Working Paper

Evaluating approaches to Family Resources Survey data linking

by Stephen McKay

DWP Department for Work and Pensions

Department for Work and Pensions

Working paper No 110

Evaluating approaches to Family Resources Survey data linking

Stephen McKay

A report of research carried out by the University of Birmingham on behalf of the Department for Work and Pensions

This document/publication is also available on our website at:
http://research.dwp.gov.uk/asd/asd5/rrs-index.asp

Any enquiries regarding this document/publication should be sent to us at:
Central Analysis Division, Department for Work and Pensions, Upper Ground Floor, Steel City House,
West Street, Sheffield, S1 2GQ

First published 2012.

ISBN 978 1 908523 63 1

Views expressed in this report are not necessarily those of the Department for Work and Pensions or any other Government Department.

Contents

List of tables

List of figures

Acknowledgements

This project involved collaboration with a wide range of people. At various times the project was led by Mehdi Hussain, Valerie Christian, Graham Jones and Simon Lunn with Ryszard Zaluski-Zaluczkowski. I also benefited from discussions with Ian Herring, Peter Matejic and Julie Sullivan. The research built on previous research conducted by Philip Clarke (ONS).

The Author

Stephen McKay is Chair in Social Research at the University of Birmingham. He conducts research on social security and living standards, and on quantitative research methods in general. He was responsible for all aspects of this study. He joined the University of Birmingham in 2007, and has worked extensively on research projects for government departments, research councils and charitable funding bodies.

Abbreviations and glossary of terms

False recipient

A Family Resource Survey (FRS) respondent who does not receive a particular benefit (according to administrative data), but who reports that they receive it. For instance a non-recipient of Disabled Living Allowance (DLA) reports receiving it when they do not – perhaps because they receive an alternative benefit such as Employment Support Allowance (ESA).

Heckman selection model

A method of correcting for selection bias. This approach provides a test of selection bias, and a means of estimating a model corrected for such a bias.

Hot-deck imputation

This imputation process identifies characteristics within a record containing the missing value to be imputed and matches them up to another record with similar characteristics. The known variable is then copied across to the missing case.

Hidden recipient

An FRS respondent who receives a particular benefit (according to administrative data), but who does not report that they receive it. For instance a recipient of DLA and Income Support (IS) might incorrectly report receiving either one of these benefits, or neither, rather than both.

Inverse Mills ratio

In statistical modelling it is the ratio of the probability density function – which is a defining characteristic of a distribution - divided by the cumulative distribution function. It is most commonly encountered in Heckman selection models, where it is based on the predicted values from the probit (or selection) equation. It provides the 'selection hazard' for the main linear regression equation.

Logistic regression

The statistical modelling technique of regression with a binary dependent variable. This is part of the family known as generalised linear models, which fits data to a logistic curve. It is also known as a logit model.

Probit

A function associated with the normal distribution and has applications in logistic regression models.

Non-response

Those with data not available through non-participation in the survey. May be divided into:

- **unit or total non-response:** data is absent for all values;
- **item or partial non-response:** there is missing data on some relevant questions.

Root Mean Squared Error (RMSE) RMSE is a simple measure of differences between two sets of numbers. Across multiple pairs of data, a lower RMSE suggests a pair is ‘closer’.

Summary

This report provides an investigation into evaluating approaches to linking Family Resources Survey (FRS) with Department for Work and Pensions (DWP) administrative data. The FRS has two important functions which are to support the production of key statistics such as Households Below Average Income (HBAI), and as a dataset that is available for use by both Government policy analysts and external researchers.

Measurement error exists on a survey when respondents do not report their true status. The purpose of data linking examined here is to use administrative data to correct measurement error. It is possible to do this with the FRS for various types of income such as benefit income. Improving accuracy is important because the FRS forms the basis of key National Statistics including HBAI. HBAI is used to measure child poverty which the Child Poverty Act, 2010, requires the Government to report to Parliament on progress towards reduction targets for 2020, as well as rates of low income for other groups and overall inequality.

We need to ask permission from respondents to link their data. An important challenge is how to treat the approximately 50 per cent of respondents not linked because they either refused consent to link their data, or a match was not achieved. This report provides a narrative, and further analysis, of various imputation approaches to adjust for measurement error in the 50 per cent of unlinked cases. The report also assesses the extent of consent bias, and methods of dealing with any resulting bias.

Any long-term solution for imputation must be coherent and address the needs of National Statistics and general users.

Methods

The methods used were as follows:

- A review of existing statistical techniques for investigating consent bias.
- Descriptive characteristics of consenters and non-consenters.
- A logistic regression model of consenters and non-consenters.
- Comparisons of FRS reported receipt and amount of benefits, against administrative data.
- A Heckman selection model to determine consent bias in amounts of benefit reported.
- Analysis of the dispersion of newly-constructed grossing weights based on adjusting consenters only to sum to population totals (rather than the existing weights which adjust both consenters and non-consenters to population totals).
- A 'Hot-deck' style of imputation used to impute amounts of benefit for the unlinked.
- Five models were compared against the baseline administrative figures: an original FRS model was included along with four alternatives based on different imputations, re-weighting or sample selections.
- Estimates of the average amount of benefit received for selected benefits for each of the five models.
- The 'Admin + survey + imputed' option included a further logistic regression model which assigned respondents to a hidden receipt group.

- A logistic regression model for benefit receipt by age and gender was developed to assess each of the five models against the correct population model. Root Mean Squared Error comparisons for model regression coefficients were used to compare models.

Findings

- Models based on imputing the level of benefits for non-consenters generally provided good results. Further imputation of hidden recipients also moves results in the right direction.
- In most models, in most settings, using the administrative data in some form provides more accurate results than simply using the standard FRS.
- Using administrative plus survey data for appropriate respondents tends to be better than using only data from consenters, and does not compare too badly against more sophisticated approaches.
- The level of 'consent bias' was relatively low, although this varied by benefit.
- Heckman selection models indicated that reported amounts of benefits are subject to selection bias for some benefits.

Conclusions/recommendations

- A 'complete cases' approach based on only linked respondents would provide a useful set of companion results to current analyses.
- The simpler approach of just using administrative plus survey data also seems to be a very practicable approach that generally outperforms a complete cases approach and should be investigated further.
- Further consideration needs to be given to the findings of this report, and any implications following the introduction of Universal Credit.
- Consideration needs to be given to an imputed general purpose dataset. This is likely to achieve significantly wider use if it becomes possible to link to Real Time Information earnings data from Her Majesty's Revenue and Customs (HMRC).

Limitations

- The results that apply to this particular set of benefits may not apply to other benefits, nor indeed to other sources of income and especially earnings.
- The application of Heckman models has limitations. There is a tendency for the assessment to change from 'no bias' to 'bias' as more variables are added to the selection model.
- The range of logistic regression models that could have been developed for Table 4.2 was severely limited by the availability of data on administrative sources, to act as the correct population model. A predictive model of benefit receipt by age and sex was developed. However, we cannot ignore the possibility that a different set of variables and models might have led to different conclusions about the best model.
- Results in Section 4.2 include logistic regression style imputation of hidden receipt of benefits. Results do not include an equivalent adjustment for the less common case of 'false recipients' whereby respondents incorrectly report that they are not receiving benefit.

1 Introduction

1.1 Background

The FRS is used both within and outside of Government to produce figures on the number of people on different levels of income. This includes estimates of those below particular income lines, such as those used to measure child poverty. It is, therefore, critical that the survey measures income as accurately as possible. For many low income families a high proportion of their income will be in the form of social security benefits.

However, for many years, it has been clear that the number of people reporting receipt of benefits is lower than would be expected from the total amount spent on different benefits. The FRS annual report for 2009-2010 shows the extent of undercounting (Clay *et al.* 2011: Table M.6 p.116). Income Support (IS) was under-reported by 31 per cent, similar to Pension Credit with 32 per cent under-reporting. Conversely, there was only a four per cent under-representation of state Retirement Pension. Attendance Allowance (AA) was under-reported by 39 per cent and Carer's Allowance (CA) by 25 per cent.

To improve the accuracy of income data, individuals responding to the FRS are asked for their consent to link their information, provided during the survey, with data on income streams of various kinds. This includes data on benefit receipt, child support (from DWPs administrative data), earnings and tax credits. This permission is asked for at the end of the FRS interview. Only those participating fully in the survey are asked for consent to link to administrative data. Those whose data is collected 'by proxy' from another household member are not asked for consent to link data. In addition, the consent question was not asked of those living in Northern Ireland until the 2010-11 survey (i.e. in the data collection fieldwork from April 2010).

The overall breakdown of weighted responses in 2009/2010 was:

- 52 per cent providing consent to data linking;
- 30 per cent declining to provide consent;
- 18 per cent interviewed by proxy and, therefore, not asked for consent.

This may be alternatively characterised as a consent rate of 52 per cent (of all cases with data) or of 63 per cent among those directly asked for consent (52/82).

Within the linked survey-administrative data there are various degrees of misreporting. In some cases there is receipt that is not reported in the survey (hidden recipients), while for others, a person says they receive a benefit when the administrative data indicates that they do not (false recipients). By using the administrative data it may be possible to reduce or perhaps eliminate the extent of undercounting benefit receipt. However, since it is possible to retain receipt of benefits while temporarily in non-household settings that are not included in the FRS sample (e.g. hospital, or abroad) it seems likely that a small undercount should be expected.[1]

1.2 Overview of approaches

This analysis is based on the FRS for 2009-2010, linked to data on benefits from the Work and Pensions Longitudinal Survey (WPLS). This has been the subject of previous work within the department. The analysis does not include other potential sources of bias such as Tax Credits and earnings. Tax Credits were not available during the research phase of the project. Earnings are not collected and recorded consistently across administrative and survey sources and offers further challenges in terms of imputation processes, and dataset dissemination. The key problem is to consider different ways of including non-consenting cases in the analysis.

Four options were identified, as follows.

1 To continue using unlinked FRS data – a kind of null hypothesis against which to consider the other possibilities.

2 To replace survey data with linked data (among the consenters), and to leave the remaining unlinked respondents data (post-edited) unchanged. This may be characterised as a form of data editing, where the best available information for each case is taken, even if such information is not available consistently across all cases.

3 To replace survey data with linked data for consenters, and not to use the unlinked data. This may be regarded as taking a 'complete cases' approach to missing data. There are further options for the weighting of such data, either:

   - Using the existing grossing weights.
   - Or, with the linked data re-grossed, to help deal with any clear biases between consenters and non-consenters.

4 To replace the survey data with linked data (among the consenters), and to adjust the unlinked data as necessary (based on inferences from the linked data). In the context of distributing relevant micro-data, this may be seen as a form of imputing. There are various ways in which such imputation may be implemented – single and multiple imputation, and different algorithms exist to impute (including hot-deck approaches and Heckman style selection models).

The key differences lie in the treatment of the unlinked cases, which are affected in different ways. First, the administrative data can simply be ignored with reliance placed on the data provided in the initial survey. It is collected consistently for all cases – although it is known there are flaws of various kinds.

Second, these unlinked cases can be kept by retaining their survey values for benefit receipt. This approach may at first seem inconsistent – replacing survey values with admin values for about

[1] Another more direct way to tackle the undercount, at least for analysis purposes, would be to re-weight the data to ensure that the right number of benefit recipients is represented in the weighted sample. At present weighting of government surveys tends to be conducted using demographic variables, and particularly gender, age group and region. Weighting by benefit receipt tends to gloss over current knowledge that a sizeable proportion of reported recipients are not actually recipients of those benefits. However if the prime purpose of analysis is the costing of changes to benefits then this kind of weighting may be the most appropriate and convenient method to use.

half the sample, but retaining the others. It is strongly suspected there is misreporting among the non-consenters in an aggregate sense, but that does not mean it is straightforward to identify where such errors occur at the individual level. This approach is tantamount to making the best use of the data on individual cases that are available, albeit that information is different for different respondents depending on their linkage status.

Third, the non-consenters can simply be thrown away leaving just the linked (or consenting) cases – sometimes known as 'complete cases' or list-wise deletion. Those remaining cases may then be re-weighted if desired. Either the main grossing weight can continue to be used for the FRS sample (using a simple uplift to maintain anticipated population number) or the calibration weights can instead be recalibrated to ensure the sample conforms to several different population totals.

