
Article:
Connelly, Roxanne orcid.org/0000-0002-3886-1506 and Gayle, Vernon (2017) An Investigation of the Consistency of Parental Occupational Information in UK Birth Records and a National Social Survey. European Sociological Review. pp. 240-256. ISSN 0266-7215

https://doi.org/10.1093/esr/jcw060

European Sociological Review, 2017, Vol. 33, No. 2, 240–256
doi: 10.1093/esr/jcw060
Advance Access Publication Date: 25 January 2017
Original Article

An Investigation of the Consistency of Parental Occupational Information in UK Birth Records and a National Social Survey

Roxanne Connelly[1,*] and Vernon Gayle[2]

[1]Department of Sociology, Social Sciences Building, The University of Warwick, Coventry CV4 7AL, UK and [2]School of Social and Political Science, University of Edinburgh, 18 Buccleuch Place, Edinburgh EH8 9LN, UK

*Correspondence author. Email: R.Connelly@warwick.ac.uk

Submitted July 2016; revised November 2016; accepted December 2016

Abstract

In the United Kingdom, new sources of administrative social science data are unfolding rapidly but the quality of these new forms of data for sociological research is yet to be established. We investigate the quality and consistency of the parental occupational information that is officially recorded on administrative birth records by undertaking a comparison with information collected from the same parents in the UK Millennium Cohort Study (MCS). We detect a large amount of missing information in the birth records and a range of inconsistencies. We present an empirical analysis of MCS data using parental social class measures derived both from the birth records and the survey to assess the effects of these discrepancies. We conclude that parental occupational information from administrative birth records should not be assumed, *a priori*, to be suitable for sociological analyses and that further research should be undertaken into their consistency and accuracy.

Introduction

The explosion in the availability of new sources of data in the early part of the 21st century is set to revolutionize research possibilities within sociology. The emergence of 'big data' and other forms of 'digital data' offer new opportunities to study individuals and societies (see for example Manovich, 2011; Burrows and Savage, 2014; Kitchin, 2014; Schroeder, 2014). Simultaneously, advances in e-research and computer science provide increasingly improved solutions for linking large data sets (see Goerge and Lee, 2001; Halfpenny and Procter, 2015).

Administrative social science data resources contain information which originate from the operation of administrative systems, typically those that are associated with public sector agencies (Elias, 2014; Woollard, 2014). These data sets offer new opportunities for empirical sociological research. Researchers in the Nordic nations have benefited from unparalleled access to administrative social science data (see United Nations, 2007), whilst at the same time their national registers have provided the basis for a strong data infrastructure. By contrast, in most other nations, sociological analyses of administrative data have been far less widespread and are far from routine. The increased research potential that would be offered by improved access to administrative data has recently been recognized in the United Kingdom, and major infrastructural investment has been

made to support the analysis of administrative data[1] (see Administrative Data Taskforce, 2012).

The new sources of administrative social science data in countries like the UK are unfolding rapidly and haphazardly, and are not supported by the framework of a national population register. The quality of these new forms of data for sociological research is yet to be established. This article is original because it engages in an innovative analysis to assess the consistency of a set of administrative data and survey data collected from the same individuals. The specific focus of this article is the assessment of the consistency of parental occupational information in UK birth records.

Within sociology there is a longstanding recognition that in industrialized societies occupations are often the most powerful single indicator of levels of material reward, social standing, and life chances (Parkin, 1971; Rose *et al.*, 2005). Occupations remain a key element of contemporary social life, and occupation-based indicators are a cornerstone of sociological research. Measures of parental socio-economic position are essential to analyses of inequalities in a wide range of areas for example social stratification, education, health, and well-being (see for example Graham, 2007, Bukodi and Goldthorpe, 2012; Sullivan *et al.*, 2013; Grätz, 2015).

