Variation within Households in Consent to Link Survey Data to Administrative Records: Evidence from the UK Millennium Cohort Study

Tarek Mostafa[1]

Abstract

This study expands our knowledge of consent in linking survey and administrative data by studying respondents' behaviour when consenting to link their own records and when consenting to link those of their children. It develops and tests a number of hypothesised mechanisms of consent, some of which were not explored in the past. The hypotheses cover: parental pride, privacy concerns, loyalty to the survey, pre-existing relations with the agency holding the data, and interviewer effects. The study uses data from the longitudinal Millennium Cohort Study to analyse the correlates of consent in multiple domains (i.e. linkage of education, health and economic records).

The findings show that respondent's behaviour vary depending on the consent domain and on the person within the household for whom consent is sought. In particular, the cohort member's cognitive skills and the main respondent's privacy concerns have differential effects on consent. On the other hand, loyalty to the survey proxied by the longitudinal response history has a significant and strong impact on consent irrespective of the outcome. The findings also show that interviewers account for a large proportion of variations in consent even after controlling for the characteristics of the interviewer's assignment area. In total, it is possible to conclude that the significant impact of some of the correlates will lead to sample bias which needs to be accounted for when working with linked survey and administrative data.

Keywords: Informed consent; data linkage; multivariate probit models; UK Millennium Cohort Study; sample bias.

Published in the *International Journal of Social Research Methodology*, Volume 19, N 6, p. 355–375.

[1] Department of Quantitative Social Science, Centre for Longitudinal studies, Institute of Education. This research was funded through the ESRC/NCRM research grant number: R00960-01-001-1. Email: t.mostafa@ioe.ac.uk.
The author would like to thank John Micklewright, Lucinda Platt, Richard Wiggins and two anonymous referees for their valuable comments.

I- Introduction

Household surveys are increasingly being linked to administrative records with the potential of greatly enriching survey content on subjects such as health, education and income. One major challenge to data linkage is non-consent. Non-consent occurs when respondents refuse permission to link their administrative records to their survey data. This problem leads to a reduction in sample size for the administrative data concerned and more worryingly to differential patterns of consent and possibly bias if consent is correlated with the characteristics of the respondents.

In the existing literature, consent was found to be related to the characteristics of the respondents, the interview features (e.g. wording, sequencing of questions, etc), and the characteristics of interviewers (Jenkins et al. 2006; Sakshaug et al. 2012; Sala et al. 2012). However, despite these recent developments the evidence is still scarce. Most of the existing research draws upon medical and epidemiological investigation and few studies focus upon multipurpose social surveys (e.g. Jenkins et al. 2006; Sala et al. 2014; Sakshaug et al. 2013).

This paper aims to advance our knowledge about consent by analysing adult respondents' behaviour when consenting to link their own administrative records in contrast to their behaviour when consenting to link someone else's records (i.e. another member of the household). These variations in consent behaviour have not been explored in the past. All previous studies focused on respondents consenting to link their own records but not those of other members of their household. The paper uses data from the UK Millennium Cohort Study (MCS) to answer the following research questions:

> **RQ1:** Do respondents behave differently when consenting to link their own administrative records in comparison to consenting to link those of their children?
> **RQ2:** Does respondents' consent behaviour vary according to the domain of consent, e.g. health, economic, education records?
> **RQ3:** What is the impact of interviewers on consent outcomes and can interviewer effects be separated from the impact of an interviewer's geographical assignment?

Furthermore, I set out to test a number of hypothesised mechanisms of consent: parental pride, privacy and data confidentiality, loyalty to the survey, pre-exiting relations with the agency holding the administrative data, and the impact of interviewers.

The findings show that non-consent in MCS ranges between 6% and 20% depending on the consent domain. Moreover, consent behaviour varies according to the person for whom consent is sought (i.e. main respondent (MR) vs. cohort member (CM)), and according to the domain of consent. In addition to this, interviewer characteristics explain a large proportion of the variation in consent even after accounting for the effect of the interviewer's assignment area.

The paper is organised as follows. Section II presents the existing literature, section III presents the theory and hypothesised mechanisms of consent; section IV describes the Millennium Cohort Study (MCS) and the methodology; section V presents the results and the last concludes.

II- Previous Literature.

Most of the existing literature on consent comes from the medical profession. In these studies (Baker et al. 2000; Dunn et al. 2004; Nelson et al. 2002; Kho et al. 2009; Silva et al. 2002; Huang 2007) consent was sought from patients to access their medical records. The main focus was to ascertain whether non-consent is influenced by patients' characteristics and whether it leads to sample bias. Most of these studies relied on binary single-equation consent models. However, in recent years a number of studies dealing with consent in complex social surveys have emerged: Jenkins et al. 2006; Sala et al. 2012; McKay 2012; Knies et al. 2012, Sakshaug et al. (2012, 2013); Sakshaug and Kreuter 2012; Korbmacher and Schroeder 2013; Kreuter and Sakshaug 2014. These studies explored consent across multiple domains and used new methods to jointly estimate consent questions.

Jenkins et al. (2006) constitutes the first major contribution to the analysis of consent in a non-medical survey. The authors analysed the impact of respondents' characteristics and interview features on the propensity to consent on four different economic domains. The methodology is also innovative as the authors used a multivariate probit procedure to jointly model consent questions. The authors found that non-consent is a source of bias and that the correlates of consent may vary across the different domains. They also argued in favour of the joint modelling of consent questions.

The choice of correlates was further expanded in Sala et al. (2012) to include the characteristics of interviewers. In this study, the authors used data from the British Household Panel Study with two consent outcomes: health and benefits. The authors included interviewers' characteristics such as their personality, attitudes to persuading respondents, and survey experience. They found a positive impact for survey experience and task specific experience. Similarly, Korbmacher and Schroeder (2013) measured the effect of interview and interviewer characteristics on the likelihood of consent using a multilevel model with respondents nested within interviewers. They found that interviewers account for a larger proportion of variation in consent in comparison with respondent socio-demographic characteristics.

On the other hand, the study by Sakshaug et al. (2012) explored consent along a number of hypothesised mechanisms covering privacy concerns, inaccurate recalling of past information, resistance towards the interview, and interviewer behaviour. They found strong support for the privacy and interview resistance hypotheses. Respondents having more concerns about data confidentiality and those with higher levels of resistance were found to be less likely to consent. A further study by Sakshaug and Kreuter (2012) examined the

magnitude of non-consent bias in linked administrative and survey data to find limited evidence for the existence of such bias.

More recently, two studies provided experimental evidence on the impact of consent question wording and placement. Sakshaug et al. (2013) examined the impact of question wording, question placement, and interviewer attributes. They found that question length did not affect the likelihood of obtaining consent. In contrast, the placement of the question in the beginning of the interview had a positive effect and interviewers who themselves would consent to data linkage were more successful in obtaining consent. Similarly, Sala et al. (2014) found that the likelihood of consent varies according to the placement of the question and that reminding those who have consented previously of their answer (i.e. dependent question) prompts them to make the same decision.