The fourth option is to keep the non-consenters in the analysis, but rather than retaining their survey values seek to instead use new 'imputed' values. Imputation of data is used where there is some degree of missing data for particular questions ('item non-response') rather than the complete record for that person being missing ('unit non-response'). Where only a few 'items' or questions are missing and there is no wish to simply drop such cases, one may choose to impute the missing responses. Imputation helps to maintain sample sizes, compared to other methods – which should provide greater efficiency than methods that simply delete cases with missing data.

1.3 Theory

According to Lumley (2010: 186) *'Multiple imputation and survey re-weighting are sometimes described as 'statistically principled' approaches to inference with missing data'*. In survey research it is more usual to use weighting rather than imputation for complete non response ('unit non-response'), and imputation for 'item non-response' where data on particular questions are missing.

The example of data linking, where it is usually not possible to link all cases, could be said to fall into either camp. This situation may be considered either analogous to unit non-response (treating admin data like another wave of data collection – where weighting is generally used) or to item non-response (there is missing data on only a few variables – use imputation).

These are all micro-level adjustments, adapting the dataset in various ways for later use. It is worth noting that for any given application, analysts may not need to adjust data at the unit-level (for each person). Instead they may make aggregate adjustments based on information from the linked data. For instance, assuming that the proportion of 'hidden recipients' is the same in the unlinked data as in the linked data is akin to option 3. This happens with take-up statistics for Pension Credit (Barton and Riley 2012, 7.3, p. 153). This embodies the strong, but not rebutted, assumption that the non-consenters share the same pattern of hidden receipt as the consenters. However, these adjustments are somewhat ad hoc, and do not provide a means of distributing a general purpose dataset that would meet a range of user needs.

1.4 Other relevant research

Existing research has tended to focus on the extent of consent bias, and the reasons for it, rather than methods of dealing with any resulting bias – though see reference below to research by Landy. An increasing number of surveys ask for consent to link data. Therefore, there may possibly be greater interest in research that considers consent bias and its effects. The new ESRC-funded UK Household Longitudinal Survey, *Understanding Society*, requested consent separately for linking to health data and to education data in its first wave (first part, with around 14,000 households). The particularly useful feature of this study was the wide range of data available, including attitudinal data as well as ‘harder’ data on benefit receipt and incomes. There was little difference in rates of consent, which were around 70 per cent, for either administrative source being linked to – respondents tended to either consent or not. Generally speaking the same kinds of association with linkage are found as for FRS – older people, London-based, and minority ethnic groups all show lower consent rates. That was not always the case, for example, with medical-based data, where different associations with age have been found in past studies (Kho *et al.* 2009). However, there were some associations between patterns of consent and attitudes towards trust and privacy. Since the FRS does not currently have attitudinal data of this kind, the material deprivation questions being perhaps the closest, our analysis must rely on the ‘harder’ data on social and economic standing.

A recent research project in Germany was able to compare linked and unlinked cases, to a degree, irrespective of consent. This was for a German labour force and social security survey. Permission was granted to identify on the administrative data those who had not consented to data linking. The initial sample was also drawn from administrative data, so that a number of key comparisons were possible. These comparisons included respondents and non-respondents, the effects of refusal to questions, and the level of consent bias. As a result this study provided an ideal opportunity to compare the administrative data characteristics of consenters and non-consenters.

Two papers are currently available – a general one (in English, unpublished, and only available in draft form at the time of writing, by Sakshaug and Kreuter), and a more statistical one (Krug 2009 – note written in German). The key conclusion of Sakshaug and Kreuter is that consent biases are (a) small in themselves, and (b) much smaller than the biases introduced by both non-response and by measurement error. From their particular sample it was possible to quantify each of these different kinds of bias, at least for key variables that were available on the administrative data. They concluded that there are certainly gains to be made from collecting consent to link data, and using linked data. They were also able to determine that survey answers given by consenters were more accurate than non-consenters.

It is, of course, not possible to replicate this finding on the British data where no administrative data is available for non-consenters. If found to be true more widely, this could represent an important way in which consenters differ from non-consenters – their survey responses tend to be closer to what is recorded on administrative systems. In particular, a standard imputation model ought to be capturing the range of the error term, making it clear that estimates for non-consenters are subject to rather more doubt than a model based on consenters.

Krug (2009) went further and considered what kinds of approaches to analysing survey data are the most reliable. A regression model was established using administrative data, which is taken to be a clear benchmark against which to compare other regressions models based on different approaches to missing data. Models are then run on the survey data (for linked cases, and imputed unlinked cases). The methods used were sample selection models, multiple imputation and complete cases only. The main conclusion was that ‘All missing data techniques under analysis show only small deviations from the benchmark’ – in other words, the model based on the population could be approximated quite well with data from the survey, adjusted in various ways (re-weighting and imputation). Four different scenarios were looked at in terms of missing information (in terms of dependent and independent variables).

No particular approach was always the best. The best approach varied with the scenario (the pattern of ‘missingness’ on both dependent and independent variables). This may mean that the appropriate approach to recommend will depend on the particular context, and that sometimes re-weighting will be superior to imputation, and the reverse. This is an important point in looking at ways forward – no one approach can be guaranteed to do best on all occasions.

The overall approach – of comparing results of the different approaches with a true regression based on population data – is one we try to follow in this analysis in Chapter 4.

In a simulation-based study, Landy (2012) concluded that imputation approaches are likely to do better than Heckman selection models. This was stated because imputation models tend to improve as more variables are added, while selection models tend to have greater problems converging as the relevant models become more complex. There is an assumption that the errors between the selection equation and the Ordinary Least Squares (OLS)-corrected equation are jointly normal, at least in the standard specification of such models.

2 Bias between consenters and non-consenters

2.1 Introduction

In this section some of the differences between consenters and non-consenters are analysed. First basic statistics are provided on each group, and then attempts to model these differences in a multivariate setting by using logistic regression are examined. Different amounts of benefit reported, comparing administrative and survey data for the linked consenters are also examined.

This section was designed to consider whether there are any very clear biases between consenters and non-consenters that need to be taken into account – for instance by weighting those groups with the lowest consent rates.

2.2 Descriptive characteristics of consenters and non-consenters

Initially different profiles of consenters were compared with (a) those declining to provide consent and (b) all non-consenters, including the proxy cases. Results for a range of demographic variables are shown in Table 2.1. There was a high degree of similarity between the consenting cases and the overall sample profile.

The non-consenters (excluding proxies) tended to vary slightly from the consenters – they were less likely to report a limiting long-term illness (30 per cent compared with 34 per cent) and more likely to be from a non-white background (18 per cent compared with 11 per cent). Profiles by age group and gender were very similar, with only slight differences by marital status. Consenters were slightly more likely to be in one-adult households than non-consenters.

When the comparison was broadened to include proxy respondents among the non-consenters – the final column of Table 2.1 – then some of the differences became stronger. In particular, while 24 per cent of consenters lived in one-adult households, this was only true of 22 per cent of explicit non-consenters, and (by definition) virtually none of the proxies. Conversely 21 per cent of the proxy cases were living in households with at least four adults, compared with only eight per cent of the consenting group and nine per cent of the explicit non-consenters. This appeared to be reflecting a higher rate of proxy interviews in larger households – and perhaps the lack of proxies (by definition) in one adult households. There were also higher rates of proxy interviews for 18-29 year olds, and singles.

Table 2.1 Demographic profiles of sample, consenters, non-consenters and proxies

Column percentages

Characteristics	All	Consenters	Non-consenters	Proxies
Men	46	46	46	60
Women	54	54	54	40
Aged 18-29	17	18	16	34
30-39	17	17	17	18
40-49	19	19	19	20
50-59	16	16	16	15
60-69	15	15	15	9
70-79	10	10	10	3
80+	6	6	7	2
Married (civil partner)	53	51	54	54
Cohabiting	12	12	11	14
Single	21	19	19	30
Widowed	7	8	8	1
Separated	2	3	3	*
Divorced (dissolved civil partnership)	6	6	6	1
1 adult in household	19	24	22	*
2	56	56	56	56
3	15	13	13	23
4+	10	8	9	21
White	85	89	82	83
Not white	15	11	18	17
Limiting illness	30	34	30	30
Others	70	66	70	70
Unweighted base	*40,249*	*21,610*	*12,002*	*6,637*
Weighted base	*42,562*	*22,029*	*12,806*	*7,726*

Results by housing tenure and region are in Table 2.2. Among the clearest associations in terms of region:

- ten per cent of consenters lived in London, compared with 16 per cent of those where consent was not provided.
- ten per cent of consenters lived in Scotland, compared with eight per cent of non-consenters.

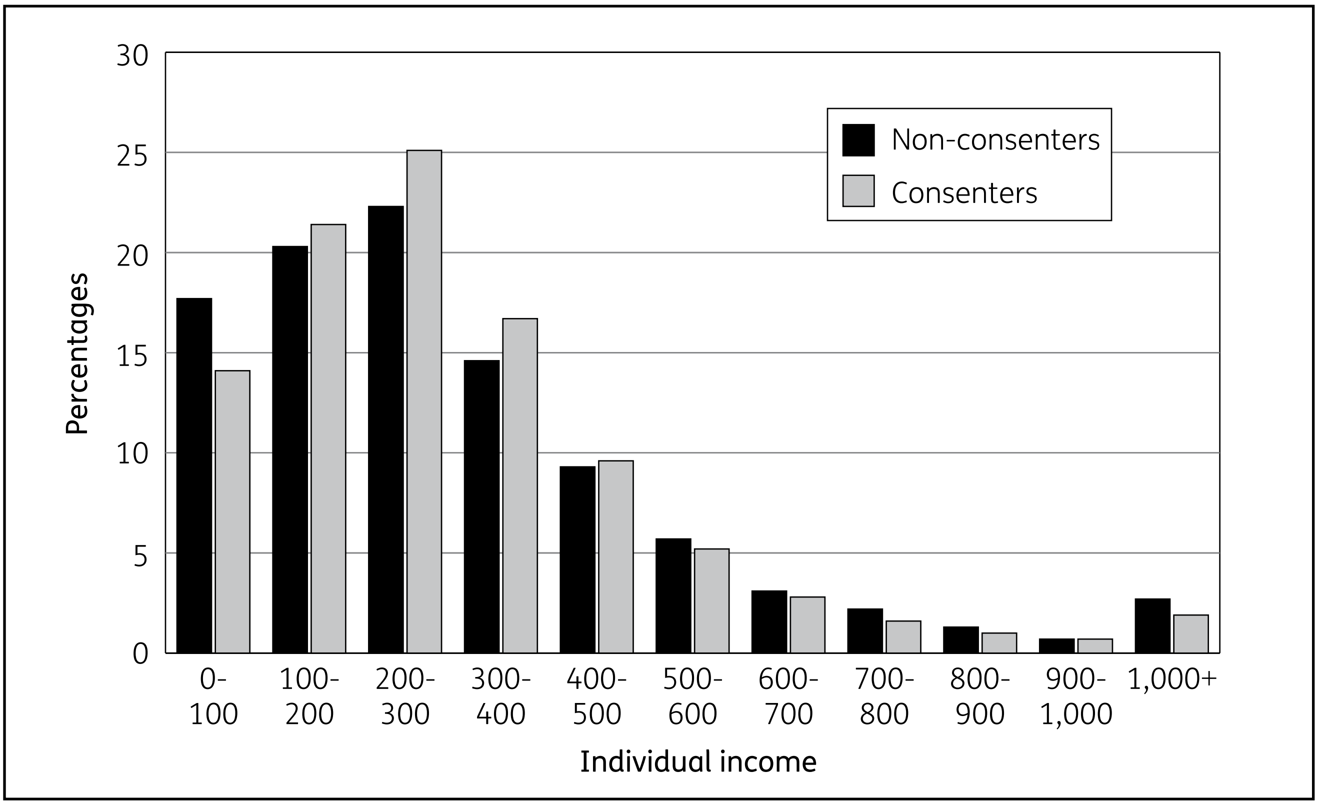
When analysing by housing tenure, Housing Association, Local Authority, and private tenants (unfurnished only) were more likely to consent than not consent. Home owners who owned their property outright were less likely to consent.