In the UK, the only source of administrative data on parental occupations, taken at the same age for all children, are birth records.[2] Parental occupational information is also available in parental marriage records. These records are of limited use since an increasing number of children are born outside of marriage, and the gap between marriage and children's birth dates varies substantially. UK census records provide another potential source of information on parental occupations. The utility of this data source is also questionable since the UK census is conducted decennially.

In the UK the systems for the collection of birth registrations vary slightly between territories,[3] but each territory collects information on the name, date, and place of birth of the child, the father's name and occupation, and the mother's name and occupation. We are in the methodologically fortunate and unusual position to have access to linked data on parental occupations reported on administrative birth records and also a short time later in a social survey interview conducted as part of the UK Millennium Cohort Study (MCS). Our analyses investigate the consistency of reports of parental occupations between these two data resources.

The MCS data are collected specifically for the purposes of research, and the data collection has been designed to maximize the validity and reliability of the data. The MCS survey is administered by a professional data collection agency, and interviews are carried out by trained interviewers collecting data specifically for the purposes of research. In large-scale nationally representative social surveys, extensive cross-checking and validation work is carried out to maximize data quality and the data quality will be clearly documented.

By contrast, the administrative birth records are not collected for the purposes of research. Goerge and Lee (2001), for example, note that the original motivation for collecting administrative data should be questioned when assessing its quality. Researchers should consider whether the information they are interested in is central to the purposes it was collected for. If certain measures are not required for the operation of an administrative system, they may not be collected conscientiously (Goerge *et al.*, 1992; Goerge and Lee, 2001). The collection of parental occupations on the birth records is not directly required for the operation of any administrative system or the delivery of a service. Therefore, the accurate collection of these data is not of immediate importance to the frontline worker collecting the information. The influence which frontline workers can have on administrative systems is highlighted clearly in Lipsky (1979).

It is also important to note that administrative data resources can take many forms. In this case, the data collected in birth records are based on the information provided by an individual, in much the same way as they would provide information in a social survey interview. In that respect, our comparisons between these data sets investigate differences in the recording of occupations by a registrar compared with a social survey interviewer, and not the differences between survey and administrative data in general. This form of administrative data collection is not unusual in the UK. The UK does not have a national register, and individuals do not have a unique identification number; therefore, information is most commonly provided to different administrative systems by the individuals themselves. In some cases, administrative data will be produced through more objective processes such as records on the amount of tax paid, the educational qualification attained in national examinations, or the model and colour of a vehicle registered to a motorist. The characteristics and accuracy of administrative data will vary according to its source, and the manner in which it is collected. Goerge and Lee (2001) emphasize that the degree of error varies between administrative data systems, and they encourage researchers to assess each new administrative social science data set individually for every new research question. Following the prescription from Goerge and Lee (2001), the central aim of this article is to undertake an evaluation of UK administrative birth records.

Whilst both the survey data and administrative data will contain inaccuracies, we strongly believe that the purposes and processes involved in the production of the survey data will usually render these data more suitable for social research than administrative birth records. Therefore, we consider the comparison of parental occupations on birth records with available social survey data to be a valid and meaningful assessment of the quality of this administrative data resource.

This article will address three main questions:

1. How consistent are maternal and paternal occupations reported on the survey and in the birth records?
2. Are parental characteristics associated with patterns of agreement and missingness of occupational information on the survey and the birth records?
3. What potential impact do disagreements have on empirical sociological analyses?