Various socio-demographic characteristics have been found to have an effect on consent even though the sign, magnitude and significance of these effects varied between studies. The propensity to consent is found to be significantly related to age, gender, and health. Older men with poorer health and ethnic majority respondents are more likely to consent (Woolf et al. 2000). Dunn et al. (2003) found similar results with higher propensities to consent among males and patients with health conditions. However, they found that younger respondents are more likely to consent than older ones. Similarly, in their review of 17 medical research reports, Kho et al. (2009) found conflicting evidence. Age has a significant effect on consent only in seven studies and women are less likely to consent only in four. However, since all these studies are focused on patients, their findings might not necessarily be valid for the general population.

In addition to respondents' socio-demographic characteristics, some studies paid attention to the respondents' personality traits such as altruism, being a private person, and having a stronger perception of risk. Consent is found to be lower among respondents who refuse to answer income questions (Sala et al. 2012; Jenkins et al. 2006; Olson 1999; Woolf et al. 2000) and among those who have fears about the confidentiality of the information they provide (Armstrong et al. 2008).

Despite the recent developments, the literature still contains a number of gaps. All the aforementioned studies dealt with consent sought from respondents for linking their own records. This paper goes beyond the existing literature by considering the case where consent is sought from respondents for linking their own records and for linking the records of someone else (i.e. their children: the cohort members in the MCS). In addition to this, a number of hypothesised mechanisms of consent are developed and tested. In particular, the differential impact on consent of the attributes of the child (cognitive skills and health) and those of the respondent (being private, loyalty to the survey, pre-existing relations with the agency holding the administrative data, and various socio-demographic characteristics) is measured.

From a methodological perspective, with the exception of Jenkins et al. (2006) and Sala et al. (2012), all of the other studies included in this literature review have modelled consent questions separately rather than jointly. Some studies only presented consent rates and break-downs by socio-demographic characteristics (Olson 1999; Gustman and Steinmeier 1999; Haider and Solon 1999). In this paper, consent questions are jointly modelled using a multivariate probit procedure which takes into account the complex design of MCS. The study focuses on four consents: a) MRs' consent to link their own health and economic records, b) MRs' consent to link the CM's health and education records.

III- Theory and Hypotheses.

I argue that six key influences affect an individual's likelihood to consent. In the case where the main respondent is asked to agree to consent for their offspring, the first major influence is their personal pride derived from their children's abilities. Other influences are the respondent's concerns regarding privacy and confidentiality, the influence appertaining to their own 'loyalty to the study' and their existing relationship with the agency holding their administrative records. In addition to these, the socio-demographic characteristics of respondents and the survey interviewers are also expected to influence consent.

Parental pride

Parents like to talk about their children. Previous studies found that children's success influences different aspects of parental wellbeing and behaviour (Birditt et al. 2010, Fingerman 2012). In this study, I hypothesize that cognitive abilities of young children might influence their parents' likelihood to consent. However, I expect that children's abilities will only affect parental consent for linking their children's records but not their own records. Moreover, I expect that the effect will be positive and higher in magnitude on consent for linking the CM's education records in comparison with consent for linking the CM's health records.

Note that this hypothesis has not been explored in any of the previous studies. I use a composite indicator of cognitive abilities at age 5 which is an arithmetic average of two scores: one on naming vocabulary and the other on pattern construction. If this hypothesis is true, then it is possible to conclude that the linked education records suffer from sample bias since these contain the performance scores of the CMs which are known to be highly correlated with cognitive skills.

Privacy and data confidentiality

One of the frequently assessed hypotheses is whether or not concerns about protecting individual information (privacy) affect consent. Respondents who are more concerned about a potential breach of confidentiality are expected to be less likely to consent. Previous studies (Singer et al. 2002 and Jenkins et al. 2006) have used income item non-response (Sakshaug et al. 2012 used a composite measure of refusals on five financial questions) as a measure of unwillingness to provide sensitive financial information while Sala et al. 2012 used a measure

of trust. The drawback of using income item non response are: first, it is a binary variable that hides variations in the willingness to provide information; and secondly it is focused on the provision of financial information rather than the provision of information in a broader sense.

This study uses a direct measure of privacy as a general predisposition or personality trait instead of using a proxy measure. In this instance MCS has a Likert scaled item which asks the MR to agree/disagree with the following statement: "I am a very private person". The possible answers are: strongly agree; agree; neither; disagree; strongly disagree; and can't say. The three categories: refusal; don't know; and not applicable were combined into one category called 'other'.

The impact of the privacy measure is expected to vary when respondents consent for linking their own records and when they consent for linking those of the CMs (RQ 1). It is possible that parents might be more protective of their children and therefore the privacy measure might have a higher impact on the CMs consent outcomes. Conversely, since the CM is the focus of the survey, respondents may feel more inclined to link the CM's records than their own. In this case the privacy measure will have a greater effect on the respondent's own outcomes.

Loyalty to the survey

The working assumption is that respondents who have missed a wave of data collection in the past (i.e. in a longitudinal survey) are less committed to the survey and less likely to cooperate with the future in-survey requests (i.e. consent in this case). Since most studies (except Sala et al. 2012) used cross-sectional datasets, it was impossible to test whether previous non-response can be symptomatic of a lack of a continued commitment to the survey.

The longitudinal data available under MCS provides a record of response co-operation over four waves. In this paper, I use a response history indicator which takes a value of 1 if a respondent failed to co-operate at least once in the previous three waves and zero otherwise (obviously all respondents were productive in wave 4 from which the consent outcomes are taken). A binary variable instead of a continuous one measuring the number of missed waves is used, because very few respondents missed two waves (180 respondents) and none missed three.

Pre-exiting relations with the agency holding the administrative data.

One of the reasons why respondents might consent to a specific data linkage is because they already receive services or benefits from the agency holding the data (Sakshaug et al. 2012). Indeed, Dunn et al. (2004), Woolf et al. (2000), and Petty et al. (2001) found that respondents suffering from health problems are more likely to consent to follow-up interviews, and to health data linkage. However, as noted above, the mechanisms of consent in social surveys might differ from those in medical ones.

This paper uses two measures to proxy this pre-existing relation. First, self-reported health for the respondent and for the CM is used as a proxy for the receipt of health services. Secondly, the receipt of benefits is accounted for using a binary variable that takes the value of 1 if the respondent is receiving benefits. Child benefits were excluded from this variable since almost all families were eligible. It is expected that the impact of self-reported health (for the respondent and for the CM) will vary depending on whether the respondent is consenting to link his/her records or those of the CM. The receipt of benefits is expected to have a positive impact across outcomes given that it accounts for a wide range of different types of benefits.

Interviewers

Recent studies such as Sala et al. (2012) and Sakshaug et al. (2012) have devoted more attention to the impact of interviewers on variations in consent. Interviewers are charged with administering the consent questions, explaining what consent to data linkage is, and what the consequences of consent are. However, given that interviewers are incentivised to minimize unit non-response, consent is not usually a main preoccupation. Therefore, it is unclear how interviewers might influence consent. Sakshaug et al. (2012) note that interviewers' attitudes toward data confidentiality will influence their likelihood of obtaining consent. Further, Sala et al. (2012) have shown that some interviewer characteristics such as survey experience do have an impact on the likelihood of consent.