Table 2.2 Location and tenure profiles of sample, consenters, non-consenters and proxies

				Column percentages
Region and housing tenure	**All**	**Consenters**	**Non-consenters**	**Proxies**
North East	4	5	3	4
North West	11	12	10	13
Yorkshire	9	10	7	8
East Midlands	8	8	7	8
West Midlands	9	8	11	17
East	10	10	9	9
London	13	10	16	14
South East	14	14	14	14
South West	9	9	9	9
Wales	5	5	5	6
Scotland	9	10	8	7
LA tenant	7	8	6	6
HA tenant	7	9	6	5
Rent – unfurnished	10	11	9	9
Rent – furnished	4	4	4	3
Mortgage	39	37	36	48
Own outright	33	31	38	29
Rent-free	1	1	1	1
Unweighted base	*40,249*	*21,610*	*12,002*	*6,637*
Weighted base	*42,562*	*22,029*	*12,806*	*7,726*

There were only limited differences in people's individual incomes by consent status. In Figure 2.1, there are a few differences in the incomes of consenters and non-consenters, with more non-consenters on the lowest and the highest incomes.

Figure 2.1 Incomes (individual net) of consenters and non-consenters



2.3 A model of consenters and non-consenters

Formerly a number of bi-variate associations between consent and background variables were examined. However, there may well be overlaps between the variables analysed – for instance, younger people being more likely to live in London, and to live in rented accommodation – so the most important associations may be difficult to establish. The following logistic regression models are used to try and identify those variables with statistically significant effects, having controlled for a range of other variables.

Two outcomes are considered – consent versus active non-consent, and consent compared with proxies and active non-consenters. The former relates to the particular survey question asked, while the latter may have greater consequences for any analysis based on this dataset. Results are shown in Table 2.3.

The effect of age differed, to some extent, in the two models. Consent was slightly less likely to be given among those aged 80+ (controlling for the other variables in the model) but otherwise there was no effect by age. Broadening the comparison group to include proxies, there was also a lower rate of consent among younger people, and especially those aged 18-29 years.

Compared with the South-West, consent was less likely to be achieved in London and the West Midlands. However, rates of consent were higher in the North-East, Yorkshire (consent versus non-consent only), North-West, East Midlands and in Scotland. The analysis also confirmed that tenants were more likely to provide their consent, while home-owners owning outright were less likely to consent.

Two final large effects are worth noting. Consent was less likely to be given among those whose ethnic background not described as being 'white'. Naturally, this is a heterogeneous group, but small sample sizes make it difficult to break down any further. There was also a strong relationship between achieving informed consent and household size. The rate of obtaining consent was much lower in larger households, and this particularly reflects the higher number of proxy respondents in such settings.

Table 2.3 Logistic regression models of consent

	Parameter estimates and significance levels	
	Consent compered to non-consent	**Consent compered to others**
Intercept	0.6087***	0.1472***
Age group (ref = 40s)		
18-29	0.0741	-0.2746***
30-39	0.00178	-0.1025**
50-59	-0.00951	0.0324
60-69	-0.0311	0.0308
70-79	-0.0706	0.0117
80+	-0.2429***	-0.2715***
Region (ref=SW)		
North East	0.4802***	0.337***
North West	0.1544***	0.0447***
Yorkshire	0.2553***	0.1889
East Midlands	0.2047***	0.1058***
West Midlands	-0.1966***	-0.134*
East	0.1183*	0.0973**
London	-0.2968***	-0.2951***
South East	-0.0987*	-0.1069*
Wales	0.004	-0.0752
Scotland	0.2418***	0.2361***
Tenure (ref = mortgage)		
LA tenant	0.2026***	0.1334**
HA tenant	0.2785***	0.2663***
Rent – unfurnished	0.2362***	0.3109***
Rent – furnished	0.0827	0.368***
Own outright	-0.2371***	-0.2747***
Rent-free	-0.2272	-0.2465*
Employment status (ref=FT Employee)		
Full-time Self-employed	-0.3067***	-0.2815***
Part-time	-0.0058	0.1964***
Unemployed	0.1446*	0.3434***
Inactive	-0.0502	0.1866***
Limiting illness	0.2051***	0.1955***
Non-white	-0.4752***	-0.414***

Continued

Table 2.3 Continued

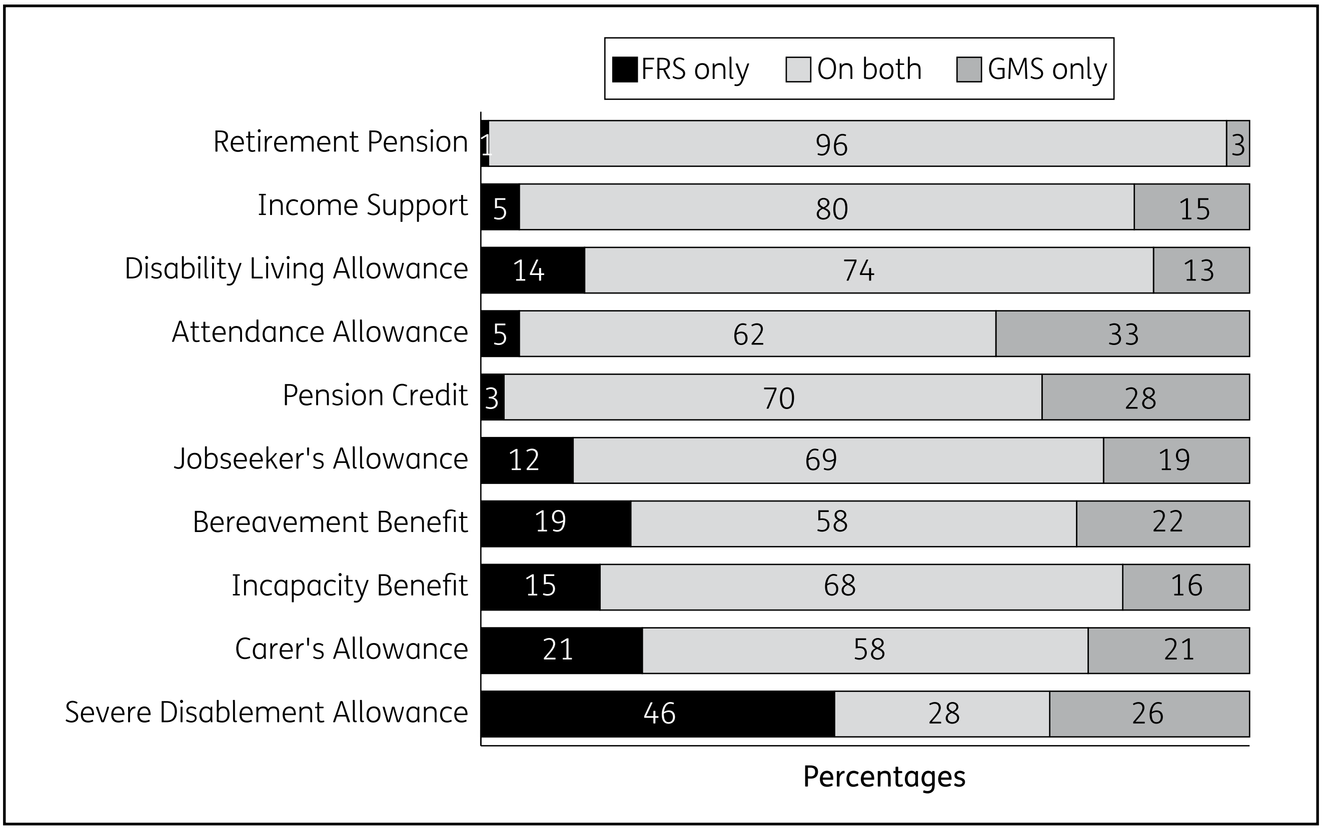
	Parameter estimates and significance levels	
	Consent compered to non-consent	**Consent compered to others**
Graduate	-0.0924**	-0.0351
Male	-0.00056	-0.1427***
1 person hh	0.0309	0.4416***
3 person hh	-0.0445	-0.2672***
4+ persons in hh	-0.1736***	-0.4913***
- 2 Log L (intercept)	45625.3	58538.3
- 2 Log L (model)	44510.2	56154.6

Note '*' indicates statistical significance at the five per cent level, '**' at the one per cent level and '***' at the 0.1 per cent level.

2.4 Reported benefit receipt and amounts of benefits reported

Figure 2.2 reproduces a chart from the methodology chapter of the FRS publication (Clay *et al.* 2011: Figure 7.1 p.108). It shows the degree to which respondents on the FRS are identified as being on Administrative data only ('GMS only' the mid-grey strip), 'FRS only' (the black strip), and 'On both' sources (lighter-grey strip).

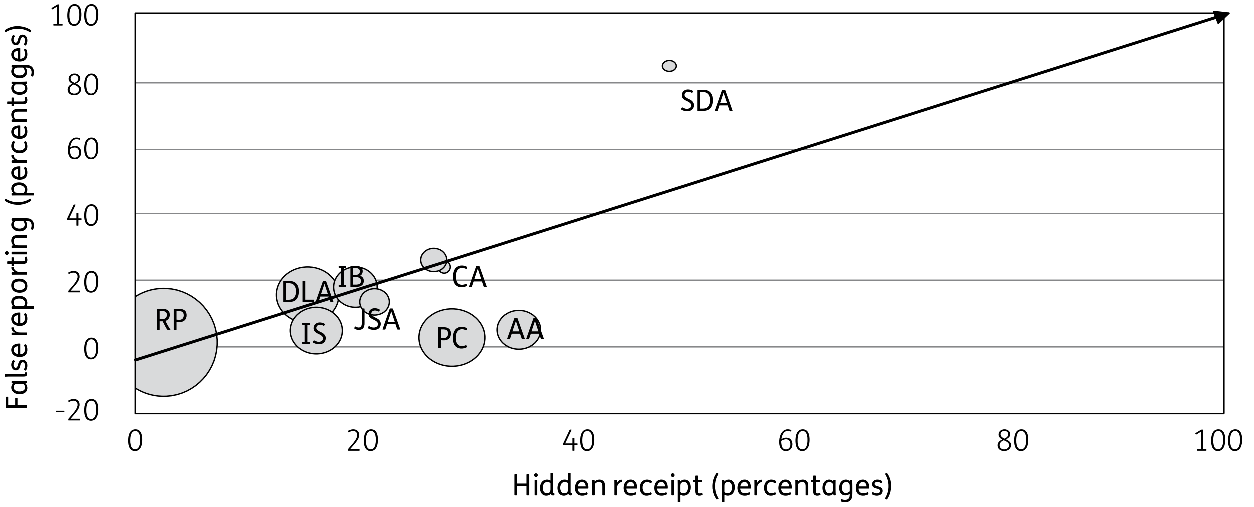
Figure 2.2 Percentage of adults shown in receipt of DWP benefits from FRS and administrative data 2008-09



Caseload and expenditure administrative totals held for DWP benefits often represent the entire recipient population and could be expected to provide a good match when compared with FRS totals. However these benefit comparisons show FRS survey under-estimation. Only Retirement Pension showed consistency between receipt recorded on the FRS and receipt recorded on administrative systems.

Figure 2.3 is derived from the information in Figure 2.1 along with information on caseload size which is reflected by the size of the bubble. The chart highlights differences for each benefit between false recipients and administrative receipt both quoted as a percentage of the total recorded receipt (on either source). Entries below the line highlight greater under-reporting of benefit on the FRS. Pension Credit and Attendance Allowance are the most extreme examples. For Pension Credit 29 per cent of the admin total is hidden, while three per cent of linked FRS cases reported receipt with no equivalent administrative record. Severe Disability Allowance showed the largest level of both 'Hidden receipt' and 'False recipients'.

Figure 2.3 Hidden receipt and false recipients derived from Figure 2.1



RP = Retirement Pension, Income Support = IS, DLA = Disability Living Allowance, IB = Incapacity Benefit, JSA = Jobseekers Allowance, PC = Pension Credit, AA = Attendance Allowance, CA = Carers Allowance, SDA = Severe Disability Allowance.

Figure 2.4 illustrates there was a strong degree of match between the rates of benefit found in the administrative data (for the linked consenters) and those reported in the FRS. The state Retirement Pension is presented in a later graph, as it can tend to dominate results. The overall match is quite good, albeit with a number of outliers. A set of horizontal points just under £100 per week of admin receipt (and above the x=y line) indicate some under-reporting of amounts.

