from the main issue considered in this article – that of replacing dropouts – initial nonresponse can also cause biased inference, so it is of interest to see how introducing sampling weights to take account of Wave 1 nonresponse affects performance. This is done by giving to each substitute the sampling weight of the dropout it is replacing. Where matched to multiple dropouts, the weight given to the substitute is the sum of the sampling weights for all associated dropouts. The results are shown both including and excluding unemployment transitions from the estimation of the propensity score and implementing matching both with and without replacement.

The overall impression from Table 5 is one of qualitative similarity with the results already presented. The main differences arise when including unemployment transitions in the propensity score model and matching without replacement. The results for the PSP and BPSP are noticeably worse than when sampling weights are not incorporated. However,

Table 4. Differences in unemployment status outcome between dropouts and matched substitutes, matching without replacement

		Covariate imbalance			Unemployed, 26 Jun 2000			Unemployed, 22 Feb 1999		
Source of substitutes:	Matched (%)	Before	After	(b − a)/b	Mean	S.E.	95% C.I.	Mean	S.E.	95% C.I.
Excluding Wave 1 → Wave 2 unemployment transitions from propensity score model										
ISP	97.9	3.2	3.4	−9.7	−0.1	1.9	[−3.0, 2.9]	0.1	1.7	[−2.5, 3.0]
BP	99.5	7.6	3.8	49.7	−11.4	1.9	[−14.4, −8.4]	−3.3	1.8	[−6.1, −0.2]
PSP	81.5	18.3	23.5	−28.1	6.8	2.6	[2.5, 10.9]	19.4	4.5	[11.9, 27.7]
BPSP	87.3	16.5	19.9	−21.1	7.3	2.4	[3.3, 11.2]	19.6	4.0	[12.7, 26.9]
ISP & BP	99.5	4.8	2.2	53.6	−5.4	1.7	[−8.2, −2.5]	−1.5	1.7	[−4.4, 1.1]
PSP & BP	99.5	11.5	4.0	64.9	−8.9	2.0	[−12.1, −5.4]	1.9	2.8	[−2.8, 6.5]
BPSP & BP	99.5	10.9	3.5	68.1	−7.4	1.9	[−10.7, −4.0]	3.2	2.7	[−1.1, 7.6]
BBP	99.5	7.3	3.1	57.4	−10.9	1.9	[−13.8, −7.6]	−3.2	1.8	[−6.2, −0.3]
Including Wave 1 → Wave 2 unemployment transitions in propensity score model										
ISP	97.9	3.2	3.4	−7.4	−0.7	2.0	[−3.8, 2.6]	−0.2	1.8	[−3.1, 2.8]
BP	99.5	7.6	4.0	46.7	−5.6	1.8	[−8.6, −2.6]	−1.1	1.3	[−3.2, 1.1]
PSP	81.5	18.3	22.2	−20.8	1.4	3.9	[−5.1, 7.6]	16.5	5.4	[7.3, 26.2]
BPSP	87.3	16.5	19.1	−15.9	3.6	3.6	[−2.7, 9.2]	17.5	4.7	[9.7, 25.9]
ISP & BP	99.5	4.8	2.3	52.3	−1.4	1.5	[−3.9, 1.3]	−0.7	1.1	[−2.5, 1.2]
PSP & BP	99.5	11.5	5.4	52.3	1.1	1.7	[−1.7, 4.1]	5.2	1.9	[1.8, 8.3]
BPSP & BP	99.5	10.9	4.5	58.7	1.6	1.7	[−0.9, 4.5]	5.9	1.6	[3.2, 8.4]
BBP	99.5	7.3	3.3	55.0	−4.2	1.7	[−7.1, −1.2]	−0.8	1.1	[−2.5, 1.1]

Note: All results based on 500 bootstrap replications. The column headed "Matched" gives the percentage of the dropouts for whom a match can be found among the pool of potential substitutes. For other columns, see footnote to Table 3.