In the MCS, interviewer characteristics are not available. However, interviewer identifiers are available and can be used in fixed effects models. One of the challenges facing the interpretation of these effects is the ability to separate the effect of the interviewers themselves from the effect of interviewer area assignment.

Socio demographic background

Apart from the previously mentioned correlates, other controls are included in the analyses. These are: the CM's gender, MR's social class, ethnicity, religion, age, marital status, number of siblings in the household, whether the interview is translated, and log OECD-adjusted income. All these socio-demographic variables come from the same MCS survey as the consent outcomes (i.e. wave 4).

IV- Data, Consent Procedures, and Methods.

The Millennium Cohort Study wave 4

The Millennium Cohort Study (MCS) is a longitudinal survey following a nationally representative, clustered and stratified sample of 19,000 children born in the UK in 2000-01. The sample was drawn from all babies born between 1st September 2000 and 31st August 2001 in England and Wales; those born in Scotland and Northern Ireland between 23rd November 2000 and 11th January 2002. It was selected from a random sample of electoral wards, disproportionately stratified to ensure adequate representation of all four UK

countries, of deprived areas and areas with high concentrations of Black and Asian families. MCS has been tracking the CMs since the age of nine months and survey data has been collected on five different occasions (i.e. age nine months, three, five, seven, and eleven years). In this paper, all consent outcomes are from the age 7 survey (wave 4). The MCS has a complex design, the sample is stratified by country (i.e. England, Scotland, Wales, and Northern Ireland), clustered at the electoral ward level, and has oversampled minorities and disadvantaged groups. In addition to this, the sample has experienced attrition over time. The number of families ever interviewed was 19,244 (some having more than one child, i.e. twins and triplets) and in wave 4 only 14,044 children participated (see the MCS technical report on response). All these features (i.e. stratification, clustering, oversampling, and unit non-response) are taken account of through the use of the *svy* procedures in Stata (see the user guide to analysing MCS data using STATA). The analytical sample consists of 14,044 respondents interviewed by 443 interviewers.

Consent Procedures

Written consent was sought for gathering information from health, education and economic records for the MRs, and for the CMs. All consent questions were answered by the MR (in most cases the mother). Hence, MRs were in charge of consenting to link their own records and those of their children. The consent outcomes are presented in Table 1.

Table 1: Consent domains.

Consent (all from MCS wave 4)	Notes
CM's Health records	Consent for linking health records (hospital admissions and records held by the NHS) from birth to age 14.
CM's Education records	Records held by Educational authorities (e.g. Department for Education in England). See MCS Guide to the Linked Education Administrative Datasets (2007).
MR's Health records	Hospital admissions and records held by the NHS.
MR's Economic records	Records held by the Department for Work and Pensions (DWP) and Her Majesty's Revenue and Customs (HMRC).

Leaflets describing data linkage and the need for consent were sent in advance of the survey. All interviews were face to face and consent forms were administered at the end of the main interview. The wording of the consent questions was the same for all respondents. The two consent questions for linking the CMs records were administrated on the same consent form and similarly for the two questions for linking the MRs own records. The procedures, the leaflets and consent forms are presented in detail in the technical report on Ethical Review and Consent (2012).

Respondents who were willing to give consent were asked to tick an endorsement box (simply containing two possibilities ‘yes’ or ‘no’) sign, print their names and date their signature. As with all parts of the survey, it was made clear to the respondents that they can refuse to participate in any element or withdraw from the study at any time by simply

expressing the wish to do so (See the Millennium Cohort Study, Ethical Review and Consent 2012).

Moreover, it was possible to give one consent but withhold another when the same form had multiple consent questions. This was done for ethical reasons and because each consent was regarded as an independent decision. In MCS, none of the consents were conditional on other consents being given as it was the case in Jenkins et al. (2006).

Methods

If respondents do hold a latent propensity to consent (Jenkins et al 2006) then consents are likely to be correlated irrespective of their domain. The correlations are also reinforced by the fact that the circumstances surrounding the interview are the same for all domains (since these are sought during the same interview). Put differently, those who consent on one domain are expected to consent on the others with higher probability than other respondents. The gaps in the theory and in the empirical literature warrant the examination of the association between consents across domains, and across different individuals for whom consent is sought (i.e. MRs and CMs). Therefore, the consent domains in this paper are modelled jointly.

Unlike univariate and bivariate probit models, multivariate probit models can handle more than two consent questions and the only limitation to their use is the rise in computational time with the inclusion of more questions. This estimation approach will allow us to measure the strength of the association between consent domains and its significance. The M-equation multivariate probit model is the following:

$$y_{im}^{*} = \beta_{im}' x_i + \varepsilon_{im}, m = 1, \dots, M$$

$$y_{im} = 1 \text{ if } y_{im}^{*} > 0 \text{ and 0 otherwise}$$

where y is the binary consent outcome for respondent i and consent outcome m with $m = 1, \dots, 4$. x is a vector of independent variables for respondent i. The x vector is the same for the four equations. ε_{im}, are error terms distributed as multivariate normal, each with a mean of zero and a variance-covariance matrix V, where V has values of 1 on the diagonal and values different to 1 off-diagonal (Cappellari and Jenkins 2003).

The Rho (ρ) elements measure the correlations of the unobserved factors for each combination of two consent domains (Jenkins et al. 2006). A significant Rho indicates that the domains are associated and therefore modelling them jointly produces more efficient results than univariate probit models.

The estimation of multivariate probit models is computationally intensive. In this paper, the model is estimated using a simulated maximum likelihood procedure with 50 Halton draws plus antithetic draws (100 draws in total) and 10 initial sequence elements dropped in each dimension. This procedure reduces the computational time and is more accurate than 1000 pseudorandom draws since it produced the same estimates but with lower standard errors (Cappellari and Jenkins 2006, p.174). The procedure of Cappellari and Jenkins (2006) was

adapted to take into account the survey features of MCS through the use of the *svy* command in Stata. The MCS features are: clustering at the electoral ward level, stratification at the country level, oversampling of minorities and disadvantaged groups in the base sample and attrition over time. Oversampling and attrition were accounted for through the use of sampling and unit non-response weights. (See the MCS Technical Report on Response, the MCS technical report on sampling and the MCS user guide on analysing MCS data in Stata).

Note that this model includes all aforementioned correlates except the interviewer effects. The reason is that multivariate probit models become very complex and computational time rises dramatically when there are a large number of fixed effects to be accounted for: in this case 443. Moreover, fixed effects cannot be included in non-linear models (Gianelli and Micklewright, 1993). Therefore, interviewers' effects were included in separate linear probability models for each consent domain.