Figure 2.4 Reported and actual amount of benefit received among linked consenters (all benefits, except Retirement Pension)

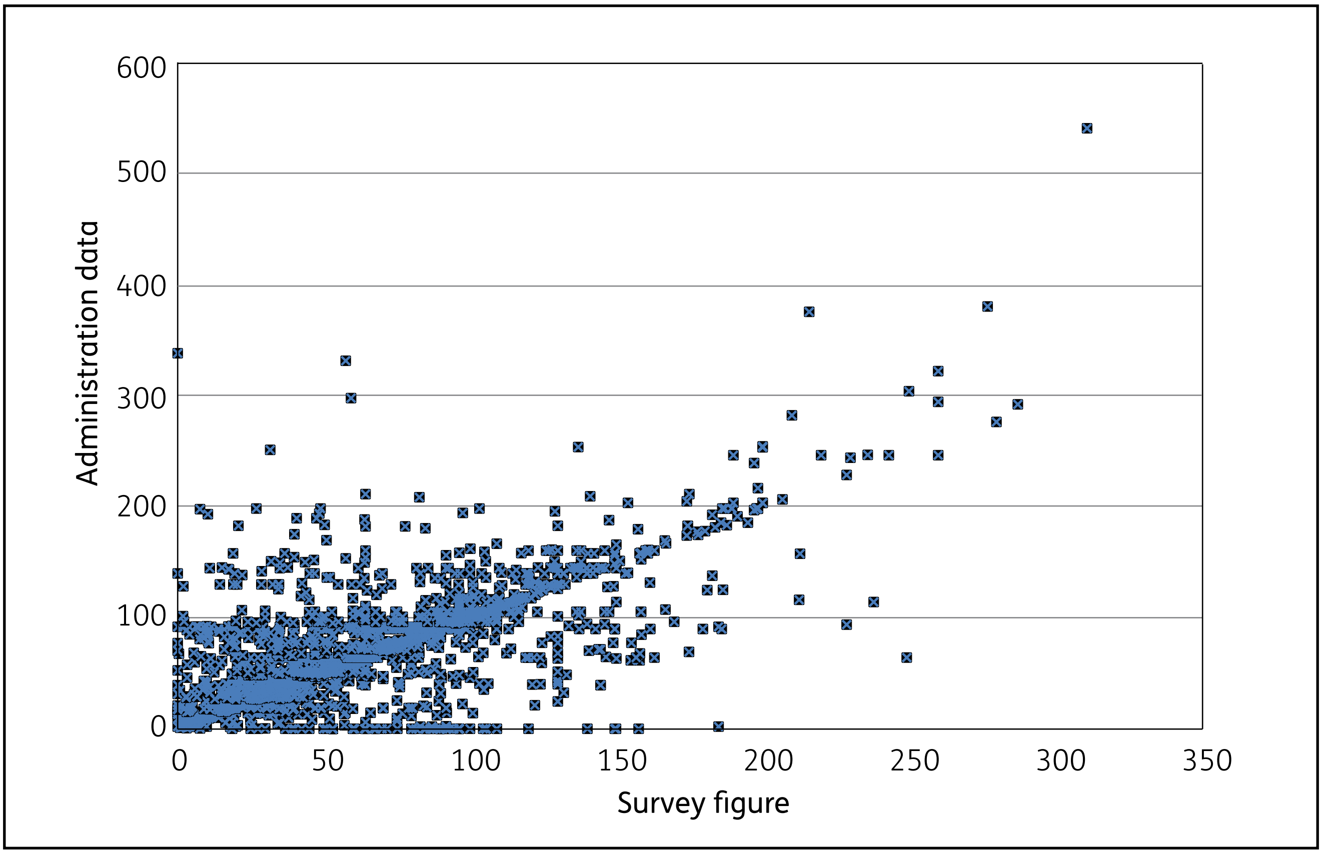
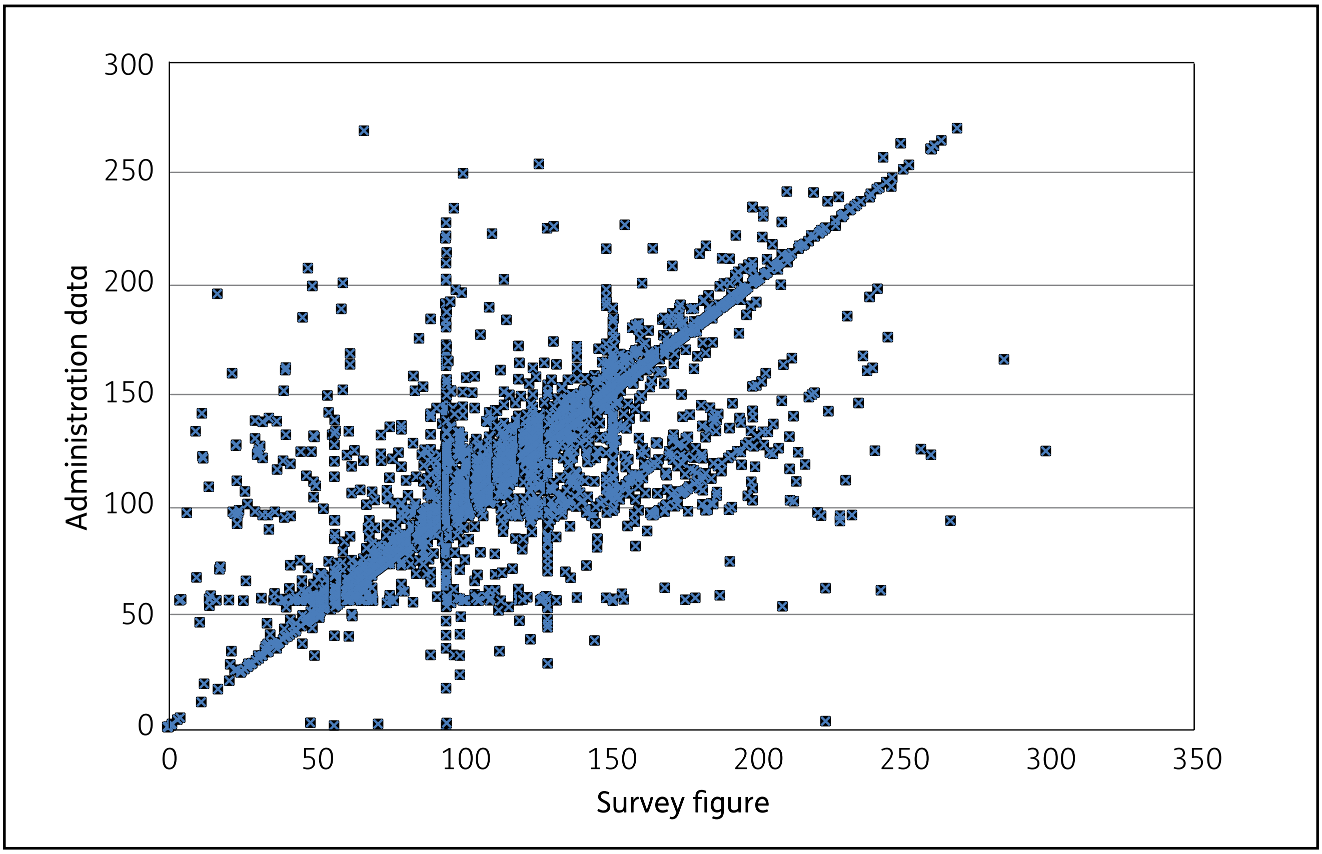


Figure 2.5 represents a similar analysis based on amounts of Retirement Pension. Again there was a strong overall positive correlation, but the effect of the 'bunching' at specific amounts (roughly at £95 and £160) is even clearer. This appears to reflect the imputation/editing of amounts of benefits at 'standard' amounts, even where the administrative data gives a more varied range of receipt. An example would be imputing a full standard rate of Retirement Pension rather than separating out an overall amount that was partly Retirement Pension and partly Pension Credit.

Figure 2.5 Reported and actual amount of Retirement Pension received among linked consenters



These figures can also be presented in terms of banded levels of mismatch between reported and actual receipt, for the linked cases (see Table 2.4). The degree of mismatch was divided into those of less than £10 per week, or less than £20 or £40, or exceeding £40. For Disability Living Allowance ((DLA), mobility) survey and administrative figures were separated by less than £10 in nine out of ten cases, with the remaining cases being separated between £20 and £40. This pattern was repeated for the care component of DLA, and for AA. This suggests possible misreporting (or possible mis-editing) of the appropriate level of these benefits being received (the care component of DLA has three separate levels with flat rates of benefit, the mobility component two levels, and Attendance Allowance has two levels equating to the two higher levels of the care component of DLA).

Table 2.4 Benefit amount mismatches by benefit type

				Row percentages
Benefit	**Within £9.99**	**£10-£19.99**	**£20-£39.99**	**£40+**
DLA – mobility	89	–	11	–
DLA – care	80	*	14	6
AA	80	*	20	–
JSA	80	7	4	9
Retirement Pension	77	7	7	9
Incapacity Benefit	73	6	6	15
Pension Credit	70	9	8	13
Income Support	50	20	14	16

Note '-' indicates no cases, and '*' means less than 0.5 per cent of respondents in that row.

Errors tended to be greatest for Income Support (16 per cent differing by £40 or more), with only 50 per cent providing an answer that was within £10 of the true rate of receipt. This compared with seven in ten for Pension Credit, the next most likely to be misreported.

There was not strong evidence of a greater mismatch in survey and administrative amounts at different times of the year. It seemed plausible that the extent of misreporting might be higher in the first part of the financial year, when the majority of changes to benefit rates take place (uprating). However, the mismatch was greatest some months later.

Table 2.5 Benefit amount mismatches by sample quarter (across all benefits)

				£ amount of mismatch
Benefit	**1 (April-June)**	**2 (July-September)**	**3 (October-December)**	**4 (January-March)**
All WPLS benefit amounts – including exact matches	£0.77	£0.93	£1.26	£0.73

Reported errors were less, however, where a person reported consulting a benefit letter regarding the amount or (to a lesser extent) where they had consulted a bank statement regarding the amount in payment.

Table 2.6 Benefit amount mismatches by checks made by respondent regarding benefit receipt (across all benefits)

				£ amount of mismatch
Benefit	**Consulted benefit letter**	**Others**	**Consulted bank statement**	**Others**
All WPLS benefit amounts	£0.25	£1.13	£0.83	£1.48

2.5 Selection models

The Heckman selection model is used where an outcome is only known for a selected group. A classic case is wages for those in work, when we want to understand how wages are affected by whether the person is in work.

There appears to be an analogous situation here. There is a first stage where people consent, and a second stage where the link between reported benefits and actual benefits is modelled. However, interest is in the second model, potentially to make inferences about the non-consenters.

There is a selection into consent and non-consent (or, linked and unlinked may be used). Heckman's insight was to think of this kind of selection bias as being akin to having an omitted variable (Heckman 1979).

2.5.1 Selection equation

$z_i* = w_i'\alpha + \varepsilon_i$

z_i* is a 'latent variable', the propensity to be linked to admin data

w_i' variables associated with linkage

α coefficients

ε_i error term

2.5.2 Outcome equation

$y_i = xi'\beta + ui$

y_i outcome (amount of benefit)

x_i' variables affecting level of benefit (esp, the survey estimate). One of the terms is lambda λ a measure of the 'inverse Mills ratio'

β coefficients

u_i error term

Overall, rho ρ = cor(εi , ui).

The outcome is only observed if the unobserved latent variable in the first equation exceeds a particular threshold. An additional variable is entered into the main outcome equation (the inverse Mills ratio – basically a measure of the residuals, and sometimes represented as lambda) that is derived from the selection equation which is usually a probit model. The outcome model will also produce a measure rho, the correlation between ε_i and u_i indicating if there is selection bias – if rho is low, then an OLS method of determining regression coefficients on the main outcome ignoring selection is sufficient.

If there are no unmeasured variables that predict selection into being linked – in other words there appears to be no selection bias and hence the value of rho is low – then the outcome equation may be run without being concerned about selection bias. The OLS equation may be run without needing to run the probit to address selection issues, if there is no selection bias.

There is a SAS procedure (PROC QLIM) to run such models[2], invoked using the following commands:

```
proc qlim data=linkeddata ;

          model lnkdwp = age sex /discrete; /* selection equation, probit */

          model benamta = benamtf nchild age / select(lnkdwp=1); /* OLS outcome */

run;
```

See Appendix for the full SAS code needed to be specified to run these models.

