

**ESRC Research Methods Programme
Working Paper No 13**

*Linking household survey and administrative record data: what
should the matching variables be?*

Stephen P Jenkins, Peter Lynn, Annette Jäckle and Emanuela Sala

Institute for Social and Economic Research, University of Essex

Revised 14 October 2004

Summary

Linkages of household survey responses with administrative data may be based on unique individual identifiers or on survey respondent characteristics. The benefits gained from using unique identifiers need to be assessed in the light of potential problems such as non-response and measurement error. We report on a study that linked survey responses to UK government agency records on benefits and tax credits in five different ways. One matched on a respondent-supplied National Insurance Number and the other four used different combinations of sex, name, address, and date of birth. As many linkages were made using matches on sex, date of birth, and post-code, or on sex, date of birth, first name and family name, as were made using matches on self-reported National Insurance Number, and the former were also relatively accurate when assessed in terms of false positive and false negative rates. The five independent matching exercises also shed light on the potential returns from hierarchical and pooled matching.

Keywords: record linkage, matching, National Insurance number, measurement error

Acknowledgements

This paper derives from a project on ‘Improving survey measurement of income and employment’ (ISMIE), funded by the ESRC Research Methods Programme (H333250031). We also benefited from ISER’s core funding from the ESRC and the University of Essex. We are grateful to our ISER colleagues, especially Nick Buck, Jon Burton, John Fildes, Heather Laurie, Mike Merrett, and Fran Williams, for their assistance in producing the ISMIE dataset, and to James Banks for providing some tabulations from ELSA. Helpful comments on an earlier version were provided by Lucinda Platt and by participants at a workshop on Data Linkage, 27 September 2004, London. We are also indebted to the Information and Analysis Directorate, DWP Information Centre, especially Catherine Bundy, Katie Dodd and Judith Ridley, for implementing the data linkages. The opinions expressed in this paper are the views of the authors alone.

Address for correspondence

Stephen P. Jenkins, Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK. Email: stephenj@essex.ac.uk

1. Introduction

Although linkage between household survey responses and administrative data records is rare in Britain (Plewis et al. 2001), it is increasingly on the agenda. For example, the English Longitudinal Survey of Ageing is to supplement survey data with information about respondents' National Insurance contributions, benefit and tax credit records held by government agencies, and information from hospital episode statistics and from mortality and cancer registration records. The Office for National Statistics and the Department for Work and Pensions have a pilot project investigating the feasibility of linking administrative record data on benefits to working-age respondents to the Labour Force Survey. The Millennium Cohort Study is to include data obtained from hospital episode statistics and birth registration records, and plans to include school records in later sweeps. In general, record linkage has several attractions for household survey producers and users: it may help diminish respondent burden, additional information may be collected, and measurement error may be reduced. Whether this potential can be fully realised is not yet known, as linkage with household surveys is in its infancy, not only in Britain but also in many other countries. We address one of the important linkage implementation issues in this paper, aiming to help make linkage of administrative records with household survey responses a more straightforward and routine procedure in future.

One question is fundamental for linkage exercises: what variables should be used to implement the link between respondents in the survey and records in the administrative source? We provide answers drawing on a study that linked UK government agency records on benefits and tax credits to household survey respondents in five different and independent ways. We found that as many linkages were made using matches on sex, date of birth, and post-code, or on sex, date of birth, first name and family name, as were made using matches on self-reported National Insurance Number (NINO), and the former were also relatively accurate when assessed in terms of false positive and false negative rates. The independent matching experiment also provided the opportunity to assess the gains from using match criteria hierarchically and by pooling matches.

The advantages of matching using a NINO are clear: a NINO is a unique personal identifier and virtually all adult Britons have one. There are, however, potential disadvantages to NINO-based matches when a NINO is derived from a survey. There is possible *non-response*: respondents may not be willing to provide a NINO or they may not know what their NINO is. There is also potential *measurement error*: respondents may report

NINOs with error or interviewers may transcribe them incorrectly. Similar remarks also apply to national identifiers for other countries: see e.g. Jabine and Scheuren (1986) about the US Social Security Number.

Instead of linking records using a NINO, one could use variables that are collected as part of the survey and which also appear in the administrative database. The advantage of this strategy is that the variables are already available, and there is no additional respondent burden. The disadvantages of the strategy are, first, that there is a chance that a match is not unique (e.g. if done by name and sex). Second, there is potential for mismatch because the survey and the administrative source may record the same type of information differently. This may reflect measurement error in either or both of the sources (e.g. a different spelling of someone's name or address), or different recording conventions. A survey may routinely record a respondent's nickname as forename, but the administrative database may use the legal first name (cf. 'Bill' versus 'William'). In addition, the date to which information refers may not be the same in the two sources. For example, in our study, a survey respondent's address refers to the address at the date of the interview but, in the DWP administrative database, the address refers to the address at the date of the most recent data scan for current benefit recipients, and the date of some earlier scan for former benefit recipients. (A 'scan' is a 100% data extract of all current claims, taken as a snapshot at a particular date.) Hence, residential mobility may lead to mismatch by address. So too might marriage, or divorce, if there is a change in family name recorded in the survey which occurred after the most recent report in the administrative database.