In attempting to account for interviewer effects on consent, the aim is to measure any improvement in the explanatory power of the model (i.e. a rise in R-squared). However, any change cannot be completely attributed to the impact of interviewers simply because the allocation of interviewers to interviewees is typically implemented on a 'nearest-to-home' basis. Therefore interviewer effects will be confounded by geography. Geographical areas could have specific characteristics such as being relatively poor, having large proportions of minorities, having high levels of unemployment, etc. In order to overcome this challenge, four different models are estimated:

Base model: is a linear probability model with the aforementioned correlates and without interviewer fixed effects.

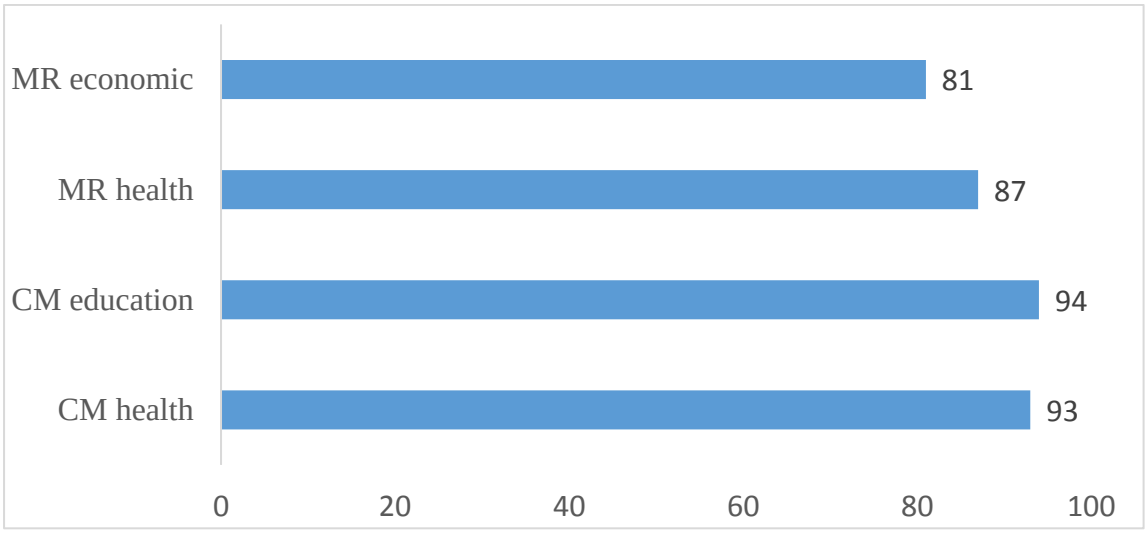
Model with assignment area characteristics: is identical to the base model and includes additional variables measuring the assignment area characteristics (i.e. proportion of minorities, proportion unemployed, log average income, and social class composition). These were computed as averages of MRs' characteristics at the level of the interviewer and were adjusted to take into account the changes in sample composition due to attrition through the use of the MCS attrition weights.

The fixed effects model: is equivalent to the base model and includes interviewer fixed effects.

V- Results.

Consent and sample characteristics

Figure 1: Consent rates (%) for the four domains.



CM stands for 'cohort member' and MR for 'main respondent'. Sample size 14,044.

Figure I shows that the consent rate for linking the CM education records is the highest (94%) followed by consent rate to link the CM health records (93%). In comparison, consent rates for linking the MR's health and economic records are lower (87% and 81% respectively).

A number of observations can be made. First, overall MRs are more likely to consent to linking their children's records than to linking their own records. This indicates that parents at the margin are not more protective of their children than of themselves and may show that they see their child as the main focus of the study. Secondly, consent outcomes are similar when they are sought for the same person (i.e. for the CM across education and health and for the MR across health and economic records). Thirdly, consent rates are the lowest for linking to the MR's economic records suggesting that fears about confidentiality are probably the highest for this domain.

Table 2: Tetrachoric correlation matrix between consent domains.

Domains	CM health	CM education	MR health	MR economic
CM health	1			
CM education	0.99	1		
MR health	0.87	0.83	1	
MR economic	0.79	0.77	0.95	1

In Table 2, tetrachoric correlations between all domains are presented. These correlations measure the degree of association between two binary variables. A high positive value means that if a respondent consented on one domain he/she is likely to consent on the other and the

reverse is true. The correlations are the highest when consent is sought for the same person (i.e. MR vs. CM). Consent for linking the CM's health records is highly correlated with consent to link the CMs education records, and the same is true for the two domains of the MR (highlighted cells). It is also worth noting that the correlation between the MR's health consent and the CM's health consent is also high, indicating that consents are also highly correlated for the same domain. The lowest correlations were for different domains and different persons (i.e. CM health and MR economic, and CM education and MR economic), even though they are still relatively high. The high level of correlations between the domains warrants the use of a joint modelling strategy.

Tables 3a and 3b provide weighted estimates of percentages (and one average) of consenters vs. non-consenters based on the key variables used in the formulation of the hypotheses. MRs who are receiving benefits and who have not missed any wave of data collection are more likely to consent regardless of the domain of consent. When it comes to privacy, those who acknowledge that they are the least private are more likely to consent. Similarly, MRs are more likely to consent to link the CM's education and health records if the CM has higher cognitive abilities. In contrast, consenters and non-consenters do not significantly differ in terms of the CM's and the MR's health statuses.

Table 3a: Characteristics of the sample.

	All respondents	CM health		CM education		MR health		MR economic	
		No consent	Consent	No consent	Consent	No consent	Consent	No consent	Consent
CM's cognitive score (continuous variable)									
Average	16.2	**15.6**	**16.3**	**15.3**	**16.3**	15.9	16.3	16.0	16.3