For model identification it is important that there are at least slightly different lists of independent variables, with one or more variables that appear in the selection equation but not in the equation of

[2] In running these models, SAS uses a full maximum likelihood approach, rather than the original two-step estimator that was set out by Heckman. The latter was developed on slower computers with a need to make the computation side more practicable.

interest (e.g. sex in the above equations). Generally it is assumed that there is a binary selection at the first stage – into consent or linkage status in our example.

The selection models were run by a selection equation that included age and age-squared, gender, living in London, illness, and having a graduate-level qualification. Heckman selection models are generally quite sensitive to the precise specification used. This proved to be the case with the revised models, as several key benefits swapped places in terms of whether there was sample selection bias detected, as follows – compared with a model consisting simply of age and gender (Table 2.7).

Table 2.7 Selection bias in Heckman models

Apparent selection bias	No apparent selection bias
Retirement Pension	Income Support
Attendance Allowance	JSA
Incapacity Benefit	***Disability Living Allowance (Care)***
Employment and Support Allowance	***Disability Living Allowance (Mobility)***
Industrial Injuries Disablement Benefit	
Pension Credit	

The assessment of whether selection bias existed for those ***benefits*** emboldened and italicised was sensitive to the choice of variables in the model.

The presence of selection bias means that the unobserved determinants of the benefit amount are correlated with the unobserved determinants of agreeing to data linkage.

2.6 Conclusions

In this section the characteristics of non-consenters, consenters and proxies were compared. In most respects the consenters and non-consenters were surprisingly similar. Insofar as there are differences between consenters and the rest of the population, many of the differences related to the situation of proxy respondents. In particular, proxies were more likely to occur in multi-adult households (by definition they cannot arise in single adult households). However, even comparing consenters against both non-consenters and proxies, these groups tended to have similar incomes and similar grossing weights – the latter indicating relatively few differences in overall response rates.

In addition to looking at consent, amounts of benefit reported were compared with the correct amount on administrative data. In most cases there was a very good match, with correlations of around 0.85, with some significant outliers. As argued elsewhere, it is misreporting of the fact of receipt, rather than of the amount of benefit, that seems the larger issue. However, Heckman sample selection models were used to consider if the amounts of benefit reported were subject to sample selection bias. There was evidence of selection bias for some benefits, but not for others, and the benefits affected varied depending on the set of independent variables included in such models. It is also possible that the non-normal distribution of amounts of benefit may be affecting the validity of such models, which are (in any case) quite sensitive to minor changes in specification.

Overall, while differences between consenters and non-consenters cannot be ruled out, the extent of any bias in the bi-variate analysis and selection models appears not to be large. The extent of bias owing to consent is likely to be relatively small compared to problems caused by, for example, non-response which appears to be a larger issue in studies where the effects can be compared.

3 The options

This section outlines alternative means of using administrative data for consenters (as before analysis is restricted to receipt of DWP benefits). These may be compared with the status quo of continuing to use FRS data.

3.1 Editing the consenters, retaining the non-consenters

One approach is to simply retain the survey data for those not consenting (or those who consent but for whom a match cannot be made) and to substitute the administrative data for those who do consent (and whose data could be matched). This may be regarded as a form of data editing, using available information to improve what we know about particular respondents. At present data editing has to include quite a number of assumptions, although it is also possible for analysts to work with unedited data if they have concerns. This approach may be perceived as having an element of inconsistency. It seems to ignore the fact that it is known (or at least strongly suspected) that patterns found among the consenters, of under-reporting receipt of benefits, will also apply to non-consenters. However, it is a different matter to apply such adjustments at a micro level, such as in a general use dataset that is distributed to analysts, compared with making aggregate adjustments on this basis.

3.2 Complete cases and re-weighting

A general attraction of using administrative data in some form (options 2-4) is that it might be possible to shorten the main FRS interview. However, only if it could be shown that asking for consent early in the interview did not adversely affect consent rates, and if there was convincing evidence that there was little or no bias between consenters and other respondents. Clearly, such a situation is a very long way away, and it seems unlikely that this idea could be progressed while still providing a reliable dataset.

Once someone has consented to having their data linked to administrative records, it would be possible to skip the questions on benefits, or at least to reduce the number of detailed questions. However, the potential for doing so is rather small, particularly if we wish to monitor how well the survey data corresponds to administrative data for such respondents.

There is potentially a large further financial saving if it was possible to obtain sufficiently robust estimates from a sample of consenters and to dispense with those who didn’t grant such permission – only interviewing the 50 per cent of people who were willing to provide consent to data linking. However, in this scenario the confidence intervals on any statistics would be rather wider, as they were based on a smaller sample size. This latter effect may well be larger than any gains in lower measurement error from having administrative data. It is already necessary to combine several years of data for some analyses, which would become particularly problematic with a much smaller sample size. We could also not rule out the possibility that only a biased half of the sample would thereby be included.

However, since it is known that there are (small) differences between consenters and non-consenters, it would seem to make sense to re-weight the consenters so that they conform more closely to the national profile. This re-grossing or re-weighting idea has been used on the US Health and Retirement Survey. Record linkage has been carried out for social security sourced data on earnings in this US source, with around three-quarters of cases linked. The dataset including these records is re-weighted to take account of differences in rates of consent/linkage. It is also worth noting that some policy tools that provide costings for benefit reforms may also want to calibrate to specific benefit totals. This is the approach within the DWPs Policy Simulation Model (PSM). One issue with a grossing regime to do this for all benefits is the wide range of benefits involved and the potential for grossing factors to become highly variable, particularly for benefits with small numbers of recipients.

3.2.1 Calibration re-weighting of the consenters

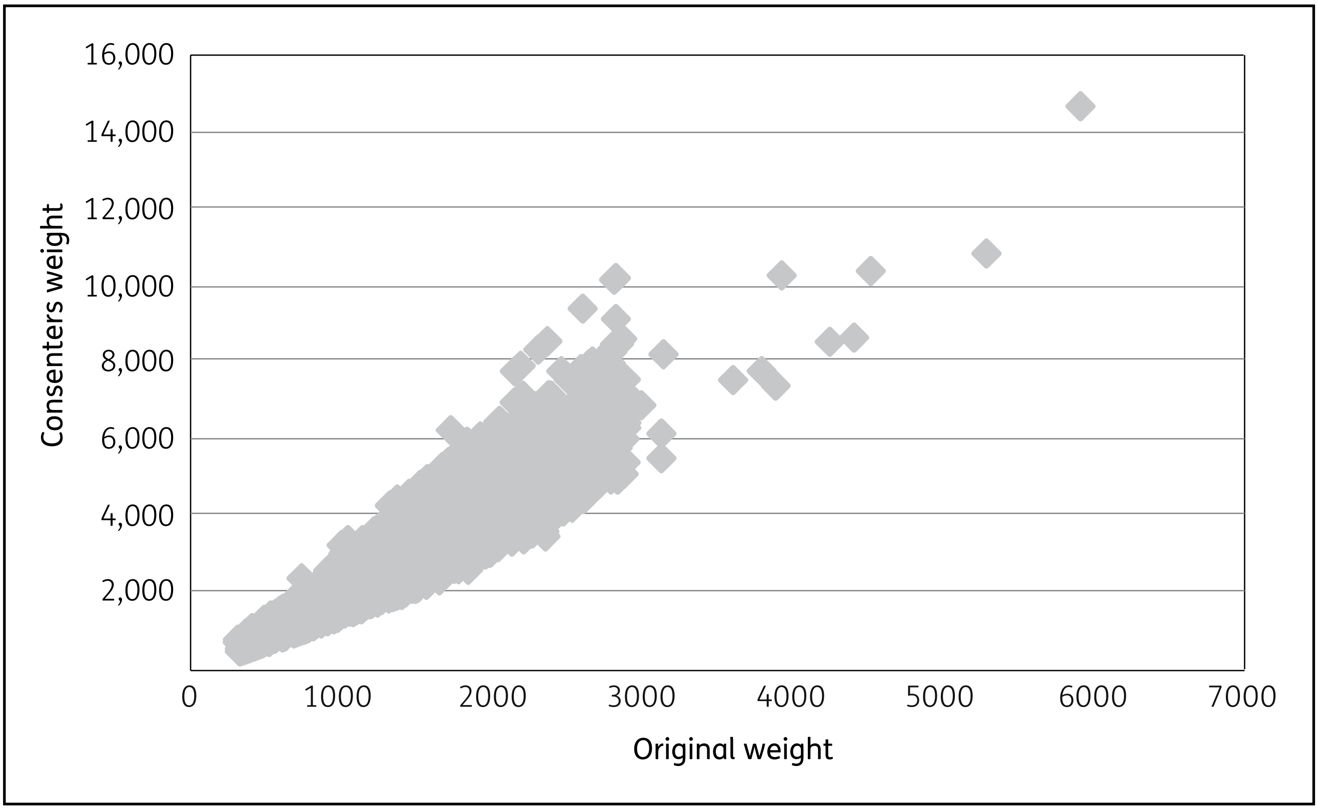
Weighting a sample survey to sum to population totals across a range of different variables is technically quite specialised but in practice commonplace on United Kingdom government surveys. Deville and Särndal (1992) set out the standard modern approaches to different forms of 'calibration weighting' (although they have older precursors in terms of 'raking' techniques). In particular they conceptualised the problem as having to make totals add up to J external constraints – which leads to many different possible solutions where the sample size N is much greater than J. The task of finding an appropriate weight often proceeds by minimising the distance between an initial weight, such as a design-based weight (dealing with any issues of over- and under-sampling in the sample design) and the final derived weights. On several government surveys, the external constraints are specified in terms of the age, gender and regional balance. The initial weight is often a design-based weight (although for our purposes the main grossing weight is the natural choice as the initial weight for deriving new weights).

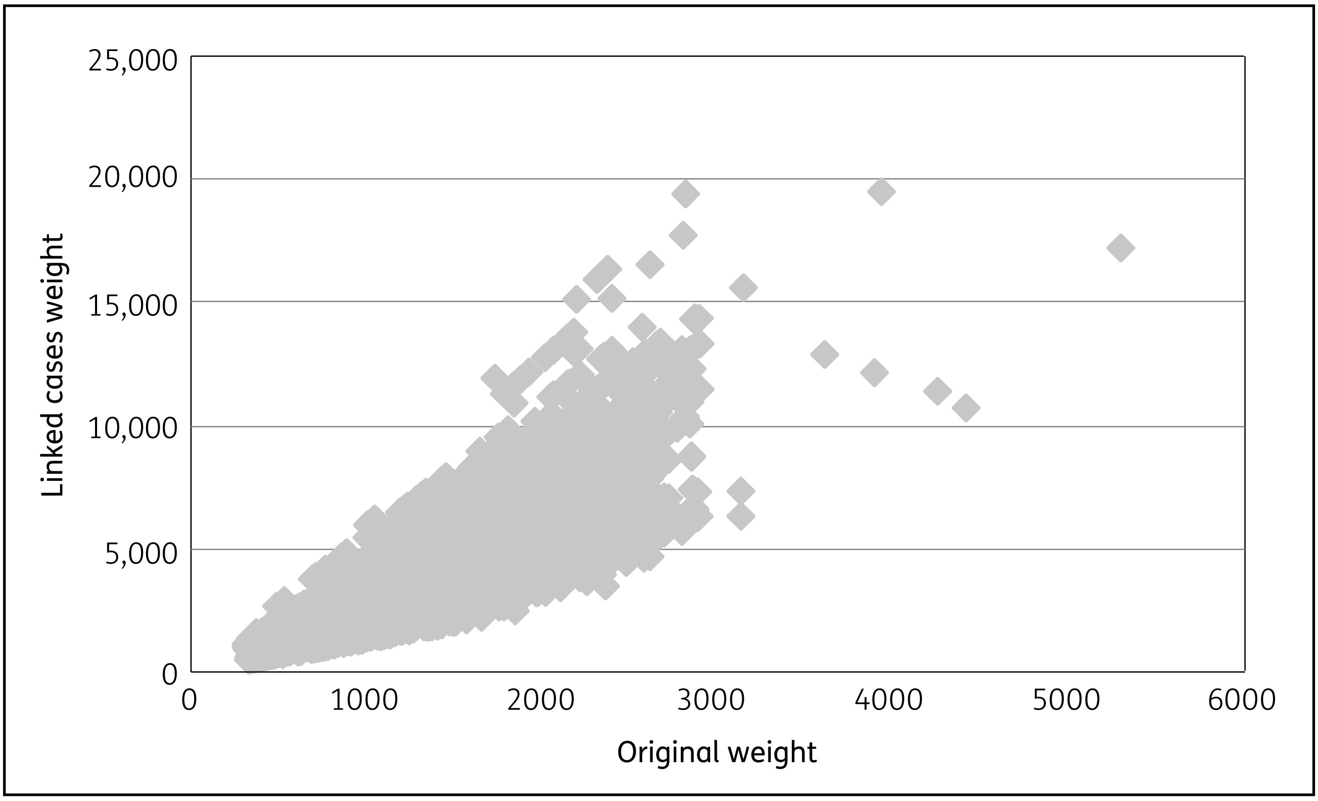
The FRS grossing regime is very complex, with a large number of external constraints. Population totals are the key constraints used, and specified at the level of the region (e.g. North East, London, Eastern). The grossing also takes account of numbers of lone fathers and lone mothers, and Council Tax banding, among other constraints (Department for Work Pensions 2005). The large sample size permits this large number of constraints in the grossing (in smaller samples, having a large number of constraints can lead to quite diverse weights or, in calibration routines based on models, to problems that such models do not converge to a sensible solution). It is hard to tell in advance if this would continue to apply to a similar number of constraints imposed on a somewhat smaller dataset based on the consenters.