There are some previous studies from the United States that have investigated the accuracy of the occupational information provided on birth records. These have generally been from within the field of epidemiology, and they have been motivated by the need to identify occupational risk factors for maternal and child health. Carucci and Prasad (1979) studied birth records in upstate New York. This study found a lack of detail in reports of mothers' occupations. This precluded the use of full occupational codes,[4] and this in turn would be a major impediment for the development of occupation-based socio-economic measures. Carucci and Prasad (1979) encountered a high degree of missing maternal occupational information. Mothers were required to give details of their last employment, and 65 per cent were described as 'housewives' on the birth record. The survey identified that over half of mothers described as 'housewives' on the birth record did have a previous occupation. Shaw *et al.* (1990) found more promising results when assessing parental occupations on Californian birth records. For 71 per cent of mothers and 80 per cent of fathers, the occupation on the birth record was the same as the occupation reported in an interview. Brender *et al.* (2008) studied parental occupations on birth records in Texas and found that mothers were frequently misclassified as 'homemakers' or unemployed when they did have previous employment.[5] Paternal occupations were missing in 22 per cent of cases. For those parents with occupational information available, 77 per cent of maternal occupations and 63 per cent of paternal occupations matched between the birth records and the interview.

We can only speculate on the reasons for the finding of increased missingness of mother's occupations on administrative birth records. We conjecture that the following three factors may be implicated. First, mothers may consider their occupation as being a 'housewife' even though this is not officially recognized as an occupation. Second, registrars may not fully explain that they are asking for last occupation, and not what the mother considers as her current activity. Third, the registration takes place shortly after the baby's birth. If the father attends the registration on his own it is plausible that he may provide less detailed information on his partner's occupation. An observational study of registrations, which included suitable follow-up interviews, would be required to comprehensively establish the reasons for the under-reporting of maternal occupations.

Data and Methods

The data that are investigated in this analysis are drawn from the UK Millennium Cohort Study (MCS) (for more details, see Connelly and Platt, 2014). The MCS is a sample of children born between the 1st of September 2000 and the 11th of January 2002 throughout England, Wales, Scotland, and Northern Ireland. The MCS currently comprises five survey waves. We use information from the first wave of data collection, when the children were around 9 months of age (SN4683, UCL Institute of Education, 2012). Data from birth records were linked to the MCS survey by statistical agencies in each of the constituent UK territories (i.e. the Office for National Statistics in England and Wales and the General Register Office in Scotland). These data are held in the 'Millennium Cohort Study Birth Registration and Maternity Hospital Episode Dataset' (SN5614, UCL Institute of Education, 2008). Full details of the data linkage process are available in Hockley *et al.* (2007).

There are 16,629[6] families included in the first MCS survey (excluding families in Northern Ireland), and 15,013 of these families were successfully linked to the birth records data, a 90 per cent linkage rate (see Tables 1 and 2). Our analyses exclude birth registrations from Northern Ireland, as the occupational information in these cases was provided in the form of Standard Occupational Classification 90 (SOC90) codes which are different to the occupational information from other territories which is provided in the form of SOC2000 codes. There is no direct conversion between the older SOC90 and the more recent SOC2000. Therefore, to avoid the possibility of introducing additional inconsistencies into the analyses, we have excluded Northern Ireland.

The MCS data are collected through a face-to-face interview, conducted in the family's home. Information is collected from main respondents (usually the child's mother)

Table 1. Descriptive statistics of consent to link and achieved linkage of MCS families in the survey and birth records, by UK territory

	MCS families	MCS families that were successfully linked (overall)	
	n	*n*	(%)
England	11,532	10,326	(90%)
Wales	2,761	2,545	(92%)
Scotland	2,336	2,142	(92%)
Total	16,629	15,013	(90%)

Note: The analyses in this article are undertaken at the family level, as our focus is parental information.

Table 2. Descriptive statistics of consent to link and achieved linkage of MCS cohort members, by UK territory

	MCS children	Children with consent to link		Children successfully linked (if consent was given)		Children that were successfully linked (overall)
	n	*n*	(%)	*n*	(%)	%
England	11,694	10,542	(90%)	10,474	(99%)	90%
Wales	2,799	2,594	(93%)	2,578	(99%)	92%
Scotland	2,370	2,179	(92%)	2,173	(100%)	92%
Total	16,863	15,315	(92%)	15,225	(99%)	90%

Note: Adapted from Hockley et al. (2007). Consent was given per child and not per family.

and partner respondents (usually the child's father). We identify and include only natural mothers and natural fathers in our sample, as these are the parents whose details were recorded on the birth record. Registration of a birth is made in person at a Registry Office, and an official registrar records the information. In Scotland, births must be registered within 21 days of the birth, and in England and Wales, births must be registered within 42 days. Births can be registered by either parent if they are married, or by the mother if the parents are unmarried.