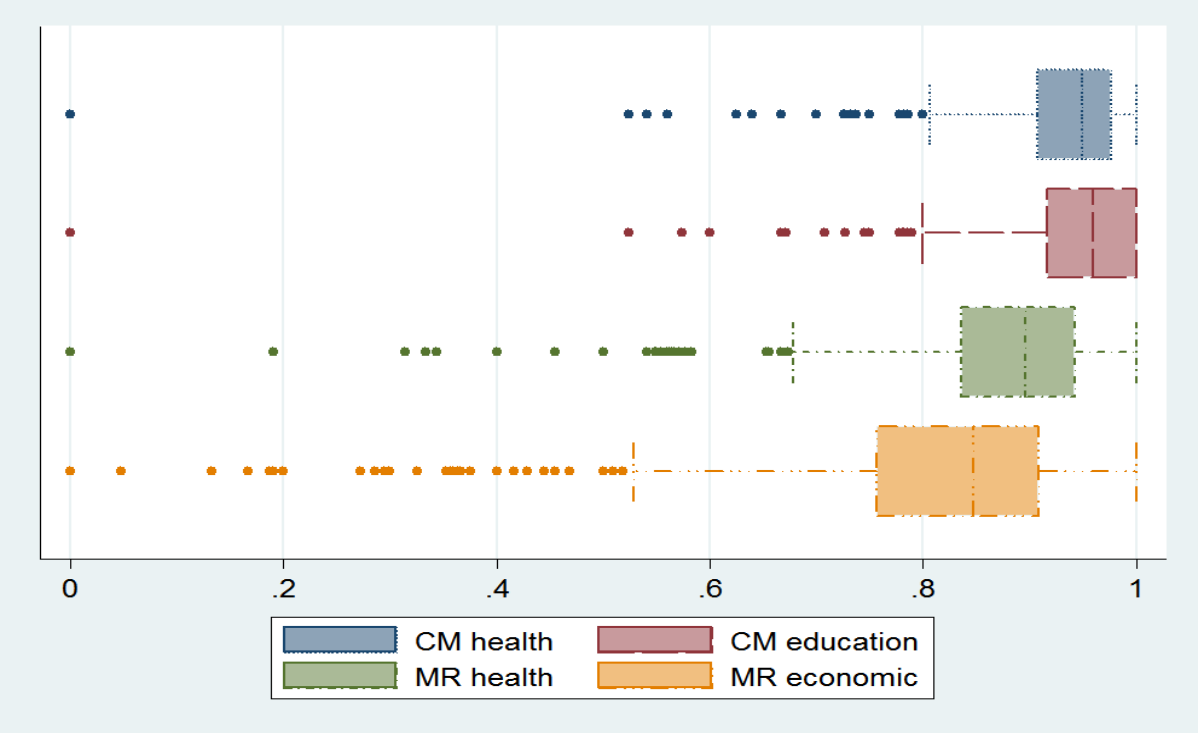
Table 3b: Characteristics of the sample.

	All respondents	CM health (Row %)		CM education (Row %)		MR health (Row %)		MR economic (Row %)	
	(Col %)	No consent	Consent	No consent	Consent	No consent	Consent	No consent	Consent
MR: I am a very private person									
Strongly agree	6.7	**7.4**	**92.6**	**6.3**	**93.7**	**16.1**	**83.9**	**22.2**	**77.8**
Agree	29.9	**7.3**	**92.7**	**5.9**	**94.1**	**14.1**	**85.9**	**19.5**	**80.6**
Neither	25.1	**5.6**	**94.4**	**4.9**	**95.1**	**12.0**	**88.0**	**18.9**	**81.1**
Disagree	28.3	**6.2**	**93.9**	**5.3**	**94.7**	**11.0**	**89.0**	**17.0**	**83.0**
Strongly disagree	4.7	**3.9**	**96.1**	**3.4**	**96.6**	**10.2**	**89.8**	**15.9**	**84.1**
Can't say	1.3	**11.3**	**88.8**	**11.1**	**88.9**	**19.2**	**80.9**	**31.2**	**68.8**
Other	4.0	**22.1**	**77.9**	**21.7**	**78.3**	**31.1**	**68.9**	**36.7**	**63.3**
Dropped out at least once									
Yes	15.7	**10.0**	**90.0**	**8.6**	**91.4**	**18.2**	**81.8**	**24.5**	**75.5**
No	84.3	**6.5**	**93.5**	**5.6**	**94.4**	**12.5**	**87.5**	**18.5**	**81.5**
Benefits									
Yes	64.0	**6.2**	**93.8**	**5.5**	**94.5**	**12.1**	**87.9**	**17.5**	**82.5**
No	36.0	**8.5**	**91.5**	**7.2**	**92.8**	**15.6**	**84.4**	**23.0**	**77.0**
CM health									
Excellent	59.7	6.6	93.4	5.7	94.3	**12.6**	**87.4**	18.7	81.3
Good	37.3	8.0	92.0	6.9	93.1	**14.8**	**85.2**	21.0	79.0
Poor	3.0	4.5	95.5	4.1	95.9	**11.7**	**88.3**	14.8	85.2
MR health									
Excellent	22.0	7.0	93.0	6.5	93.6	12.5	87.6	19.7	80.3
Good	65.0	7.3	92.7	6.2	93.8	13.8	86.3	19.8	80.2
Poor	13.0	5.8	94.2	5.0	95.1	13.2	86.8	17.7	82.3

Comparisons (in bold) are significant at the level of 1%. All other comparisons are non-significant. Sample size = 14044.

Figure 2 presents a boxplot depicting variations in success rates in obtaining consent among interviewers. The success rate is defined as the number of obtained consents divided by the number of completed interviews for each interviewer. The number of completed interviews varied between 2 and 86 while success rates varied between 0 and 100 percent. The figure shows that there are substantial variations in success rates among interviewers with a relatively large number of outliers. Success rates in obtaining consent are the most dispersed for the MR's economic consent followed by the MR's health consent. This perhaps reflects the fact that economic linkage is the most controversial among the four consents. Moreover, the bottom quartile of interviewers has the largest dispersions irrespective of the domain. It is also worth noting that all outliers belong to the lowest quartile. The existence of important variations between interviewers in terms of success in obtaining consent warrants the modelling of interviewer effects.

Figure 2: Interviewers' success rates in obtaining consent.



Sample size = 14,044 respondents interviewed by 443 interviewers. The dots represent the outliers, the whiskers delimit the bottom and top quartiles, the box itself contains the middle two quartiles, and the middle vertical line is the median.

Regression results

Table 4: Results of a multivariate probit model jointly modelling the four consent domains.

	CM's health records		CM's education records		MR's health records		MR's economic records	
CM's gender, reference: male								
Girl	-0.044	(0.035)	-0.054	(0.037)	0.012	(0.031)	-0.0074	(0.031)
Highest socio-economic status, reference: managerial and professional								
Intermediate	0.084	(0.066)	0.11	(0.072)	-0.078	(0.050)	-0.046	(0.045)
Small employers and self-employed	0.035	(0.074)	0.11	(0.072)	-0.053	(0.056)	-0.14***	(0.055)
Lower supervisory and technical	-0.017	(0.075)	0.012	(0.085)	-0.026	(0.072)	0.068	(0.063)
Semi-routine and routine	0.045	(0.062)	0.047	(0.069)	-0.013	(0.053)	-0.0082	(0.047)
Main respondent's age	0.0058	(0.003)	0.0059	(0.004)	-0.0062**	(0.003)	-0.0070***	(0.002)
Main respondent's marital status, reference: Single								
In a couple	-0.018	(0.059)	0.033	(0.054)	0.016	(0.045)	-0.044	(0.044)
Combined labour market status, reference: both in work								
At least one in work	-0.068	(0.051)	-0.0065	(0.055)	-0.027	(0.042)	-0.0035	(0.036)
Both not in work	0.030	(0.099)	0.041	(0.105)	0.12	(0.083)	0.11	(0.081)
Housing tenure, reference: Own								
Rent	0.0068	(0.051)	-0.047	(0.054)	0.028	(0.044)	0.034	(0.044)
Other	0.058	(0.117)	0.033	(0.126)	-0.17	(0.109)	-0.16	(0.103)
Main respondent's ethnic group, reference: white								
Non-White	-0.25***	(0.076)	-0.29***	(0.081)	-0.31***	(0.069)	-0.31***	(0.064)
Main respondent's religion, reference: Christian								
Non-Christian	0.0053	(0.091)	0.039	(0.091)	-0.026	(0.076)	-0.048	(0.079)
None	0.061	(0.040)	0.046	(0.042)	0.065*	(0.038)	0.090***	(0.033)
Number of siblings in household	0.10***	(0.019)	0.11***	(0.023)	0.040**	(0.016)	0.038**	(0.015)
Log OECD adjusted income	-0.022	(0.048)	-0.016	(0.044)	0.051	(0.034)	0.051	(0.033)
Receipt of benefits, reference: No								
Yes	0.13***	(0.042)	0.12***	(0.043)	0.17***	(0.037)	0.16***	(0.035)
Were the interviews translated? reference: No								
Yes, main respondent's	0.064	(0.147)	0.019	(0.156)	0.23	(0.143)	0.19	(0.127)