A new set of grossing weights were calculated for consenters as a whole, and for linked cases, using what Deville and Särndal (1992) labelled as the 'logistic method'. This iteratively fits the new weights, and prevents there being any possibility of negative weights. The main grossing weight (GROSS3) was used as the starting point, so that the new weights should be relatively close to this initial weight. This turned out to be the case – see scatterplots in **Figure 3.1**.

The new grossing weight based on the consenters shows a rather smaller spread than one calculated for just the linked cases, and a higher correlation with the original grossing weight, GROSS3 (there was a Pearson correlation of 0.95 for the consenters' new grossing weight with the original grossing weight, and 0.85 for that based only on linked cases). The distribution of the new weights is the widest, and carries the greatest risk of outliers, with the linked cases compared with the consenters' cases (and compared with the original grossing weights).

Figure 3.1 Associations between original and new grossing weights





3.3 Imputation

There are many different approaches to imputing single missing values. These can be divided into ‘single imputation’ approaches, and the increasingly popular approach known as ‘multiple imputation’. A single imputation involves one value for each missing data item; multiple imputations calculate multiple values for each missing data item (as part of a modelling strategy bringing together results from those several versions of the imputed dataset).

3.3.1 Imputation methods

In this section we consider a number of different methods of imputing missing data. These include regression imputation and hot-deck imputation, to name two of the most common approaches that impute a single value to a missing observation. It is also worth noting that the Heckman selection model is also a kind of imputation approach, with a selection correction term applied to what is essentially a regression imputation approach.

Mean substitution and its close cousin **regression imputation**. These use either the mean of existing cases (or, sometimes, the mode or median), or the conditional mean from a regression equation with relevant variables. This generally has the disadvantage that standard analysis of the imputed data will appear to have an unreasonably low standard error by increasing the sample size, even though (arguably) we have not added more information. And the former can lead to considerable clustering at the mean.

Regression imputation can either be conducted in a deterministic way, using the regression coefficients to produce the best-fit value, or including some degree of randomisation through adding a residual to the prediction (such a residual could be randomly drawn from a normal distribution or selected from an actual residual from respondent data.) Regression imputation is probably a reasonable method to the extent that the model is based on a true depiction of the relationship between the missing data and the set of independent variables making up the regression. It is less likely to be effective with variables with elements of non-normality (such as categorical variables) – and, in particular for our purposes, whether someone receives a social security benefit or not.

Hot-deck imputation is probably the most common method of imputing single missing values in the main United Kingdom/Great Britain surveys (and especially on FRS). It consists of replacing any missing observations with the value from a case that is similar on a number of background variables (such as gender, age group, location), or is a ‘nearest neighbour’ using some kind of distance metric. One advantage is that actual values will be chosen to impute with, and there is no need for any kind of distributional assumption to be made. This is particularly important with some kinds of data, including benefits, where the data has rounding effects, maximum and minimum amounts, amounts that occur particularly often, and so on.

Multiple imputation

Multiple imputation involves reproducing several different versions of the dataset, with some randomisation of how the missing values have been imputed. A new dataset with those multiple values is then analysed using fairly standard statistical approaches. Little and Rubin (1987) analysed different kinds of imputation approach, and recommended multiple imputation for making inferences based on survey data.

In SAS it is possible to somewhat automate this process using PROC MI and PROC MIANALYZE – the overall approach is outlined in an appendix to this paper.

3.3.2 Description of mechanisms for missing data

Data may be said to be missing on one of three different mechanisms:

Missing completely at random (MCAR). The chance that X is missing is not related to the value of X, nor to the value of any other variables in the dataset. Analysis of such data remains unbiased. This is highly unlikely to be the case in practice.

Missing at random (MAR). The chance that X is missing does not depend on the value of X after controlling other variables. It is possible to obtain unbiased results with more advanced methods, such as multiple imputation.

Missing Not at Random (MNAR). Doesn't meet MAR assumptions. There are no easily available approaches to deal with this scenario, though it may be possible to model the missing data process. In this scenario the missing data is sometimes known as 'non-ignorable'.

There is no means of being sure which missing data mechanism has affected any particular dataset since the values of the missing data are, by definition, unknown. Therefore, analysts generally proceed under the assumption that data is missing at random. The approach of multiple imputation is designed to give unbiased results when data is MAR.

3.3.3 Statistical purity and survey practicalities

Some of the more advanced methods are able to recover unbiased parameter estimates from models when the data is missing at random. An approach such as multiple imputation is perhaps optimal for such situations, but the results themselves will be specific to each model. The challenge facing this project is, however, that of distributing a general-use dataset. Multiple imputation is arguably less suitable for a wide base of users as most will be more familiar with a single value imputation (e.g. based on regressions or hot deck algorithms). Having to use multiple datasets would clearly complicate the task of secondary analysis. It would certainly require a step change in attitudes to handling secondary data. It is open to question whether that day will ever come, but it certainly has not arrived yet.

There are also problems, discussed below, that are specific to imputing amounts for benefits. Nevertheless, it should always be open to external analysts to apply methods like Multiple Imputation, as the data should be sufficiently transparent that it is clear how it has been processed.

Existing Family Resources Survey imputation

A considerable amount of imputation is conducted on the raw data to arrive at the final dataset distributed to users. The number of changes made to benefits during imputation is very large – around 24,000 such changes in 2009-10, or around 20 per cent of the values of benefits. What seems to be an apparent increase over the previous year is the result of imputing amounts of winter fuel payments and council tax benefit following changes in what is collected in the main survey.

FRS imputation is of a number of different kinds, with some 'bulk edits' but a larger number of imputations based on the hot-deck approach. It is quite rare for imputations to exceed 25 per cent of the values of any particular variable – in 2009-2010 there were nine such variables, and each of them attracted rather less than 1,000 missing responses. They included such variables as 'Amount included in rent for water/sewerage' (45 per cent imputed, 558 missing cases) and 'Amount of tax in last 12 months' for the self-employed (33 per cent imputed, 516 missing values).

These figures indicate that the FRS dataset commonly used by users incorporates a great deal of editing and imputation. It seems rare in published work for users to go back to the 'raw' data and adopt alternative imputation methodologies – although the transact 'data table' does make this possible. So users are certainly familiar with post-imputed data, and there seems to be limited interest in re-doing different imputation algorithms[3].

The imputation challenge for Family Resources Survey linked data 2009-2010

Table 3.1 shows, for some of the more common benefits, the number of reported values, and number of administrative values for the successfully linked cases. In most cases the mismatch in terms of average reported amounts is quite low. However, the gap between the number of survey values and the number with either admin or survey values indicates a sizeable proportion of 'hidden recipients', except for the Retirement Pension.

Table 3.1 The imputation challenge for FRS linked data 2009-2010

Benefit	Number of survey values	Number of admin data values[1]	Either set of data	Average survey value	Average admin data value
DLA care	2,174	1,045	2,430	46.75	41.07
DLA mobility	2,111	1,146	2,361	39.24	42.19
Pension Credit	1,966	1,511	2,336	48.32	51.33
Retirement Pension	11,872	6,125	11,918	107.90	109.58
AA	1,004	699	1,267	57.57	58.86
JSA	1,086	608	1,235	62.11	60.46
Income Support	1,574	941	1,720	75.06	83.09

[1] Number of admin data values for linked consenters.

To attempt an imputation of missing amounts of benefits for hidden recipients would be a more radical order of magnitude greater than any existing imputations within the FRS. It would imply that more than half of the amounts of benefits would need to be imputed, and (of course) this would affect a very high number of values.

It should also be pointed out that the receipt of benefits, and their amounts, must conform to administrative or statutory rules, and not simply to statistical rules. Amounts of benefits often cannot exceed particular values, and there are strong inter-dependencies between benefits. These include rules about overlapping benefits (e.g. a Jobseeker's Allowance (JSA) recipient cannot also receive Carers Allowance), and there are also linkages between benefits – all else being equal, a DLA recipient is likely to receive more IS than a non-recipient. Therefore, there cannot be straightforward imputation algorithms that do not take account of these regulations. In other words it seems likely that there has to be an element of 'Deductive methods impute a missing value by using logical relations between variables and choosing the most plausible value' Durrant (2009). Imputation, Durrant notes, is inevitably linked to the specific context and the aim of particular analyses.

Even so, the significant simplification of several benefits that will arise from the introduction of Universal Credit may make life somewhat easier, in that for many cases the overall amount will relate to one benefit.

[3] In principle respondent and linked data could be made available to users in a similar way to the FRS 'transact data' which shows pre- and post-edited data.

3.3.4 Categorical and continuous variables

Use of administrative data has found small, though important, differences in reported amounts. However, there seems to be a greater degree of mismatch in terms of receiving particular benefits – cases where receipt is claimed but not corroborated, or not claimed despite being represented on the administrative data (i.e. false and hidden recipients).

However, imputation techniques tend to be attuned toward continuous data, and/or have normality as a key assumption. There are further issues to be confronted when trying to impute binary variables (see Horton *et al.* 2003). The needs of analysts are more likely to be met by some kind of rounding – for example a 60 per cent chance of receipt might well default to actual receipt on the basis that it is more likely than not to happen – rather than having implausible values for categorical variables. But this may lead to bias in the resulting models (Horton *et al.* 2003). So, the use of fractional variables may be better for some modelling purposes, as they more appropriately capture uncertainty about people's status – though they would create problems for those requiring more straightforward analyses such as cross-tabulations. It is also worth emphasising that such approaches would generally only be based on imputing receipt of benefit, and a separate stage would then be needed to generate the appropriate amount of benefit received. Existing FRS imputation processes for specific benefits are not currently independent of answers to questions about benefits. Benefit imputation algorithms make use of responses to questions that are closely related (in terms of entitlement rules or policy rationale) and knowledge of common respondent errors when reporting benefit receipt to the FRS interviewer.

3.3.5 Some conclusions regarding imputation

It was initially assumed that imputation, and particularly multiple imputation, would be an attractive technique for dealing with the unlinked cases. Closer examination has raised a number of problems for such a strategy, in particular associated with the particular challenges of imputing benefits where amounts are not simple variables but subject to myriad rules of different kinds. This makes it difficult to use imputation methods that do not take account of these features. That probably leaves hot-deck imputation as being a plausible approach to imputing amounts for non-consenters. Hot-deck imputation is the main method that was employed in this analysis when testing different approaches. This was done as a kind of 'proof of concept', and a longer period of testing would improve on the simple hot-deck imputations that were used in this analysis. Hot-decking was used to impute 'true' values of benefit receipt for non-consenters. The reported benefit amounts were used as the main selection criteria for the donor cases. In other words, a non-consenter reporting £X as their survey amount would generally have the true admin amount £Y determined by selecting a donor case (who was a consenter) who had also reported benefit receipt of £X in the survey.