In sum, the choice of matching variables for linkage between survey and administrative data is not clear cut, and requires information about the numbers of matches made using different match criteria, and about their relative accuracy. This paper provides new UK evidence about these issues. We compare NINO-based matching with matching based on four other criteria for linking respondents to a large household survey (the 'Improving survey measurement of income and employment' survey) with administrative records on receipt of benefits and tax credits held by the Information Centre of the Department for Work and Pensions (DWP). The distinctive features of our work are its examination of the relative performance of five independent matching criteria (and of their combination) and the evaluation of NINO-based matching in particular, and analysis of linkages with household surveys rather than linkages between different administrative sources. (Cf. linkages between data from school surveys such as the Annual School Census and Decennial Census small area data (as used by Burgess and Wilson 2004), linkages

between patient and hospital records as in the Oxford Record Linkage Study (Gill 2001), linkages of respondents to decennial Censuses as in the ONS Longitudinal Study, and the administrative data linkages undertaken in the Nordic countries.) Although our analysis is based on a British household survey, the issues that we address are of wider relevance. The match criteria that we use are similar to those that are available in most household surveys in most countries.

Related research includes Brudvig (2003) whose US experimental study concluded that Social Security Numbers – the US equivalent of UK NINOs – reported by survey respondents were accurate (the overall validation rate was 95%). The study did not compare SSN matching with matching on other variables, however. Nor did the few previous UK studies that linked survey and administrative data: see e.g. Noble and Daly (1996) studying Disability Living Allowance claimants and the Department for Work and Pensions (2003) study linking eligible non-recipients of the Minimum Income Guarantee (MIG) who were respondents to the 2000/2001 Family Resources Survey with administrative records on benefits. Note also that both these studies focused on narrowly-defined subgroups of the population, disabled persons and low-income pensioners. Our study uses a more widely-defined population sample.

The ‘Improving survey measurement of income and employment’ (ISMIE) survey and methods of linkage with the DWP data are described in Section 2. In Sections 3 and 4, we compare the linkage rates of the various match criteria, and assess their relative accuracy by comparing the prevalence of false positive and false negative matches. Our investigation of the sources of mismatch and measurement error provides guidance about how to refine matching criteria in linkage exercises using household survey data. An additional dimension on which the match criteria may be compared is the extent to which linked-subsample datasets derived from them differ in their composition. We address this issue in Section 5, using multivariate analysis. Section 6 contains a summary and conclusions.

2. Linkage of data from the ISMIE survey and DWP administrative records

2.1 The ISMIE survey

The household survey data were derived from the ‘Improving survey measurement of income and employment’ (ISMIE) survey, a follow-up to the 2001 wave of BHPS-ECHP panel. This panel was derived from a random sample of private households, the UK component of the

European Community Household Panel Survey (ECHP-UK). This began in 1994, with annual interviews thereafter. Following the major reorganisation in ECHP design in the mid-1990s, a sub-sample was drawn from the ECHP-UK and surveyed jointly with the primary samples of the British Household Panel Survey (BHPS) from 1997 onwards. Households were eligible for selection if all adult members had been interviewed in the previous wave, and one of the following applied: (a) the household reference person was unemployed currently or in the last year; (b) the household reference person was receiving lone parent benefit; (c) the housing was rented; or (d) means-tested benefits were received. These criteria were intended to provide an over-representation of 'low income' households, though the realised sample contained a notable number of households with middle-range income and some with high incomes. See Jäckle et al. (2004) for further details.

Funding for the BHPS-ECHP subsample expired in 2001. This provided an opportunity to interview respondents once more for purely methodological purposes: a validation study based on comparisons of survey responses with administrative records, and an experimental study comparing the effects of dependent and independent interviewing. Funding for the additional interview round and the research was secured through the ESRC Research Methods Programme, and ISMIE fieldwork took place in spring 2003.

Interviews were sought with all BHPS-ECHP panel members who had responded in survey year 2001, i.e. 1,167 individuals aged 16+ in 785 households. Eligible movers were followed to their new address. The achieved sample with complete interviews was 1,033 adults, i.e. 89% of the eligible sample. The ISMIE questionnaire was the same as that given to the main BHPS sample in Autumn 2002, except that some modules were added for the purposes of the methodological work, and some others (e.g. about health) were excluded in order to minimize total respondent burden and to economise on survey costs. For further details of the ISMIE survey, see Jäckle et al. (2004).

At the end of the individual interview, the interviewer read a preamble stating that additional analysis was being undertaken that year especially to assess the quality of data collected in the survey, and then respondents were asked whether they were happy to give us permission to link their answers with the administrative records held by the Department for Work and Pensions and Inland Revenue about their benefits and tax credits (but not about their income tax). Everyone who gave consent was asked to tell the interviewer their National Insurance Number, with respondents requested to consult a payslip or other records such as a pension or benefit book or NINO card. (Whether they did or not was recorded.) The CAPI script checked that the NINO provided was of the correct format (six digits with a two-letter

prefix and a one-letter suffix). Data linkages were sought for all consenting respondents, regardless of whether they had reported receipt of benefits. As a significant minority of respondents had never received benefits, and so were not cases on the DWP database, the maximum possible linkage rate was less than 100%. We return to this issue below. (We also sought consent from employed respondents for linkages with employer data: see Jäckle et al. (2004) for further details.)

2.2 The data linkage: match criteria and the DWP database

Our data linkages were based on matches between consenting ISMIE survey respondents and information held in the DWP's '100% Generalized Matching Service' Primary Data file. The file contains a record for each person who is currently receiving, or has received, any one of 15 benefits. These include Child Benefit, Housing Benefit, Working Families Tax Credit, several types of disability benefit, Income Support, Job Seeker's Allowance and the state retirement pension. See Jäckle et al. (2004, Appendix 3) for further details.