Yes, partner's	-0.0074	(0.248)	-0.082	(0.220)	0.24	(0.185)	0.30*	(0.159)
Yes, both	0.69**	(0.269)	0.65**	(0.279)	0.92***	(0.262)	1.01***	(0.236)
CM's health status, reference: excellent								
Very good, good	-0.050	(0.040)	-0.062	(0.043)	-0.031	(0.031)	-0.044	(0.028)
Fair, poor	0.22*	(0.127)	0.19	(0.131)	0.055	(0.099)	0.15*	(0.089)
Main respondent's health status, reference: excellent								
Very good, good	0.013	(0.048)	0.065	(0.050)	-0.045	(0.036)	0.018	(0.035)
Fair, poor	0.10	(0.073)	0.18**	(0.079)	-0.0097	(0.059)	0.094*	(0.053)
Past response history, reference: participated in all waves								
Absent in at least one wave	-0.18***	(0.045)	-0.14***	(0.051)	-0.19***	(0.039)	-0.18***	(0.039)
Main respondent: I am a very private person, reference: strongly agree								
Agree	0.014	(0.082)	0.039	(0.086)	0.083	(0.061)	0.100*	(0.056)
Neither	0.12	(0.080)	0.10	(0.081)	0.16**	(0.060)	0.11*	(0.057)
Disagree	0.098	(0.082)	0.073	(0.086)	0.22***	(0.056)	0.21***	(0.057)
Strongly disagree	0.35***	(0.123)	0.33***	(0.123)	0.29***	(0.093)	0.25***	(0.084)
Can't say	-0.13	(0.185)	-0.17	(0.187)	-0.022	(0.150)	-0.23*	(0.135)
Other	-0.55***	(0.117)	-0.61***	(0.120)	-0.32***	(0.094)	-0.25***	(0.090)
CM's cognitive score	0.0077**	(0.004)	0.012***	(0.004)	0.00090	(0.003)	0.00098	(0.003)
Constant	1.07***	(0.334)	0.96***	(0.331)	0.83***	(0.261)	0.59**	(0.237)
Rho 21 CM's education records & CM's health records							0.99***	(0.002)
Rho 31 Main respondent's health records & CM's health records							0.87***	(0.012)
Rho 32 Main respondent's health records & CM's education records							0.84***	(0.013)
Rho 41 Main respondent's economic records & CM's health records							0.78***	(0.016)
Rho 42 Main respondent's economic records & CM's education records							0.78***	(0.016)
Rho 43 Main respondent's economic records & Main respondent's health records							0.96***	(0.005)
N				14044				

Standard errors in parentheses, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In what follows, the regression results are interpreted along the lines of the hypothesised mechanisms.

Parental pride: Table 4 shows that the CM's cognitive skills do have a positive and significant (at $p<0.01$) effect on the likelihood to consent to educational data linkage. This result provides a degree of confirmation for the parental pride hypothesis. Parents are more likely to consent to educational data linkage if their children have higher cognitive skills. In contrast, the CMs cognitive skills have a weak impact on the CM's health consent and no impact on the MRs economic and health consents.

Moreover, since the linked educational data contains the CMs' key-stage 1 performance scores which are highly correlated with cognitive skills, it is possible to conclude that under achievers are likely to be under-represented in the linked dataset. However, the total amount of bias will also depend on the additional bias arising from non-linkage (i.e. failure to link certain records even if consent was given).

Privacy and data confidentiality: the results show that in general those who disagree with the statement that they are 'very private' are more likely to consent than those who strongly agree. However, the statistical significance of the effect varies depending on whether the MRs are consenting for themselves or on behalf of the CM. When consenting to link the CM's records, only those who strongly disagree with the statement are more likely to consent. All other categories have non-significant effects. In contrast, when consenting to like their own records, almost all categories have a significant effect which is monotonically increasing with the decline in 'being private'.

By using an ordered categorical variable to measure privacy as a predisposition or personality trait, it is possible to see that the impact on consent is gradual and varies according to the person for whom consent is sought. The findings show that parents are not necessarily more protective of their children since the impact of 'being private' is almost non-significant on the CM's outcomes. While on the other hand, privacy concerns do influence the MRs' decision to link their own records. The reason behind these findings is that the CM is probably seen as the focus of the survey, while MRs see themselves as non-central to the study.

Note that since privacy is unlikely to be related to the values of the variables contained in the administrative records (whether economic, educational or health related), the strong impact of privacy on consent is unlikely to cause sample bias. However, if privacy is a variable of interest in a substantive analysis combining survey and administrative data, then the most private respondents are likely to be underrepresented.

Loyalty to the survey: Those who have dropped out from the survey in the past, at least once, are less likely to consent on all outcomes. These effects are all significant at $p<0.01$ irrespective of the domain of consent or the person for whom consent is sought. This finding confirms the loyalty assumption. Respondents who are less committed to the survey are less likely to cooperate with the in-survey requests such as consent. This finding is interesting for

two reasons. First, it shows that there is a latent propensity to cooperate which underpins participation in the survey and cooperation in sub-studies. Secondly, it shows that non-consenters are likely to be non-respondents on previous waves. Hence, survey agencies might want to allocate more resources to cases where non-response has happened in the past in a bid to reduce non-response and non-consent in the future.

Pre-exiting relations with the agency holding the administrative data: The receipt of benefits has a strong positive and significant ($p<0.01$) effect on the likelihood to consent irrespective of the outcome. In contrast, self-reported health for both the CM and the MR have mostly non-significant effects on consent. Hence, there is partial evidence to support the pre-existing relationship hypothesis.

Socio demographic background: Three socio-demographic variables have a significant impact on the likelihood to consent. First, non-white ethnic minority respondents are less likely to consent than their white counterparts. The impact of belonging to the ethnic minority group is negative, significant and strong in magnitude irrespective of the outcome. Secondly, religion has a weak effect on the MRs' likelihood to consent to link their own records, with non-religious respondents being slightly more likely to consent. Thirdly, age has a weak negative effect on consent to link the MRs' own records but not those of the CM. All other variables, have statistically non-significant effects.