In practical terms, it is known that there is a strong correlation between the survey figures reported and corresponding administrative data, it seems likely that the reported survey value is going to be the most useful strata for identifying particular 'donors' of data for imputation to the non-consenters. Further empirical testing (Chapter 4) will need to analyse whether this is any kind of improvement on the survey figure provided. There is, however, quite a high proportion of missing data for any kind of imputation to deal with, and effectively there are about as many donor cases as there are recipient cases.

Some analysts may want to use the data and use a technique such as multiple imputation for specific analytical tasks. However, for more general purposes it seems preferable to use single value approaches to imputation. The nature of benefit receipt, and its associated administrative rules implies that hot deck imputation, although venerable, is likely to avoid some of the problems that beset other methods. This is partly owing to the non-parametric nature of hot-deck imputation, whereas other methods seek to impose some distributional assumptions on the problem. I have sympathy with the following view: '*This author believes that the real problem of imputation is the interaction with editing..., writers prefer to simplify the problem so that it is amenable to mathematical analysis*' (Sande 1982: 151-2). Again a possible route is to consider the use of linked data as a form of editing rather than a separate stage of imputation – insofar as the two may be distinguished at all.

3.3.6 Approach taken to imputation in this report

In this research project, hot-deck imputation was used to impute the amount of benefit received, using the consenters to provide data for the non-consenters. Towards the end of the analysis, we use regression imputation to identify hidden recipients of benefits (on a probabilistic statistical basis). The former adjustment can be considered to be a relatively standard approach to imputation. And here, with limited differences between survey responses and administrative data, the impact of the adjustment is relatively minor. Attributing benefit receipt to those who do not say they receive benefits is conceptually a more radical adjustment at a unit record basis as it goes well beyond existing FRS imputation. However an adjustment of such a kind may be appropriate to fully estimate, or adjust for, any inherent bias in FRS-based outputs including National Statistics.

4 Evaluation

This section evaluates different options that have been suggested for addressing the key problem of data linking – what to do about those who do not consent to such linking. Regression analyses where receipt of different benefits forms the dependent variables are examined. In the second section the average amounts of different benefits reported are examined.

4.1 Average amounts of benefits received

In **Table 4.1** the average amounts of different benefits received are shown. Also considered are results from national data, the standard FRS dataset, replacing consenter data with admin data, imputing for non-consenters, looking only at complete cases, and last those consenters re-weighted to key national parameters. Five different benefits were analysed where sample sizes were sufficiently robust.

The first feature to note is that, summed across the different benefits, all of the proposed approaches did better than simply relying on the FRS data. The gap between the worst new approach and the current approach was actually much larger than the differences within the proposed new options. This suggests that using any of the options, at least in terms of benefit amounts, would be an improvement on the current system.

The precise 'winner' depends on the selection of benefit, and results for Retirement Pension make a substantial difference compared with the other benefits. Even so, it seems that using administrative data for the consenters, and a simple hot-deck imputation[4] for the non-consenters, produces results that are closest to the known national totals. There is little to choose between the other approaches.

Table 4.1 Average amount of benefit received

£ per week

Benefit	National data (August-2009)	FRS data	Admin and survey	Admin and imputed	Consenters only	Consenters re-weighted
AA	60.03	57.26	58.14	58.46	58.76	58.77
Pension Credit	55.66	48.11	51.36	52.30	51.77	52.89
JSA	59.73	60.98	60.20	63.42	60.76	60.58
IS	84.61	73.53	80.10	83.50	82.29	82.53
RP	102.35	108.43	108.06	107.77	109.77	110.61
RMSE[1]		6.72	3.88	3.41	3.96	4.06

Note: weighted by gross3, apart from final column.

[1] The root mean squared error (RMSE) is a measure of the differences between the values predicted by a model or an estimator and the values actually observed (in the population model, in this case). It sums the squares of the differences between model and population parameters, and then the square root is taken. Smaller values imply a closer fit to the national data.

[4] The hot deck algorithm looked separately at each different type of benefit, and grouped amounts of benefits report within them (in £20 bands).

4.2 Models of receiving benefits – effectiveness of different approaches compared with population data

It is possible to use administrative data (from the Department for Work and Pensions 'tabtool') to analyse the numbers of men and women in different age bands who receive different social security benefits. Population data on numbers of men and women of different ages can be used to identify the rate of benefit receipt broken down by gender and age. That provides sufficient data to run a logistic regression model of those receiving benefits – and one that provides a benchmark against which to compare different algorithms that may be run on the survey data.

In **Table 4.2** the performance of different approaches to data linking with the population 'benchmark' model were contrasted. Age group and gender were the only independent variables it was possible to include. Of the models shown, those based on the existing FRS had the highest error. Replacing survey data with admin data for linked cases had the lowest error, while an algorithm that imputed receipt to the survey data on unlinked cases (where receipt was not already reported) performed second best.

Using complete cases (with original weights) did next best, followed by using the re-weighted consenters.

Table 4.2 Comparing logistic regression results: Receipt of Attendance Allowance

Logit coefficients and error

Variables	Population (bench-mark)	Existing FRS	Admin + Survey	Admin + (Survey + Imputed)[1]	Consenters only	Consenters re-weighted
Intercept	-2.387	-2.689	-2.500	-2.414	-2.329	-2.317
Aged 65-70	-1.407	-1.409	-1.349	-1.306	-1.382	-1.389
Aged 75-80	0.818	0.663	0.693	0.669	0.573	0.580
Aged 80-85	1.582	1.284	1.382	1.431	1.397	1.367
Aged 85-90	2.244	1.952	2.131	2.136	2.150	2.115
Aged 90+	2.633	2.378	2.614	2.776	2.817	2.819
Male	-0.370	-0.301	-0.373	-0.399	-0.479	-0.470
RMSE		0.227	0.110	0.112	0.148	0.155

[1] In this column the dependent variable is equal to the admin data where present, to the survey data if that shows receipt, and to an imputed group of what are the most likely to be 'hidden recipients'.

Appendix A shows similar tables using the same specification of logistic regression model for Pension Credit (PC), Jobseekers Allowance (JSA), Income Support (IS) and Disability Living Allowance (DLA).

The next table (Table 4.3) provides an overall summary from these models of benefit receipt. For Pension Credit the imputed version of the model performed best – but performed worst for receipt of IS. For all models apart from IS, the approach of using administrative data for the consenters (where linked) and the survey data for all other cases also provided results close to the population benchmark models (but again was weak for IS). The results based on the existing FRS did about as well as just looking at the consenters (with the original weights). The results for consenters subject to re-weighting were marginally the worst, although the results for the last three approaches represented in the latter three columns of the table actually tended to be quite similar.

Overall, the approaches of Admin plus Survey seemed generally the best, but the difference between that and Admin plus (Survey plus Imputed) was extremely marginal in most cases and indeed the latter approach had the smallest overall error on these four benefits and was clearly the optimum for receipt of Pension Credit.

Table 4.3 Logistic regression results: margins of error

					RMSE
Variables	**Existing FRS**	**Admin + Survey**	**Admin + (Survey + Imputed)**	**Consenters only**	**Consenters re-weighted**
Accuracy (RMSE)					
AA	0.227	0.110	0.112	0.148	0.155
Penson Credit	0.205	0.164	0.058	0.247	0.255
JSA	0.282	0.282	0.286	0.323	0.333
IS	0.253	0.284	0.288	0.217	0.245
Average (RMSE)	0.242	0.210	0.186	0.234	0.247
Ranking (1st - 5th)					
AA	5	1	2	3	4
Pension Credit	3	2	1	4	5
JSA	1=	1=	3	4	5
IS	3	4	5	1	2
Average (rank)	3.13	2.13	2.75	3.00	4.00

5 Conclusions

The analysis suggested the following key conclusions.

- The level of 'consent bias' was relatively low – and it was proxy cases that appear to introduce the largest deviations between consenters and non-consenters. Household size had a large effect on the likelihood of having proxy interviews, and therefore not being able to obtain consent. There were also powerful independent associations with region, housing tenure and non-white status. However, there were few differences in incomes between consenters and the rest of the sample.
- Heckman selection models indicated that reported amounts of benefits are subject to selection bias for some benefits, but not for others. It was unclear how far these results were affected by the non-normality of benefit amounts. Small changes to the selection models can have large effects on those benefits thought to introduce selection bias into the amounts of benefit reported:
  - While this indicates there may be selection bias for some benefits – so that any adjustments to reported amounts for non-consenters should not be based only on the pattern for consenters – this is not consistent across different benefits or different model specifications.
- In most models, in most settings, using the administrative data in some form provides more accurate results (closer to the known national picture) than simply using the standard FRS. There are exceptions that occasionally surface. The basis for comparison with national results was limited to what was available for administrative data. Gender and age only were used.
- Using the administrative plus survey data for appropriate respondents tends to be better than simply using data on consenters even where that data is re-weighted. Indeed, the original weights applied to the consenters did about as well, sometimes better, as having a new set of weights. It was not clear why this surprising result should happen, and may suggest a deeper search for a more meaningful set of weights may be needed.
- It was possible to obtain more accurate results for average amounts of benefits by using imputed rather than reported values among the non-consenters in the micro-data. This did, however potentially remove some correct information from non-consenters (replacing the survey value for non-consenters with an imputed value based on the consenters). It also meant that a lot of data collected from non-linked respondents was not being used.
- Further analysis imputing the actual receipt of benefit by logistic regression appeared to give good results (imputing hidden recipient status among the non-consenters using logistic regression approaches). However, it is a large step to provide datasets with imputations in such circumstances.

These points imply that an imputation-based approach is most likely to be fruitful for establishing a new dataset or conducting certain kinds of analysis (particularly multiple imputation for specific analytical questions). The simpler approach of just using administrative plus survey data also seems to be a very practicable approach that generally outperforms a complete cases approach (with or without any re-weighting), and does not compare too badly against more sophisticated approaches. However, the limitations of each approach and their overall consequences, discussed in the next section need to be considered.

It is also worth noting that working-age means-tested benefits and Tax Credits are now being reformed and will be replaced with Universal Credit, starting from 2013. While the results of this analysis seem robust, it would certainly be worth looking at whether they hold for Universal Credit, as this is a very major reform of the system. That point aside, there has now been a reasonable amount of research on data linking, and it seems time to suggest ways forward.

5.1 Ways forward

This research has been based on analysis of a particular range of social security benefits. It is worth cautioning that the results that apply to this particular set of benefits may not apply to other benefits nor indeed to other sources of income and particularly to earnings. While there is confidence that linkage to administrative data will improve measurement of income from benefits, it does not follow that better information on benefits generates better overall estimates of total incomes. If benefit income tends to be under-reported, and income from earnings tends to be over-reported (for instance), then it could not be concluded that a better measure of benefits leads to a better measure of incomes. Even so, for many purposes it is attractive to have the best possible measures of benefit income for example, for estimating levels of take-up, or analysing the effects of changes to particular benefits on overall spending on benefits.

The central problem of using available administrative earnings data is that it is measured on an annual basis whereas the income concepts in the FRS/HBAI are based on a much shorter time window. The introduction of HMRC introduction of Real Time Information (RTI) earnings data, which will be used within Universal Credit entitlement assessments, offers the opportunity to have a measure of earnings that is more in line with current practice. Measuring incomes on an annual basis would be likely to show lower levels of inequality than incomes on a monthly basis.

There are a number of ways of responding to issues of consent and the high proportion of data that still needs to be collected from the respondents directly. The analysis presented above, in common with that from a small number of other projects, suggested that there is not a single approach that always provides the optimum solution. However, generally speaking each of the main approaches analysed provided an improvement on the 'do nothing' option – at the expense of greater analytical effort required.

The 'complete cases' approach was a good benchmark to establish. The inferences generated from models that only use all valid cases were likely to be different than those from the true population – unless the data was MCAR. That assumption seems unlikely to hold even if the extent of 'consent bias' was relatively small compared with the other kinds of error that affect survey data. A further attraction of this approach was that it required the least further analysis to implement. It possibly provided a convenient set of results for some analyses, perhaps being shown in an annex rather than in the main substantive analytical sections. Levels of benefit were slightly under-reported, so rates of poverty will appear to be somewhat lower using the linked data or when using complete cases only.

Any approach that goes further than looking at complete data (the consenters) will require a number of judgements to be made about appropriate models – whether they are imputation or selection models. Models based on imputation are likely to be the most appropriate. It is worth noting that any approach taken within DWPs towards a new dataset is likely to be followed by most outside analysts. Few, if any, have ever tried to use different imputation methods for reported benefit amounts even though this is clearly possible from the data that is distributed.