Table 5. Differences in unemployment between dropouts and matched substitutes in week commencing 26 June 2000, allowing for sampling weights

	Matching with replacement			Matching without replacement		
Source of substitutes:	Mean	S.E.	95% C.I.	Mean	S.E.	95% C.I.
Excluding Wave 1 → Wave 2 unemployment transitions from propensity score model						
ISP	0.3	2.9	[−4.4, 4.9]	0.4	1.9	[−2.5, 3.6]
BP	−10.2	3.4	[−15.7, −4.9]	−10.6	1.9	[−13.6, −7.4]
PSP	−10.7	12.1	[−27.3, 7.2]	5.2	2.6	[0.7, 9.5]
BPSP	1.2	4.3	[−6.1, 8.1]	6.4	2.4	[2.2, 10.4]
ISP & BP	−4.2	2.9	[−8.7, 0.3]	−4.8	1.8	[−7.6, −1.9]
PSP & BP	−8.3	3.2	[−13.4, −2.9]	−8.8	2.0	[−12.1, −5.5]
BPSP & BP	−6.4	3.1	[−11.6, −1.1]	−7.2	2.0	[−10.5, −3.9]
BBP	−9.5	3.4	[−14.9, −4.1]	−10.1	1.9	[−13.2, −6.7]
Including Wave 1 → Wave 2 unemployment transitions in propensity score model						
ISP	1.0	2.2	[−2.5, 4.4]	2.0	1.9	[−0.9, 5.0]
BP	−0.7	2.6	[−5.1, 3.7]	−3.1	1.8	[−6.1, 0.0]
PSP	9.9	17.9	[−27.4, 23.0]	8.0	2.5	[3.8, 12.0]
BPSP	12.5	3.5	[6.3, 17.8]	8.9	2.3	[4.9, 12.7]
ISP & BP	−0.1	2.0	[−3.4, 3.1]	0.0	1.6	[−2.4, 2.7]
PSP & BP	2.8	2.5	[−1.6, 6.9]	2.8	1.8	[−0.1, 5.9]
BPSP & BP	3.2	2.4	[−0.7, 7.0]	3.3	1.7	[0.7, 6.3]
BBP	−0.7	2.7	[−5.0, 3.5]	−1.8	1.8	[−4.7, 1.3]

given the already-noted caution associated with the PSP and BPSP, we should not attach too much significance to this finding. More tellingly, drawing substitutes from the combined PSP and BP yields a bias that is not statistically significant (albeit on the margins of statistical significance), as was the case when not incorporating Wave 1 nonresponse weights.

6. Conclusion

This article uses survey data to examine how sample attrition can bias estimates of an outcome recorded in register data and whether replacing sample dropouts with similar-looking substitutes can overcome this bias. The purpose of doing this is to gain an insight into how effective survey substitutes might be in the more general case where the outcome of interest is not recorded in register data.

A number of conclusions follow from the analysis. First, when register data containing an outcome of interest can be linked to sample frame (and survey) data, it is possible to test different models of attrition for that outcome. If this shows attrition to be ignorable for that outcome, using information available at the time of sampling to reweight those individuals who continue to respond to the survey can address attrition bias. On the other hand, if attrition is shown to be nonignorable for that outcome, reweighting in this way is not appropriate and another approach is needed. Where possible, this test should be carried out routinely before deciding on a strategy for dealing with attrition bias. Clearly, this will only be practicable for those outcomes captured in the register data. However, to the extent that the process generating attrition for one outcome may be similar to that for *related* outcomes, the test may have wider relevance. In the application considered in this article, for example, it seems plausible that if attrition is nonignorable when considering unemployment, it may also be nonignorable when considering other indicators of labour market engagement.