Each record contains personal details derived from information collected when a benefit claim was made. If someone received several benefits simultaneously, the personal details refer to those associated with the 'most reliable' benefit (based on a hierarchy from Job Seekers Allowance, Income Support, through to Child Benefit). Details are updated when new scans of benefit receipt databases indicate that they have changed. For the ISMIE project, the scans of current benefits and benefit histories refer to those made in the week beginning 13 October 2003, and the scans of Housing Benefit and Tax Credit details refer to the period 1999–2003. Information for each recipient about dates of receipt and amounts paid is held by the DWP in separate files, each linked to the Primary File using the individual's NINO as the key. The personal details held in the Primary Data file include NINO, title (Mr, Miss, Ms and Mrs; and hence sex), date of birth (day, month, year), first name, family name, address, and postcode. All of these variables were potentially available from the ISMIE survey too, and were the basis of our linkage experiment.

Five independent matching exercises were used to link consenting ISMIE survey respondents to the DWP Primary Data. The match criteria were characterized by the following sets of variables:

Criterion 1: NINO.

Criterion 2: Sex, date of birth, postcode.

Criterion 3: Sex, date of birth, forename, family name.

Criterion 4: Sex, postcode, forename, family name.

Criterion 5: Sex, forename, family name, address line 1.

Matches by NINO did not use the suffix letter as NINOs are unique without this. UK postcodes have two parts. The first, the ‘outward code’, is one or two letters denoting the Area followed by one or two digits, denoting the District. The second part, the ‘inward code’, is a digit followed by two letters (the Unit). There are 9,473 postal sectors (defined by outward code plus inward code digit) in Britain, with an average of about 2,530 addresses per sector (Lynn and Lievesley, 1991). An example of ‘address line 1’ is ‘12 Errol Street’.

In each of the five exercises, exact (deterministic) matching was used; there was no probabilistic matching (Gill 2001). All variables from the survey were used verbatim apart from the cleaning and formatting already implemented as part of routine panel maintenance and follow-up. The variables in the DWP Primary Data file were also used verbatim, though it should be noted that addresses and postcodes had already been cleaned and processed into a consistent format using QuickAddress Software (QAS™).

Because the five linkage exercises were undertaken independently, we could also combine the results to simulate the effects of using various hierarchical match criteria. We focused on two criteria involving NINOs (criterion 1 followed by criterion 2, and vice versa), and two criteria based on non-NINO matching (criterion 2 followed by criterion 3, and vice versa). The latter two criteria are similar to the criteria used by the Department for Work and Pensions (2003) study. Finally, we also considered the effects of pooling the results of all five linkages.

3. Linkage rates

Before undertaking record linkages for ISMIE respondents, we had to gain informed consent from them. Consent rates were relatively high: see Table 1. About 78% of the sample provided consent, with no differences in the rates for men and women. Respondents aged 50+ were slightly more likely than respondents aged less than 50 to consent (79% compared with 76%). The rates for those aged 50+ are slightly higher than for respondents to wave 1 of the English Longitudinal Survey of Ageing (all of whom are aged 50+), which may reflect differences in the wording of the consent requests, differences in sample composition, or differences in panel conditioning (the ELSA respondents had been interviewed once before; ISMIE respondents had been interviewed up to seven times previously). Some 88.7% of consenting ISMIE respondents supplied a NINO, with little difference in the fraction for men

(87.4%) and women (89.4%). Put another way, 68.8% of the ISMIE sample provided both consent and a NINO. Among respondents aged 50+, the rate was 69.4%, which is appreciably higher than for the corresponding ELSA sample (for whom the rate was 60.4%). For a detailed analysis of ISMIE respondents' consent and NINO supply propensities, see Jenkins et al. (2004).

Table 1
Percentage of ISMIE respondents who consented to DWP data linkage,
and who consented and supplied a NINO

	All who gave consent to data linkage			All who gave consent and supplied a NINO		
	Men	Women	All	Men	Women	All
Aged < 50	75.9	76.3	76.1	65.1	69.5	67.8
Aged 50+	79.3	78.1	78.6	70.5	68.4	69.4
ELSA (aged 50+) *	[76.5]	[74.6]	[75.5]	[63.1]	[58.1]	[60.4]
All	77.6	77.2	77.6	67.8	69.0	68.8

Notes. *N* = 1,033 (429 men, 604 women). *: Numbers in brackets refer to corresponding percentages for respondents to wave 1 of the English Longitudinal Study of Ageing, all of whom are aged 50+.

The type of response to the NINO question is summarized in more detail in Table 2 for respondents who provided consent to DWP data linkage. The main reason stated for not supplying a NINO was that the respondent did not know it, rather than a refusal to provide it. The breakdowns also suggest that NINOs supplied are likely to be reliable. Among respondents who did supply a NINO, just over two-thirds (67.4%) referred to a payslip or other document, and 30.8% supplied the number from memory but were confident that the number was correct. Only 1.8% stated that they were not sure about the NINO supplied. Consultation of documents to check the NINO supplied was markedly higher among respondents aged 50+ (81.2%) than among respondents aged less than 50 (54.3%). This suggests either that older people were less confident in remembering their NINOs or simply that pension books were more readily available than payslips.

Table 2
Type of response to NINO request by consenting ISMIE respondents

	All who gave consent	<i>column percentages</i>		
		All	Aged < 50	Aged 50+
Provided from payslip or other document	59.7	67.4	54.3	81.2
Remembered and respondent certain	27.3	30.8	43.0	17.9
Remembered but respondent not certain	1.6	1.8	2.7	0.9
Not provided: don't know	9.9			
Not provided: refused	1.5			
All	100.0	100.0	100.0	100.0
(N)	(802)	(711)	(365)	(346)

We now turn to examine linkage success rates for the NINO and the other four match criteria. It should be remembered that there are two potential reasons for a linkage not being made. Either the relevant ISMIE respondent had never received one of the benefits or tax credits for which the DWP database has information (a 'true non-match'), or the respondent had received one of the benefits or tax credits but could not be linked using the five match criteria (a 'false non-match'). We estimate that the expected true non-match rate is about one third, because about two-thirds of the ISMIE respondents reported receiving at least one of the relevant benefits or tax credits at the previous interview.