Distributing a dataset with a single set of imputed values is rather more practical than multiple datasets with different imputations, even if multiple imputation is to be preferred on theoretical grounds and provides greater information on the uncertainty generated by the imputation process. In the analysis above, models based on imputing the levels of benefits for non-consenters generally provided good results. Further imputation of hidden recipients by different techniques such as logistic regression also moves results in the right direction. However, it is a much bigger analytical step to attribute benefits on a statistical basis to those who said they didn't receive them. The above analysis was mostly based on correcting amounts of benefits among those who declared receipt.

For the results in Section 4.2 there was further analysis and logistic regression style imputation of hidden receipt of benefits was conducted – but there was no work attempting the difficult task of identifying false recipients.[5] For most benefits of interest, hidden recipients tend to be rather more numerous than false recipients – hence the general tendency of FRS to undercount benefit receipt.

An approach based on retaining the survey data for non-consenters, and using the administrative data for consenters, seems to provide a good balance of accuracy with only limited analytical time being needed to manipulate the data in various ways. This kind of analysis would be a worthwhile addition to existing analyses, and provides a good idea of the kinds of changes that are possible when using administrative data.

When it is possible to have earnings data on a comparable basis to the current measurement of average and low incomes, it would make sense to invest in an imputed general purpose dataset. However, the gains to be had from distributing a dataset with imputation of benefit amounts only are less clear. In the meantime some results (e.g. on take-up) may well be improved by having access to administrative data. Once Universal Credit is established as the main benefit for those of working age, and assuming that increases the availability of earnings data, there will be a good opportunity to construct a suitably imputed and linked dataset.

[5] Preliminary, quite positive, results of doing this were presented at DWP in December 2011, and the main SAS programme that was part of the outputs includes the code for implementing this kind of imputation.

Appendix A
Models of receiving benefit

In this section we report some fuller details of the regression models described in Chapter 4.

Table A.1 Comparing logistic regression results: Receipt of Pension Credit

Logit coefficients

Variables	Population (bench-mark)	Existing FRS	Admin + Survey	Admin + (Survey + Imputed)	Consenters only	Consenters re-weighted
Intercept	-1.632	-1.973	-1.812	-1.5677	-1.410	-1.414
Aged 60-65	-0.325	-0.214	-0.334	-0.344	-0.370	-0.358
Aged 70-74	0.246	0.310	0.256	0.2891	0.109	0.088
Aged 75-80	0.439	0.284	0.364	0.4017	0.279	0.276
Aged 80-85	0.774	0.728	0.790	0.8757	0.639	0.622
Aged 85-90	1.138	0.933	1.051	1.199	0.973	0.944
Aged 90+	1.475	1.120	1.072	1.4841	0.895	0.885
Male	-0.214	-0.315	-0.290	-0.284	-0.326	-0.323
RMSE		0.205	0.164	0.058	0.247	0.255

The fit for Pension Credit is relatively good using imputed values of receipt. This may be reflecting the ability of the imputation model to identify a reasonably accurate proportion of otherwise ‘hidden recipients’.

Table A.2 Comparing logistic regression results: Receipt of JSA

Logit coefficients

Variables	Population (bench-mark)	Existing FRS	Admin + Survey	Admin + (Survey + Imputed)	Consenters only	Consenters re-weighted
Intercept	-3.677	-4.081	-4.014	-3.9031	-3.904	-3.905
Aged 18-24	0.490	1.036	0.932	0.9824	0.849	0.880
Aged 35-44	-0.218	-0.170	-0.173	-0.1404	0.085	0.080
Aged 45-49	-0.304	-0.184	-0.119	-0.1008	0.043	0.087
Aged 50-54	-0.407	-0.232	-0.127	-0.053	0.037	0.040
Aged 55-59	-0.561	-0.370	-0.322	-0.299	-0.146	-0.144
Aged 60-64	-2.320	-2.122	-1.966	-2.0221	-2.162	-2.166
Male	0.850	1.083	1.029	1.0226	1.074	1.057
RMSE		0.282	0.282	0.286	0.323	0.333

Table A.3 Comparing logistic regression results: Receipt of Income Support

Logit coefficients

Variables	Population (bench-mark)	Existing FRS	Admin + Survey	Admin + (Survey + Imputed)	Consenters only	Consenters re-weighted
Intercept	-2.533	-2.947	-2.993	-2.834	-2.707	-2.730
Aged 18-24	-0.300	0.050	-0.018	0.049	0.008	-0.009
Aged 35-44	-0.017	0.177	0.188	0.223	0.148	0.193
Aged 45-49	-0.124	0.089	0.139	0.227	0.114	0.157
Aged 50-54	-0.095	0.017	0.166	0.146	0.127	0.182
Aged 55-59	-0.007	0.217	0.287	0.344	0.256	0.298
Male	-0.676	-0.757	-0.752	-0.761	-0.702	-0.727
RMSE		0.253	0.284	0.288	0.217	0.245

Table A.4 Comparing logistic regression results: Receipt of DLA (either component)

Logit coefficients

Variables	Population (benchmark)	Existing FRS	Admin + Survey	Consenters only	Consenters re-weighted
Intercept	-2.581	-2.662	-2.694	-2.748	-2.763
Aged 18-24	-1.075	-1.247	-1.236	-1.602	-1.602
Aged 25-29	-1.137	-1.054	-1.169	-1.110	-1.102
Aged 30-34	-0.966	-0.917	-1.010	-0.952	-0.976
Aged 35-39	-0.694	-0.673	-0.830	-0.727	-0.719
Aged 40-44	-0.460	-0.279	-0.300	-0.128	-0.115
Aged 45-49	-0.234	-0.214	-0.252	-0.065	-0.063
Aged 55-59	0.260	0.153	0.110	0.437	0.430
Aged 60-64	0.497	0.418	0.455	0.644	0.663
Aged 65-69	0.605	0.461	0.503	0.679	0.688
Aged 70-74	0.450	0.484	0.533	0.491	0.486
Aged 75-79	0.112	-0.092	-0.016	-0.035	-0.037
Aged 80-84	-0.561	-0.186	-0.408	-0.663	-0.674
Aged 85+	-1.696	-0.408	-0.881	-3.061	-3.004
Male	-0.046	-0.132	-0.096	-0.072	-0.079
RMSE		0.362	0.236	0.400	0.389

Appendix B
Outline of running multiple imputation in SAS

1. Imputation. Fill in the missing values with multiple imputations.

   ```
   PROC MI

   DATA=/*data set with missing values*/

   OUT=/*data set with values imputed*/

   NIMPUTE=/*# of imputations per missing value*/;

   VAR /*...variables in imputation model...*/;

   RUN;
   ```

2. Analysis. Next fit models as if the data were complete but with the BY statement to fit separate models for each version of the dataset.

   ```
   PROC/*REG or LOGISTIC or...*/

   DATA=/*imputed data set*/

   MODEL /*dependent variable*/ = /*independent variables*/;

   ODS OUTPUT

   /*parameter estimate keyword*/=parameters

   /*parameter covariance keyword*/=parameter_covariances;

   BY _IMPUTATION_;
   RUN;
   ```

   The ODS statement creates a new data file of relevant parameter estimates for each imputed data set.

3. Synthesis. The final step combines the results from the different imputed data sets. In the previous step these were saved into data sets called parameters and parameter_covariances

   ```
   PROC MIAnalyze

   PARMS=parameters

   COVB=parameter_covariances;

   VAR intercept /*regressors*/ ;

   RUN;
   ```

   The output is a single set of estimates and standard errors, as well as confidence intervals and t tests. The standard errors account for the variation across imputed data sets, as well as the usual sampling variation.

Source – edited version of SAS help system notes.

References

Allison, P. (2002). *Missing Data*. Thousand Oaks, CA: Sage.

Barton, A., Riley, K (2012). *Income Related Benefits: Estimates of Take-up in 2009-10*, 7.3, p. 153: DWP.

Clay, S., Herring, I., Metcalf, S., Sullivan, J. and Vekaria, R. (2011). *Family Resources Survey United Kingdom*, 2009-10 London: DWP.

Crockett, A. (2011). *Weighting the Social Surveys*, Essex: UK Data Archive and Institute for Social and Economic Research (Updated by: Reza Afkhami, Anthony Rafferty, Vanessa Higgins, Alan Marshall) Version: 1.9.

Deville, J.-C. and Särndal, C.-E. (1992). 'Calibration estimators in survey sampling'. *Journal of the American Statistical Association* 87: 376-382.

DWP (2005). *The New Family Resources Survey Grossing Regime* London: DWP (currently available from archive at:
http://webarchive.nationalarchives.gov.uk/+/http://www.dwp.gov.uk/asd/frs/reports/new_grossing_regime.pdf

Durrant, G. (2009). 'Imputation methods for handling item nonresponse in practice: methodological issues and recent debates', *International Journal of Social Research Methodology*, 12:4, 293-304.

Fellegi, I. P. and Holt D. (1976). 'A Systematic Approach to Automatic Edit and Imputation'. *Journal of the American Statistical Association* Vol. 71, No. 353 (Mar., 1976), pp. 17-35.

Heckman, J. (1979). 'Sample selection bias as a specification error', *Econometrica* 47(1) pp 153-161.

Horton, N.J., Lipsitz, S.P., and Parzen, M. (2003). 'A potential for bias when rounding in multiple imputation'. *The American Statistician* 57(4), 229-232.

Kho, M., Duffett, M., Willison, D., Cook, D. and Brouwers, M. (2009). 'Written informed consent and selection bias in observational studies using medical records: systematic review', *British Medical Journal*: 822.

Ibrahim, J.G., Chen, M.H., Lipsitz, S.R., and Herring, A.H. (2005). 'Missing-data methods for generalised linear models: A comparative review'. *Journal of the American Statistical Association*, 100(469), 332–346.

Kalton, G. and Kasprzyk, D. (1986). 'The Treatment of Missing Survey Data'. *Survey Methodology*, 12, 1-16.

Krug, G. (2009). *Fehlende Daten beim Record Linkage von Prozess- und Befragungsdaten : ein empirischer Vergleich ausgewählter Missing Data Techniken* (Missing data in the record linkage of process and survey data : An empirical comparison of selected missing data techniques) Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg [Institute for Employment Research, Nuremberg, Germany] IAB Discussion Paper number 200907.
http://doku.iab.de/discussionpapers/2009/dp0709.pdf

Landy, R. (2012). *A simulation of 1946 birth cohort data using three different missing data mechanisms- complete case, multiple imputation and Heckman selection*. Presentation at Methodological challenges associated with Non-Response and Missing Data in ageing populations Friday 3 February 2012, Institute for Social and Economic Research (ISER), University of Essex
http://www.methods.manchester.ac.uk/ageingcohort/seminar-3/Landy.pdf

Little, R.J.A. and Rubin, D.B. (1987). *Statistical analysis with missing data*. New York, Wiley.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

Lumley, T. (2010). *Complex Surveys: A guide to analysis using R*. New Jersey: Wiley.

Sakshaug and Kreuter, *Assessing the Magnitude of Administrative Non-Consent Biases in the German PASS Study*, draft paper.

Sande, I. G. (1982). 'Imputation in Surveys: Coping with Reality'. *The American Statistician* Vol. 36, No. 3, Part 1 pp. 145-152.

This report provides an investigation into evaluating approaches to linking Family Resources Survey (FRS) data with Department for Work and Pensions (DWP) administrative data.

Measurement error exists on a survey when respondents do not report their true status. The purpose of data linking examined here is to use administrative data to correct measurement error.

Minimising measurement error will improve accuracy for the FRS. This is important because it has two important functions:

- supporting the production of key National statistics such as Households Below Average Income (HBAI); and
- a dataset that is available for use by both Government policy analysts and external researchers.

The report also assesses the extent of consent bias and methods of dealing with any resulting bias.

If you would like to know more about DWP research, please contact:
Carol Beattie, Central Analysis Division, Department for Work and Pensions,
Upper Ground Floor, Steel City House, West Street, Sheffield, S1 2GQ.
http://research.dwp.gov.uk/asd/asd5/rrs-index.asp

DWP Department for
Work and Pensions

Published by the
Department for Work and Pensions
June 2012
www.dwp.gov.uk
Working Paper no. 110
ISBN 978-1-908523-63-1