Second, the results have provided evidence on the usefulness of strategies to address the problem of attrition bias. When the ideal pool of potential substitutes is available (individuals who would have responded at Wave 1 but not at Wave 2), the approach works well. This demonstrates the validity of the approach in principle. The results arising from more realistic cases were more mixed. Using the BP as a source of potential substitutes worked poorly when unemployment transitions were excluded from the propensity score model but worked well when transitions were included. This is an informative result since this approach is a form of reweighting. As expected, reweighting based on variables available in the sample frame does not address the nonignorable attrition characterising these data. However, where these weights can be constructed to reflect outcomes since the time of sampling, a reweighting approach can successfully address the problem of nonignorable attrition. Such weights achieve their success by incorporating variables that are additional to those available at the time of drawing the sample and are correlated with the attrition process.

The results of using the PSP as the sole source of potential substitutes were not encouraging. Dropouts were shown to differ from their matched substitutes with regard to both their background characteristics and their unemployment outcomes. However, pooling the PSP with the BP gave good results, but not significantly better than those using BP alone. While there does not seem to be much to recommend the use of a PSP from the point of view of bias reduction, the pooled PSP/BP results had smaller standard errors; including a PSP alongside the BP may therefore be attractive if the precision of estimates is a key concern.

The precision of estimates is further increased by matching without replacement. This prevents any substitute from being used more than once. When matching without replacement, the pooled PSP/BP sample out-performs the BP alone when it comes to later unemployment outcomes but not when it comes to unemployment at the time of the Wave 1 survey.

In summary, the results have shown the biggest gains to arise from taking account of outcome information observed in register information after the time of sampling. Incorporating this additional information substantially improved the extent to which reweighting non-dropouts could address attrition bias. Efficiency gains were achieved by combining this reweighting with substitution of some dropouts with individuals taken from a sample drawn from the original population of individuals thought unlikely to respond to the second wave of interviews (i.e., the pooled PSP/BP approach). Increasing the response rate among the PSP or the BP brings only marginal benefits.

Finally, there are three additional points to highlight about the approach considered in this article. First, while it may be possible to identify substitutes that have histories for a particular outcome similar to those of the dropouts, when it comes to outcomes not recorded in administrative data, histories will only be observable by collecting retrospective information in the survey interview. Consequently, such histories for the substitutes will necessarily rely on respondent recall. This is a limitation that affects the use of survey substitutes but does not affect reweighting approaches.

Second, the results in this article are specific to the dataset considered here and do not necessarily generalise to other datasets with different characteristics, including different attrition processes. It would be informative to repeat the analysis on suitable alternative datasets. However, it should be noted that, while the particular dataset considered here is unique, the use of survey data linked to administrative records is not uncommon with labour market evaluations (see Riccio et al. 2008 for a recent example), so the findings in this article have a broader relevance.

Third, the substitutes identified when addressing attrition with regard to one particular outcome may differ from the substitutes that would be identified when considering an alternative outcome. This cautions against regarding the approach as addressing attrition across multiple outcomes, particularly when these outcomes are not closely related to each other. In principle, the process of identifying substitutes should be carried out anew for the analysis of each separate outcome. However, where the analytic focus is on a set of related outcomes, it may be that identifying a single set of substitutes will suffice.

7. References

Bonjour, D., Dorsett, R., Knight, G., Lissenburgh, S., Mukherjee, A., Payne, J., Range, M., Urwin, P., and White, M. (2001). New Deal for Young People: National Survey of Participants: Stage 2. Employment Service Report, 67, Sheffield.

Bryson, A., Knight, G., and White, M. (2000). New Deal for Young People: National Survey of Participants: Stage 1. Employment Service Report, 44, Sheffield.

Cassel, C.M., Särndal, C.-E., and Wretman, J.H. (1983). Some Uses of Statistical Models in Connection with the Nonresponse Problem. In Incomplete Data in Sample Surveys, Vol. III: Symposium on Incomplete Data, Proceedings, W.G. Madow and I. Olkin (eds.), New York: Academic Press.

Chapman, D. (1983). The Impact of Substitutions on Survey Estimates. Incomplete Data in Sample Surveys, W. Meadow, I. Olkin, and D. Rubin (eds). Vol II. New York: National Academy of Sciences and Academic Press.