The 'pooled' linkage rate, i.e. counting all matches on at least one criterion, was 57.3%, which is roughly nine percentage points lower than the rate expected if there were no erroneous non-matches. This suggests that there are false non-matches, but it is difficult to assess their prevalence further because there are no comparable matching exercises against which to benchmark the results. The linkage rate for matches between respondents to the US Health and Retirement Study (HRS) and earnings records held by the Social Security Administration (made using Social Security Numbers) was 75% (Olson 1999). However, this rate is not comparable with the overall ISMIE one (or the NINO-based rate discussed below), as the expected true non-match rate is much lower in our study. In the HRS the expected true non-match rate is near zero: virtually all US adults aged 50+ have had some labour earnings during their working life and hence an SSA record. In the Department for Work and Pensions (2003) study that matched low-income pensioner respondents from the Family Resources Survey with DWP records, the expected true non-match rate was also negligible, because virtually all of the respondents would have been receiving retirement pension or a winter fuel

payment (and therefore a case in the DWP records). The actual match rate was 96% (2003, p. 55).

The linkage rates for each of the various independent and hierarchical criteria are shown in Table 3. (These are the raw linkage rates, and potentially include mismatches, which are discussed further below.) Among the independent matching exercises, the greatest linkage rate was for matching based on sex, date of birth and postcode (criterion 2), followed closely by matching based on NINO (criterion 1) and sex, postcode, forename and family name (criterion 3). The rates are 49.7%, 48.2%, and 47.9%, respectively, when expressed as a fraction of the ISMIE sample size (Table 3, column 1), or 64.0%, 62.1%, and 61.7%, when expressed as a fraction of the number of consenting respondents (column 2). Matching by criterion 4, and especially by criterion 5, led to noticeably worse linkage rates, suggesting that date of birth is a particularly important matching variable (in combination with sex) or, alternatively, that address and name data are subject to more variation in how they are recorded. We return to this issue below. Almost three-quarters of all consenting respondents were matched by at least one criterion ('pooled' matching).

The high potential return to using hierarchical matching is shown in the lower panel of Table 3. Employing additional criteria identified a significant number of additional matches, for both NINO-based and non-NINO-based hierarchical matches. In both cases, the linkage rate was only about one percentage point below the rate achieved from pooled matching (56.4% and 56.1% compared with 57.3%).

Columns 3 and 4 of Table 3 show that many of the differences between linkage rates for the NINO-based match and for matches based on sex and date of birth (criteria 2 and 3) arose because NINO-based matches require a NINO to have been supplied. Interestingly, the linkage rates for criteria 2–5 were all lower for respondents who did not supply a NINO than for those who did. This might be indicative of a general tendency to supply lower quality data, or it may be that respondents who receive benefits are more likely to supply a NINO. Among respondents who supplied a NINO, the linkage rate was 70%, which corresponds almost exactly to the fraction of these same respondents who reported receiving at the previous interview one of the benefits for which information is held by the DWP database.

Table 3
Record linkage rates (%) for ISMIE respondents

Criterion and matching variables	ISMIE sample	All who gave consent to data linkage	Supplied NINO	Did not supply NINO
	(1)	(2)	(3)	(4)
<i>Independent matching</i>				
1. NINO	48.2	62.1	70.0	–
2. Sex, date of birth, postcode	49.7	64.0	64.3	61.5
3. Sex, date of birth, forename, family name	47.9	61.7	62.6	55.0
4. Sex, postcode, forename, family name	41.7	53.7	54.4	48.4
5. Sex, forename, family name, address line 1	33.7	43.4	44.3	36.3
<i>Pooled matching: at least one of the above</i>	57.3	73.8	74.5	68.1
<i>Hierarchical matching</i>				
1 followed by 2, or 2 followed by 1	56.4	72.6	74.1	61.5
2 followed by 3, or 3 followed by 2	56.1	72.1	72.7	68.1
<i>N</i>	1033	802	711	91
<i>(as % of all who gave consent)</i>		(100)	(88.7)	(11.3)

Notes. Table includes potential mismatches (see Section 4).

Table 3 might also be interpreted as saying that matching by non-NINO criteria is a potential strategy for record linkage in the future, particularly given that securing a NINO from each survey respondent is a problem. The veracity of this conclusion depends on the accuracy of the various linkages. Before turning to this issue, we consider the overlaps between the sets of respondents for whom linkages were made.

Table 4 lists the combinations of linkage outcomes from the five independent matching exercises. Of the respondents who gave linkage consent, 26% were not linked by any of the five independent criteria, 4% were linked by one criterion, 15% by two criteria, 4% by three criteria, 15% by four criteria, and 36% were linked by all five (the modal linkage combination). The degree of overlap between the respondents identified by even the most successful match criteria is perhaps surprisingly small. For example, 155 respondents (19% of all consenting respondents) were matched either by criterion 1 or by criterion 2, but not by both. Put another way, this highlights again the potential return to hierarchical or pooled matching procedures.