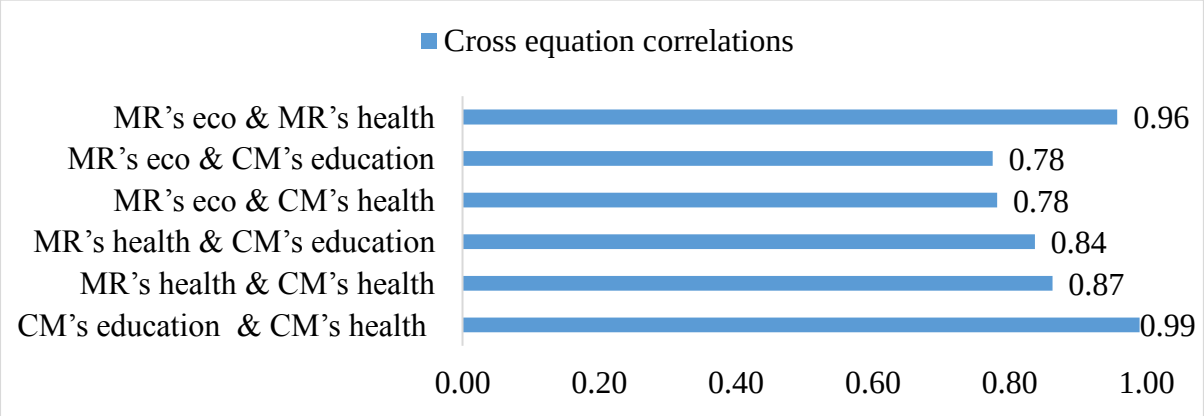
Since ethnicity is highly correlated with economic prospects, health and educational outcomes; the loss of ethnic minorities will lead to the loss of particular administrative records (lower income, lower educational achievements, and more health problems) since minorities on average have lower outcomes than the majority group. However, the total level of bias in any linked survey and administrative data depends on non-consent bias and possible non-linkage. Non-linkage occurs when it is not possible to link a case even though the MR has given consent. Non-linkage happens mostly because of incorrect identifiers and it could be non-random.

In figure 3, the estimated cross-equation correlations are presented. These correlations measure the strength of the association between the unobserved factors explaining each consent. A strong association reflects the existence of a latent propensity to consent and that the unobserved circumstances surrounding the interview are the same for all outcomes (since consents were sought in the same interview and there were no outcome-specific circumstances).

The figure shows that all correlations are strong in magnitude with the strongest being the ones relating consents sought for the same person (CM vs. MR). There are two explanations for this result. First, MRs behave differently when consenting for themselves and when consenting on behalf of the CMs. Hence there is a stronger latent propensity to consent linking the outcomes sought for the same person. Secondly, the unobserved circumstances surrounding the interview might have differential effect depending on the person for whom consent is sought. Thirdly, the strength of the correlations also reflects the fact that consent

questions for the same persons (MRs and CMs) were administrated on the same paper forms (see The Millennium Cohort Study, Ethical Review and Consent 2012). Hence, those who consented for one domain on the same form are likely to have consent on the other.

Figure 3: Cross-equation correlations for each combination of two outcomes.



All cross-equation correlations are significant at the level of 1%.

In table 5, the results from a number of models including interviewer effects are presented. As suggested by Gianelli and Micklewright, (1993) fixed effects should not be included in non-linear models. Therefore all four models are estimated using single equation linear probability procedures. The **base model** is a linear probability model with the aforementioned correlates and without interviewer's fixed effects. The **area effects model** is identical to the base model and includes additional variables measuring the assignment area characteristics (i.e. proportion of minorities, proportion unemployed, log average income, and social class composition). The **fixed effects model** is equivalent to the base model and includes interviewer fixed effects.

Table 5: Interviewer effects for each consent outcome.

Consent outcomes	Base model	Area effects	FE
	R squared		
CM health	0.04	0.05	0.17
CM education	0.05	0.05	0.19
MR health	0.05	0.05	0.17
MR economic	0.05	0.05	0.21

Sample size = 14044 respondents interviewed by 443 interviewers.

The findings show that the explanatory power of the base model is weak. All covariates explain between 4 and 5 percent of the variations in consent. When the area characteristics are included, the R-Squared are broadly unchanged indicating that the characteristics of the assignment area do not account for much of the variation in consent. The reason behind this finding is that assignment areas are very heterogeneous (large within variations) and very similar to one another (small between variations). When interviewer fixed effects are included, the R-squared are 3 to 4 times larger. This indicates that interviewer characteristics and behaviour account for a large proportion of variations in consent. This is in line with

previous evidence (Sakshaug et al. 2012, and Sala et al. 2012) where interviewers' experience, age, education and critical views of data linkage were found to affect consent.

Since interviewers' characteristics are unlikely to be correlated with those of respondents and with the health, education and economic outcomes contained in the administrative records, the impact of interviewers on consent is unlikely to be a source of sample bias in the linked datasets. However, if consent rates are low, then survey agencies should give more attention to the interviewers' characteristics because their impact is much larger than those of the respondents and because survey agencies can influence interviewer behaviour through training and interviewer allocation.

VI- Conclusion.

This study expanded our knowledge of consent by analysing adult respondents' behaviour when consenting to link their own administrative records and when consenting to link those of their seven year-old children. The study explored a number of theories and hypothesised mechanisms of consent which have not been examined in the past. In particular, it focused on: parental pride, privacy concerns, loyalty to the survey, existing relations with the agency holding the data, and the impact of the interviewers. The analysis used data from the Millennium Cohort Study, a multi-topic longitudinal social survey.

In summary, the findings show that main respondents behave differently when consenting to link their own records and when consenting on behalf of the cohort members. For instance, parents of children with high cognitive skills are more likely to consent on linking their children's educational records. In contrast, the child's cognitive skills do not affect the parents' likelihood to link their own health and economic records. Moreover, being a private person has a more significant effect on the MRs outcomes than those of the CM. When it comes to loyalty to the survey, respondents who have missed a wave in the past are found to be less likely to consent irrespective of the outcome. In contrast, partial evidence was found in support of the impact of past relationship with the agency holding the administrative data. Among the socio-demographic characteristics of respondents, ethnicity was found to have the strongest impact irrespective of the outcome. Non-white respondents are less likely to consent.

The cross-equation correlations showed that the highest level of association is between outcomes sought for the same respondent (i.e. MRs consenting for linking their own records vs. MRs consenting for linking the CMs records). When interviewers' effects were included through the use of fixed effects models, the explanatory power of the models increased by 3 to 4 times. This indicates that the interviewers' characteristics and behaviour have a large effect on consent.

In terms of fieldwork practices, the findings suggest that it is possible to identify the respondents who are less likely to consent (ethnic minorities, respondents with higher privacy concerns, and respondents who have dropped out from the survey in the past). In addition to

this, the findings show that interviewers have a strong impact on consent. Therefore, in the case of low consent rates, the matching of interviewers and respondents and the allocation of interviewers, possibility with more survey experience, to difficult cases might improve consent rates. However, more research is needed in order to have a clearer view of how interviewers affect consent.