Chapman, D. (2003). To Substitute Or Not To Substitute – That Is The Question. Survey Statistician, 48, 32–34.

Chapman, D. and Roman, A. (1985). An Investigation of Substitution for an RDD Survey. Proceedings of the American Statistical Association, Survey Research Methods Section, 269–274.

Dehejia, R. and Wahba, S. (2002). Propensity Score Matching Methods for Nonexperimental Causal Studies. The Review of Economics and Statistics, 84, 151–161.

Dolton, P. (2002). Reducing Attrition Bias using Targeted Refreshment Sampling and Matched Imputation. University of Newcastle Mimeo.

Fitzgerald, J., Gottschalk, P., and Moffitt, R. (1998). An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics. Journal of Human Resources, 33, 251–299.

Hausman, J.A. and Wise, D.A. (1979). Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment. Econometrica, 47, 455–474.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. Econometrica, 47, 153–161.

Hirano, K., Imbens, G., Ridder, G., and Rubin, D. (2001). Combining Panel Data Sets with Attrition and Refreshment Samples. Econometrica, 69, 1645–1659.

Kish, L. (1965). Survey Sampling. New York: Wiley.

Lavori, P.W., Dawson, R., and Shera, D. (1995). A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data. Statistics in Medicine, 14, 1913–1925.

Lepkowski, J. and Couper, M. (2002). Nonresponse in the Second Wave of Longitudinal Household Surveys. Survey Nonresponse, R. Groves, D. Dillman, J. Eltinge, and R. Little (eds). New York: Wiley, 3–26.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139–157.

Little, R. and Rubin, D. (1989). The Analysis of Social Science Data with Missing Values. Sociological Methods and Research, 18, 292–326.

Little, R. and Rubin, D. (2002). Statistical Analysis with Missing Data, (Second Edition). New York: Wiley.

Lynn, D. (2004). The Use of Substitution in Surveys. Survey Statistician, 49, 14–16.

Lynn, P., Clarke, P., Martin, J., and Sturgis, P. (2002). The Effects of Extended Interviewer Efforts on Nonresponse Bias. Survey Nonresponse, R. Groves, D. Dillman, J. Eltinge, and R. Little (eds). New York: Wiley, 135–147.

Riccio, J., Bewley, H., Campbell-Barr, V., Dorsett, R., Hamilton, G., Hoggart, L., Marsh, A., Miller, C., Ray, K., Vegeris, S. (2008). Implementation and Second-year Impacts for Lone Parents in the UK Employment Retention and Advancement (ERA) Demonstration. Department for Work and Pensions Research Report 489.

Rosenbaum, P.R. (1987). Model-based Direct Adjustment. Journal of the American Statistical Association, 82, 387–394.

Rosenbaum, P. and Rubin, D. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika, 70, 41–50.

Rubin, D. (1976). Inference and Missing Data. Biometrika, 63, 581–592.

Rubin, D. and Zanutto, E. (2002). Using Matched Substitutes to Adjust for Nonignorable Nonresponse Through Multiple Imputations. Survey Nonresponse, R. Groves, D. Dillman, J. Eltinge, and R. Little (eds). New York: Wiley, 389–402.

Singer, E. (2002). The Use of Incentives to Reduce Nonresponse in Household Surveys. Survey Nonresponse, R. Groves, D. Dillman, J. Eltinge, and R. Little (eds). New York: Wiley, 135–147.

Van den Berg, G., Lindeboom, M., and Dolton, P. (2006). Survey Nonresponse and the Duration of Unemployment. Journal of the Royal Statistical Society, Series A, 169, 585–604.

Vehovar, V. (1999). Field Substitution and Unit Nonresponse. Journal of Official Statistics, 15, 335–350.

Vehovar, V. (2003). Field Substitutions Redefined. Survey Statistician, 48, 32–34.

Received October 2006
Revised October 2009