Table 4
Linkage outcomes among consenting ISMIE respondents

Linkage outcomes*	All who gave consent to data linkage		All who gave consent and supplied a NINO	
	Frequency	Percentage	Frequency	Percentage
00000	210	26.2	181	25.5
00100	7	0.9	3	0.4
00101	2	0.3	0	0
01000	16	2.0	4	0.6
01110	20	2.5	7	1.0
01111	49	6.1	18	2.5
10000	11	1.4	11	1.6
10010	1	0.1	1	0.1
10011	1	0.1	1	0.1
10100	47	5.9	47	6.6
10101	10	1.3	10	1.4
11000	68	8.5	68	9.6
11110	74	9.2	74	10.4
11111	286	35.7	286	40.2
All	802	100.0	711	100.0

Notes. * Outcomes for criteria 1, 2, 3, 4, and 5 (in that order), with ‘0’ meaning not matched, and ‘1’ meaning matched. For example ‘10010’ means respondent matched by criteria 1 and 4, but not by 2, 3 or 5. The match criteria are defined in the text and summarised in Table 3. The table includes potential mismatches (see Section 4).

4. Linkage accuracy

The accuracy of linkage by a particular criterion m may be assessed along two dimensions. First, one wants to minimize the proportion of actual matches by m that are erroneous matches. This is the *false positive rate*, calculated for criterion m as the number of mismatches by m divided by the total number of matches by m . Second, one also wishes to minimize the proportion of non-matches by m that are erroneous. This *false negative rate* is calculated for criterion m as the fraction of non-matches by m that were genuine matches according to criteria other than m . For a given number of matches, one match criterion is unambiguously better than another if the first has a lower false positive rate and a lower false negative rate than the second. If this is not the case, unambiguous rankings of match accuracy involve additional judgements about the appropriate trade-off between the risk of missing more matches and the risk of more false positives.

We estimated false positive and false negative rates by pooling information from the five independent matching exercises. For example, for NINO matches, the false positive rate (criterion 1) was derived from information on cases with match patterns of form ‘1xxxx’ in Table 4, and the false negative rate was derived from information on cases with match patterns of form ‘0xxxx’ (where ‘x’ refers to ‘0’ or a ‘1’). Estimates were calculated for criteria 1–3 (but not for criteria 4 and 5 given their relatively low match rates), and for the hierarchical and pooled criteria discussed earlier. When calculating false negative rates, the appropriate treatment of the 210 cases not matched on any criterion (pattern ‘00000’ in Table 4) is a moot point: as explained earlier, many of these respondents were likely to be true non-matches (non-recipients of benefits). We report estimates based on the assumption that all these individuals were non-recipients of benefits. Supposing instead that they were all benefit recipients increased the magnitude of every estimate but did not change the ordering of the criteria by false negative rate.

We assumed that matches made by three or more of the five independent matching criteria were genuine matches (except in one NINO-related situation discussed shortly), and visually inspected listings of information about all remaining cases to assess whether an actual match (or non-match) was true or false. Although this introduced an element of researcher judgement, assessment was almost always clear cut in practice. For example, when the survey and DWP postcodes differed, they usually did so by only one or two characters, and it was clear from the name, address, and birth date information, that the correct person had been identified according to one or more other criteria. Problems with assessment of address information are discussed further below.

The exceptional NINO-related situation was when the matching process led to two different individuals in the DWP Primary Data (with two different NINOs) being associated with a single respondent in the ISMIE survey. This arose with 14 respondents (13 with match pattern ‘11111’ and one with ‘11000’). We could determine that, in eight cases, the NINO from the survey was incorrect and hence there was a mismatch by criterion 1 but a genuine match by other criteria. In three cases, there was a mismatch by criterion 3, and in one case, mismatch by criterion 5.

The estimates of the false positive and false negative linkage rates are shown in Table 5. In several of the table cells, a range has been reported rather than a single estimate. In each of these cases, estimation involved comparisons of address information, and a visual inspection could not resolve with certainty whether there was a genuine match or genuine

mismatch. (Recall from the Introduction that addresses could legitimately differ between the survey and DWP databases because of residential mobility.)

Table 5
Estimates of linkage accuracy

Matching method	False positive rate		False negative rate	
	%	(<i>N</i>)	%	(<i>N</i>)
<i>Independent matching</i>				
1. NINO	2.2 – 11.6	(498)	30.9	(304)
2. Sex, date of birth, postcode	0	(513)	23.9 – 27.3	(289)
3. Sex, date of birth, forename, family name	0 – 10.9	(495)	30.6	(307)
<i>Hierarchical matching</i>				
1 followed by 2	1.9 – 9.9	(583)	4.1	(219)
2 followed by 1	0.5 – 8.6	(583)	4.1	(219)
2 followed by 3	0 – 8.1	(579)	4.7	(213)
3 followed by 2	0 – 9.3	(579)	4.7	(213)
<i>Pooled matching</i>				
Match by at least one of 1–5	0 – 8.6	(592)	0	(210)

Notes. Independent, hierarchical and pooled matching defined in the text. False positive rate for criterion m = percentage of matches by m that were mismatches according to criteria other than m . False negative rate for criterion m = percentage of non-matches by m that were genuine matches according to criteria other than m . Estimates of false negative rates assume that all 210 cases with match pattern ‘00000’ were not benefit recipients (see text). N refers to the number in the denominator of the relevant rate calculation.