Last but not least, the findings indicate that the linked administrative data is likely to suffer from sample composition bias due to non-consent. This is of a particular interest for the MCS data users. The sample is likely to lose children with lower cognitive skills. The effect will be larger on educational records, since these records contain the performance scores of the cohort members. Similarly the high and significant impact of ethnicity means that samples are likely to lose non-white minorities. Since ethnicity is highly correlated with educational, health and economic outcomes, the data contained in the linked administrative records will be affected by non-consent. However, the total level of bias contained in the linked survey and administrative data depends on non-consent and on the extent of non-linkage (the failure to link data even if consent was given) which might alleviate or exacerbate the initial non-consent bias.

References

Armstrong, V. Julie, B. Helen, C. Michelle, M. Moran-Ellis, J. and Shepherd, R. (2008). Public Perspectives on the Governance of Biomedical Research: A Qualitative Study in a Deliberative Context. London, UK: Wellcome Trust.

Baker, R. Shiels, C. Stevenson, K. Fraser, R. and Stone, M. (2000). What Proportion of Patients Refuse Consent to Data Collection from Their Records for Research Purposes?'' British Journal of General Practice 50, 655-56.

Birditt, K. Fingerman, K. and Zarit, S. (2010). Adult Children's Problems and Successes: Implications for Intergenerational Ambivalence. *Journal of Gerontology: Psychological Sciences. 65B(2),* 145–153.

Cappellari, L. and Jenkins, S. (2003). Multivariate probit regression using simulated maximum likelihood. *The Stata Journal.* 3(3), 278–294.

Cappellari, L. and Jenkins, S. (2006). Calculation of multivariate normal probabilities by simulation, with applications to maximum simulated likelihood estimation. *The Stata Journal*. 6(2), 156–189.

Dunn, K. Jordan, K. Lacey, R. Shapley, M. and Jinks, C. (2004) Patterns of consent in epidemiologic research: evidence from over 25,000 responders. *American Journal of Epidemiology,* 159, 1087–1094.

Fingerman, K. Cheng, Y. Birditt, K. Zarit, S. (2012). Only as Happy as the Least Happy Child: Multiple Grown Children's Problems and Successes and Middle-aged Parents' Well-being. *The Journals of Gerontology, Series B: Psychological Sciences and Social Sciences,* 67(2), 184–193.

Gianelli, C. and Micklewright, J. (1993) Estimating Fixed Effect Binary Choice Models with Long Panels: A Practical Approach to the Conditional Logit Model. *Statistica anno LIII,* 3, 453-466.

Gustman, A. L. and Steinmeier, T. L. (1999) What people don't know about their pensions and social security: an analysis using linked data from the Health and Retirement Study. *Working Paper w7368*. National Bureau of Economic Research, Cambridge.

Haider, S. and Solon,G. (1999) Non-random selection in the HRS Social Security earnings questions. Unpublished. University of Michigan, Ann Arbor.

Huang, N. Shih, S. Chang, H. and Chou, Y. (2007). Record Linkage Research and Informed Consent: Who Consents? *BMC Health Services Research* 7, 18.

Jenkins, S. Cappellari, L. Lynn, P. Jäckle, A. and Sala, E. (2006). Patterns of consent: evidence from a general household survey. *Journal of the Royal Statistical Society Series A*, 169(4), 701-722.

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 338:b866.

Korbmacher, J. and Schroeder, M. (2013). Consent When Linking Survey Data with Administrative Records: The Role of the Interviewer. *Survey Research Methods*, 7, 115–131.

Kreuter, F. and Sakshaug, J. (2014). The effect of benefit wording on consent to link survey and administrative records in a web survey. *Public Opinion Quarterly*, 7(2) 133 – 144.

Knies, G. Burton, J. sala, E. (2012). Consenting to health record linkage: evidence from a multi-purpose longitudinal survey of a general population. *Health Services Research*, 12, 52.

McKay, S. (2012) Evaluating approaches to Family Resources Survey data linking. DWP Working Paper 110.

Millennium Cohort Study, Ethical Review and Consent (2012).
http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=1601&itemtype=document

MCS Technical Report on Sampling (4th edition, CLS, 2007).
http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=409&itemtype=document

MCS Technical Report on Response (3rd edition, CLS, 2010).
http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=607&itemtype=document

MCS Guide to the Linked Education Administrative Datasets (1st edition, CLS, 2011).
www.cls.ioe.ac.uk/shared/get-file.ashx?id=1342&itemtype=document

Nelson, K. Garcia, R. Brown, J. Mangione, C. Louis, T. Keeler, E. and Cretin, S. (2002) Do patient consent procedures affect participation rates in health services research? *Medical Care*, 40, 283–288.

Olson, J. A. (1999) Linkages with data from Social Security administrative records in the Health and Retirement Study. *Social Security Bulletin*, 62, 73–85.

Petty, D, Zermansky AG, Raynor DK, et al. (2001) No thank you: why elderly patients declined to participate in a research study. *Pharmacy World and Science*, 23, 22–7.

Sakshaug, J. (2013). 'Using paradata to study response to within-survey requests'. In Kreuter, F. *Improving surveys with paradata. Analytic uses of process information*. Wiley series in survey methodology. Hoboken: Wiley, 171-190.

Sakshaug, J. Couper, M. Ofstedal, M. and Weir, D. (2012). Linking survey and administrative records mechanisms of consent. *Sociological Methods and Research*, 41(4) 535-569.

Sakshaug, J. and Kreuter, F. (2012). Assessing the magnitude of non-consent biases in linked survey and administrative data. *Survey Research Methods*, 6(2) 113-122.

Sakshaug, J. Tutz, V. and Kreuter, F. (2013). Placement, wording, and interviewers: Identifying correlates of consent to link survey and administrative data. *Survey Research Methods*, 7(2) 133-144.

Sala, E. Burton, J. and Knies, G. (2012). Correlates of obtaining informed consent to data linkage: respondent, interview and interviewer characteristics. *Sociological Methods and Research*, 41(3) 414–439.

Sala, E. Burton, J. and Knies, G. (2014). Propensity to consent to data linkage: experimental evidence on the role of three survey design features in a UK longitudinal panel. *International Journal of Social Research Methodology*, 17(5) 455-473.

Silva, M. S. Smith, W. T. and Bammer, G. (2002) The effect of timing when seeking permission to access personal health services utilization records. *Annals of Epidemiology.* 12, 326–330.

Singer, E. Van Hoewyk, J. and Neugebauer, R. (2003). Attitudes and Behaviour. The Impact of Privacy and Confidentiality Concerns in the 2000 Census. *Public Opinion Quarterly* 67, 368-84.

User Guide to Analysing MCS Data Using STATA (1st edition, CLS, 2011). http://www.cls.ioe.ac.uk/shared/get-file.ashx?id=1372&itemtype=document

Woolf, S. H., Rothemich, S. F., Johnson, R. E. and Marsland, D.W. (2000) Selection bias from requiring patients to give consent to examine data for health services research. Archives of family medicine, 9, 1111–1118.