The MCS has a complex sample design which should be appropriately represented in statistical analyses (see Plewis *et al.*, 2004). When making descriptive comparisons between the two data sources, we present unadjusted results, as we are interested specifically in comparing the information available for the same families in these two different data sources. When undertaking multivariate analyses, however, we represent the complex survey design. The full unadjusted results of all models are provided in the supplementary materials, and the substantive conclusions generally remain consistent in the adjusted and unadjusted models. In this analysis, we have used the standard weights that are deposited with the data (see Ketende and Jones, 2011) because they provide general and robust adjustments. In other analyses, it might be desirable to construct bespoke weights with the aim of making specialized adjustments.

Occupational Information

Maternal and paternal occupational information is collected in the MCS survey using the following questions. If the respondent is either currently working, has a paid job but is on leave, or has worked in the past but is not currently working, they are asked, 'What is your main job?' The respondents are then asked, 'What do you mainly do in your job?' These questions are asked to all respondents if they have previously stated that they have worked in the past, even if they are not currently working. All parents who have held a job at some point in their lives should report occupational information. The interviewer collects the occupational details and these are recorded as free text within a computer system.

We have gained an understanding of the practical process of how occupational information is collected by registrars in the birth records through email correspondence with the relevant national statistical agencies and through meeting and discussing the data collection process with a registrar. In comparison to the standardized questions used in the survey, registrars do not use a standard set of questions to collect the occupational information. Registrars ask for the mother and father's occupation, in an attempt to collect information on the present or last known occupation. If an individual is unemployed or retired, they are asked for details of their last job. If 'housewife' is given as an occupation,

registrars are told to inform the parent that this is not an occupation in the sense of a profession, employment, business, or calling, and they are encouraged to probe for a previous occupation. The registrars are allowed to enter the term 'housewife' or 'house person' as an occupation if the parent insists. The occupational details which registrars collect are entered as free text into a computer system.

For both the survey and the birth records, the occupational information is coded to a standard occupational classification by a third person (i.e. not by the survey interviewer or the registrar). The occupational information collected in the survey were coded to the SOC2000 (Office for National Statistics, 2000) after collection using the Computer Assisted Structured Coding Tool (CASCOT, Elias *et al.*, 1993; Jones, 2004). This tool suggests occupational codes based on the text of a job title, but a coder must decide if this code is suitable and select a more suitable code if one is required. There is a small element of interpolation involved in the process of coding occupations, but it is largely formulaic. The occupational information on the birth records was also coded to SOC2000 codes using computer-assisted programmes. In some cases, this means that the coder has to adjudicate and decide on the most suitable occupational code for the occupational information available. To date, we are not aware of any results of side-by-side calibration tests of CASCOT and the government occupational coding programmes. Both the survey data coders and the birth records data coders employed verification checks where a proportion of the coding was checked by an additional coder.

In this article we consider the consistency of occupations based on the four-digit SOC2000 codes available in the social survey and the birth records. For most research purposes, detailed occupational codes will be converted into an occupation-based measure (see Connelly *et al.*, 2016). Therefore, we also consider the agreement between the occupations coded to the eight class version of the UK National Statistics Socio-Economic Classification (NS-SEC, Office for National Statistics, 2010). Ideally NS-SEC is produced using standard occupational codes and information on employment status[7] (Rose *et al.*, 2005). Employment status information is collected in the birth records; however, in the data set employment status, information is only available for Scottish births.[8] The Scottish employment status information available is not presented in a standardized form which would permit its use in coding NS-SEC in the officially prescribed manner. Therefore we have coded NS-SEC using only occupational information, by allocating occupations to NS-SEC categories without reference to employment status, which is known as the simplified method (see Rose *et al.*, 2005). To ensure comparability and to maintain clarity, we also use the simplified method when coding NS-SEC from the survey data, although suitable employment status information is available in the MCS.[9]

Analysis

Question 1: How consistent are maternal and paternal occupations reported on the survey and in the birth records?