The match pattern ‘10100’ – actual matching by NINO and also by sex, date of birth, forename, and family name – illustrates the problems with addresses. The 47 respondents had different ‘address line 1’s in the survey and DWP file. However, inspection revealed that three cases had virtually identical address line 1 and postcode (so the errors probably reflected transcription errors), 23 were in the same postal Area and District (i.e. had the same outward code), 15 were in the same postal Area, and there were six other cases. We are inclined to believe that most of the respondents were correctly identified, since most residential mobility in Britain is short distance. (For example, Böheim and Taylor (2000, Table 1) report that 66.1% of residential moves are within the same local authority district.) Readers sharing our belief should take the estimates of false positive rates as lying towards the lower end of the range shown, and vice versa for the false negative rate.

The lowest false positive rate among the independent matching criteria was for matches by sex, date of birth and postcode (criterion 2): it was a remarkable 0%. The rates for NINO matches and criterion 3 were several percentage points higher (depending on how the

information about addresses is treated). The rate in the former case was at least 2.2%, highlighting the fact that NINOs derived from surveys are subject to measurement error.

NINO measurement error is illustrated by the data for the 32 respondents who supplied a NINO and for whom there was a match on one or more criteria other than the NINO. In 10 cases, the first two letters of the NINO were in error; for example the letters ‘M’ and ‘N’ were swapped in seven cases. In 15 cases, digits were transposed (for example ‘0’ as the first digit rather than the sixth) or apparently transcribed incorrectly (for example ‘8’ rather than ‘5’). In five cases, the six digits of the survey NINO were ‘999999’, suggesting a ‘don’t know’ entry by the interviewer. In four of these cases, the NINO was reportedly derived from a payslip or other document, and in the other case, it was remembered with confidence. Indeed, in only two of the 32 cases was the NINO remembered but the respondent uncertain about it. These examples suggest that the source of NINO measurement error may be with the interviewer rather than with the respondent.

The lowest false negative rates among the independent matching criteria were for matches by sex, date of birth, and postcode: between 23.9% and 27.3%. The rate for matches by sex, date of birth, forename, and family name was 30.6%, which is virtually the same as the rate for NINO matches (30.9%). The rate for NINO matches reflects the fact that a significant number of respondents did not supply a NINO – the problem of non-response cited in the Introduction. If all 62 of these cases had supplied a NINO, and a genuine match had been made using this, then the NINO false negative rate would fall substantially, to 19.2%.

The false negative rate for criterion 3 would have been lower if there had been fewer mismatches on forename and surname. To illustrate the scope for reducing this type of mismatch, consider the respondents with match pattern ‘11000’. Of the 68 cases, 39 non-matches by criterion 3 (and 4 and 5) arose because of differences in forename alone, and half of these appeared to be where the survey recorded a nickname. In seven cases, the forename was spelled differently, often only by one letter (for example ‘Anne’ *versus* ‘Ann’). Sixteen non-matches arose because of differences in family name alone (typically note a simple difference in spelling), and 13 for other reasons, i.e. 43% of the 68 cases. Pre-processing of name data therefore has some potential for improving match accuracy, but this potential is constrained. (For an overview of US Census Bureau software for this and related tasks, see Winkler, 2001.) An alternative, or addition, to pre-processing, would be relax the exact match on name using look-up tables based on common abbreviations or variants (e.g. surname plus initials).

Choice of the best independent match criterion on the basis of linkage accuracy is clear cut, according to Table 5. Criterion 2 – matching by sex, date of birth, and postcode – has both the lowest false positive rate *and* the lowest false negative rate. (It also had the highest raw linkage rate.) Observe that a shift to using hierarchical matching criteria reduced the false positive rate associated with any match criteria involving the NINO (though the change is small). But false positive rates did not fall universally. By contrast, false negative rates for hierarchical matches were clearly smaller than for the independent matches, reflecting a decrease in the number of true non-matches (i.e. a fall in the numerator of the fraction). When matches from the five independent criteria were pooled, there were still some possible false positive cases after our clerical inspections (cases with different addresses). The false negative rate for pooled matching was zero (by assumption).

5. Does the composition of linked-data samples vary by match criterion?

The true probability of having a record in the DWP Primary file varies systematically with differences in respondent characteristics, if only because we would expect families with children to be receiving child benefit, and respondents of pension age to receive a state retirement pension, and, generally, low-income respondents to receive some form of benefit. The relationship between the probability and characteristics is unobserved, but may be estimated from the linked data files. (There are additional factors complicating the estimated relationship such as differential consent propensities.) Although we cannot say that the fitted relationship for any given criterion is close(st) to the true relationship, we can investigate whether each of the various match criteria indicates the same relationship.

Put another way, does the composition of a linked-data sample vary depending on which match criterion has been used to create the linkage? In particular, if one were to rely on NINO-based matching, would some types of respondents be more likely to be found in the linked sample, than if some other criterion were used? Cf. Haider and Solon (1999) and Olson (1999) who investigated whether the sub-sample of HRS respondents for whom linked SSA earnings histories were available was representative of the full HRS sample. Unlike us, they were able to refer to representativeness *per se* because virtually all HRS respondents would have had an SSA record.

We addressed the issues by regressing the probabilities of record linkage on respondent characteristics – sex, age, household type, area of residence, educational qualifications, and log household income. (The last two variables were measured at the

previous interview.) Equations for the probability of linkage by each of criteria 1–3 were estimated jointly, using trivariate probit regression, thereby allowing unrestricted correlations between the cross-equation error terms. Each error variance was normalized to unity. (See Cappellari and Jenkins 2003 for estimation details.) The estimation method also provides a straightforward means of testing whether the impact of a given covariate on the probability of a match differs by criterion. Estimated probit coefficients and their standard errors are shown in Table 6. In the rightmost column are p -values from Wald tests of null hypotheses that each covariate has the same impact on each probability.

The null hypothesis of equal coefficients was unambiguously rejected for only one characteristic, age (p -value = 0.0001). The next smallest p -values were for ‘other’ household type (0.09) and residence in London or South East (0.08); all others were greater than 0.22. The probability of linkage rises slightly less steeply with age for NINO-based matching than for criterion 3 and criterion 2.

For each of the three criteria considered, men were less likely to be matched than women. And compared to single householders, all other household types were more likely to be matched; so too were respondents with educational qualifications below A-level standard. The lower household income was, the higher was the match probability.

The cross-equation error correlations were each about 0.9 and precisely estimated. The strong positive associations between the unobservable factors in each equation, and of similar magnitude, are further evidence of similar sample compositions in the linked data sets derived from the three match criteria.

Overall, the estimates indicate that, as expected, there are systematic associations between linkage probabilities and respondent characteristics – primarily reflecting the differential probabilities of benefit receipt according to these characteristics. Reassuringly, however, the patterns of association are similar for each of the three match criterion, and so the three linked data samples derived from the matching exercises have a similar composition. The different relationship with age is the exception to this, but the cross-criteria differences are relatively small in magnitude.

Table 6
The probability of record linkage by match criteria 1–3 (trivariate probit regression)

Regressors	Criterion 1		Criterion 2		Criterion 3		Wald test of equal coefficients: <i>p</i> -value*
	Coeff.	Robust SE	Coeff.	Robust SE	Coeff.	Robust SE	
Sex: male	−0.293	(0.079)	−0.306	(0.080)	−0.204	(0.080)	0.188
Age (years)	0.015	(0.003)	0.026	(0.003)	0.019	(0.003)	0.000
Household type: couple	0.375	(0.136)	0.480	(0.137)	0.330	(0.136)	0.344
Household type: couple with kid(s)	0.565	(0.153)	0.791	(0.161)	0.648	(0.164)	0.222
Household type: lone parent	0.612	(0.156)	0.800	(0.160)	0.644	(0.161)	0.336
Household type: other	0.609	(0.315)	0.275	(0.286)	0.760	(0.329)	0.092
Lives in London or South East region	0.027	(0.103)	0.179	(0.109)	0.197	(0.103)	0.079
Educational qualification: A-level or more	−0.157	(0.089)	−0.197	(0.089)	−0.222	(0.090)	0.656
Log(household income)	−0.300	(0.082)	−0.273	(0.081)	−0.283	(0.082)	0.895
Constant	1.100	(0.584)	0.290	(0.582)	0.737	(0.588)	0.130
Cross-equation error correlations:							
ρ_{21}			0.886	(0.018)			
ρ_{31}			0.876	(0.019)			
ρ_{32}			0.887	(0.018)			
Log pseudo-likelihood			−1,414.5				

Notes. $N = 1031$. Simulated maximum likelihood estimates, GHK simulator, 50 draws. Standard errors (SE) are adjusted for multiple respondents per household. Reference household type is single. Likelihood ratio test statistic for $H_0: \rho_{21} = \rho_{31} = \rho_{32} = 0$ is 1089.7, p -value = 0.0000. *: Null hypothesis is that the coefficient on the variable of interest is equal in each of the three equations. The match criteria are defined in the text and Table 3.

6. Summary and conclusions

When linking respondents to household surveys with records from administrative databases, the benefits gained from using unique identifiers like the NINO need to be assessed in the light of potential problems such as NINO non-response and measurement error. Other personal variables common to the survey and the administrative data may also be used to create linkages, but they too have potential disadvantages. Not only is there potential measurement error, but some information may differ in the two sources for legitimate reasons. (In our study, names and addresses could refer to different dates.) Whether NINO-

based matching, or matching by some other criterion, leads to higher and more accurate linkage rates is therefore a moot point.

Our study of linkages between ISMIE survey data and DWP benefit and tax credit records using five independent match criteria has highlighted this issue and provided new evidence about the relative merits of different combinations of matching variables. The results suggest that linkages based on sex, date of birth, plus either post-code or first name and family name, yield a raw linkage rate as high as that for NINO-based linkages, and the linkages are relatively accurate. Moreover differences in the composition of the linked-data samples derived using these three criteria are negligible.

Our simulations of hierarchical matching underline the potential rewards to using additional variables for data linkage, whether as a supplement to, or indeed instead of, NINO-based matching. For example, combining a match on sex, birth date and postcode plus either NINO or forename and family name led to a raw linkage rate as high as the pooled linkage rate derived when the results of all the independent matching procedures were pooled. The fact that high linkage rates can be achieved without using NINO matching is useful information for future linkage design strategies, given the additional burdens involved with collecting NINOs.

If future matching exercises do use NINOs nonetheless, then one route to improving linkage success rates would appear to be to raise the proportion of respondents who are willing and able to supply a NINO. However, since almost 90% of ISMIE respondents gave their consent to DWP data linkage (a prerequisite for asking the NINO supply question), the potential for raising the NINO supply rate further may be limited. To reduce false positive rates, NINO measurement error needs to be reduced. Our study has shown that most of the errors appear to have arisen from interviewer transcription error rather than respondent error. The incorporation of more sophisticated checking routines in CAPI scripts, or self-entry by a respondent, are ways to reduce this source of error.

How else might linkages between survey responses and administrative records be improved? Utilisation of software to clean and pre-process name and address data (such as reviewed by Winkler 2001 for the USA) can help reduce inconsistencies between variables in household surveys and administrative record data. Our study underlined the potential for this for name data, but also suggested that its scope is constrained: a significant minority of non-matches (e.g. in surname) arose in ways that would not have been easily caught by cleaning algorithms. Our linkage rate for matches using address line information would have been higher if the QASTM program had been applied to the survey data as well as to the DWP data.

However, since addresses in the two sources may refer to different dates for legitimate reasons, again the application of software algorithms may have only a limited effect. The more that benefit file scans can be coordinated with the timing of the household survey fieldwork, the less that this may be a problem. Observe too that some of the problems described in this paragraph could also be mitigated if survey and administrative sources each contained histories of respondent's names and addresses, rather than a single observation for each.

It may be useful to investigate the relative merits of matching variables other than those used here. For example, the DWP Primary Data also includes telephone numbers for respondents. These numbers may also be routinely collected by survey agencies. There are of course potential problems as well: a significant minority of respondents may not have telephones, or change numbers relatively often (for example when changing mobile phone provider), and they may be subject to measurement error in the same way that NINOs are.

Overall, the positive conclusion of our study is that record linkage between household survey responses and administrative data is feasible, and even relatively simple and cheap matching procedures (as in our study) can yield good results when judged in terms of numbers of matches and their accuracy. To get better results requires some investment in matching technologies. (In addition to the software cited earlier, greater use might also be made of probabilistic matching.) The returns to these investments will be greatest if the investments can be coordinated between the major household surveys, in order to take advantage of generic similarities in information collected that could also be used for matching.

References

Böheim, R. and Taylor, M.P. (2000) From the dark end of the street to the bright side of the road? Investigating the returns to residential mobility in Britain. ISER Working Paper 2000-38, Colchester: University of Essex.

<http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2000-38.pdf>

Brudvig, L. (2003) Analysis of the Social Security Number validation component of the Social Security Number, privacy attitudes, and notification experiment. Final report on third component. Washington DC: US Bureau of the Census Planning, Research and Evaluation Division. http://www.census.gov/pred/www/rpts/SPAN_SSN.pdf

- Burgess, S. and Wilson, D. (2004) Ethnic segregation in England's schools. CASEpaper 79. London School of Economics: Centre for the Analysis of Social Exclusion. <http://sticerd.lse.ac.uk/dps/case/cp/CASEpaper79.pdf>
- Cappellari, L. and Jenkins, S.P. (2003) Multivariate probit regression using simulated maximum likelihood. *The Stata Journal*, **3**(3), 278–294.
- Department for Work and Pensions (2003) *Income-related Benefit Estimates of Take-up in 2000/2001*. London: Department for Work and Pensions. http://www.dwp.gov.uk/asd/income_analysis/tu0001.pdf
- Gill, L. (2001) *Methods for Automatic Record Matching and Linking and their Use in National Statistics*. National Statistics Methodological Series No. 25. London: Office for National Statistics. <http://www.statistics.gov.uk/StatBase/Product.asp?vlnk=9224>
- Haider S. and Solon, G. (1999) Nonrandom selection in the HRS Social Security earnings questions. Unpublished paper. Ann Arbor MI: University of Michigan. <http://www.econ.lsa.umich.edu/~gsolon/workingpapers/nonresp.pdf>
- Jabine, T.B. and Scheuren, F.J. (1986) Record linkages for statistical purposes: methodological issues. *Journal of Official Statistics*, **2**(3), 255–277. <http://www.jos.nu/Contents/issue.asp?vol=2&no=3>
- Jäckle, A., Jenkins, S.P., Lynn, P. and Sala, E. (2004) Validation of survey data on income and employment: the ISMIE experience. ISER Working Paper No. 2004–14. Colchester: University of Essex. <http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2004-14.pdf>
- Jenkins, S.P., Cappellari, L., Lynn, P., Jäckle, A., and Sala, E. (2004) Patterns of consent: evidence from a general household survey. ISER Working Paper 2004-xx, forthcoming. Colchester: University of Essex.
- Lynn, P. and Lievesley, D. (1991) *Drawing General Population Samples in Great Britain*. London: SCPR.
- Marmot, M., Banks, J., Blundell, R., Lessof, C., and Nazroo, J. (2003) *Health, Wealth and Lifestyles of the Older Population in England: the 2002 English Longitudinal Study of Ageing*. London: Institute for Fiscal Studies. <http://www.ifs.org.uk/elsa/report.htm>
- Noble, M. and Daly, M. (1996) The reach of disability benefits: an examination of the Disability Living Allowance. *Journal of Social Welfare and Family Law*, **18**(1), 37–51.

- Olson, J.A. (1999) Linkages with data from Social Security administrative records in the Health and Retirement Study. *Social Security Bulletin*, **62**(2), 73–85.
<http://www.ssa.gov/policy/docs/ssb/v62n2/v62n2p73.pdf>
- Plewis I., Smith G., Wright G. and Cullis, A. (2001) Linking child poverty and child outcomes: exploring data and research strategies. Research Working Paper No 1. London: Department for Work and Pensions. <http://www.dwp.gov.uk/asd/asd5/WP1.pdf>
- Winkler, W.E. (2001) Record linkage software and methods for merging administrative lists. Statistical Research Report Series No. RR2001/03. Washington DC: Bureau of the Census, Statistical Research Division.
<http://www.census.gov/srd/papers/pdf/rr2001-03.pdf>



ESRC Research Methods Programme
Cathie Marsh Centre for Census and
Survey Research
Faculty of Social Sciences and Law
The University of Manchester
Manchester
M13 9PL
United Kingdom

Tel: 0161 275 4891
Email: methods@man.ac.uk
Web: <http://www.ccsr.ac.uk/methods/>

Director: Professor Angela Dale
Administrator: Ruth Durrell