Missing Occupational Information

The percentage of valid and missing occupational information on the birth record for mothers and fathers in our analytical sample is reported in Table 3. Overall 90 per cent of mothers and 73 per cent of fathers had valid SOC2000 codes in the survey. In the birth record, 62 per cent of the mothers and 86 per cent of the fathers had valid SOC2000 codes. In five cases for mothers and 11 cases for fathers, an occupational code was given on the birth record that was not a valid SOC2000 code, we recoded these cases as missing 'other'. In line with the findings of Carucci and Prasad (1979) and Brender *et al.* (2008), there is a large amount of missing occupational information for mothers on the birth record. Fourteen per cent of fathers had missing occupational information on the birth record in our sample, whereas 38 per cent of mothers had missing occupational information. Of those mother's with missing occupational information, 15 per cent were recorded as undertaking 'full time care of home/relative' and a further 20 per cent were recorded as having 'occupation not stated'.

Table 4 shows the percentage of valid occupational information available in both the survey and birth record. In our sample, 68 per cent of families have valid occupational information for fathers on both the birth record and survey, and 61 per cent of families have valid occupational information for mothers on both the birth record and the survey. In the survey, 22 per cent of families only have valid occupational information for their mother, and only 5 per cent only have valid occupational information for their father. In the birth record the situation is reversed, 6 per cent only have valid occupational information for their mother, and 30 per cent only have valid occupational information for their father. In 29 per cent of cases, a valid occupation was reported for the mother in the survey when they had missing occupational information on the birth record; this only occurred in 4 per cent of cases for father's

Table 3. Descriptive statistics of the percentage of valid and missing occupational information in the analytical sample

	Mother		Father	
	n	(%)	*n*	(%)
Survey				
Valid SOC2000	13,505	(90%)	10,913	(73%)
Missing	1,508	(10%)	4,100	(27%)
Total	15,013	(100%)	15,013	(100%)
Birth records				
Valid SOC2000	9,265	(62%)	12,884	(86%)
Invalid SOC2000	5	(>1%)	11	(>1%)
Missing: Inadequately described	218	(1%)	403	(3%)
Missing: Occupation not stated	3,036	(20%)	1,508	(10%)
Missing: Retired	1	(>1%)	7	(>1%)
Missing: Student	247	(2%)	164	(1%)
Missing: Full-time care of home/relative	2,213	(15%)	9	(>1%)
Missing: No previous job	6	(>1%)		
Missing: Other	22	(>1%)	27	(>1%)
Total missing	5,748	(38%)	2,129	(14%)
Total	15,013	(100%)	15,013	(100%)

Note: 'Missing: Other' includes non-valid SOC2000 codes and cases where SOC2000 was missing with no explanation.

Table 4. Descriptive statistics of the percentage of valid occupational information for mothers and fathers on the survey and birth records jointly

	n	(%)
Father		
Valid SOC2000 on survey and birth record	10,241	(68%)
Missing on birth record only	672	(4%)
Missing on survey only	2,643	(18%)
Missing on birth record and survey	1,457	(10%)
Total	15,013	(100%)
Mother		
Valid SOC2000 on survey and birth record	9,168	(61%)
Missing on birth record only	4,337	(29%)
Missing on survey only	97	(>1%)
Missing on birth record and survey	1,411	(9%)
Total	15,013	(100%)
Survey		
Mother and father	10,146	(68%)
Father only	767	(5%)
Mother only	3,359	(22%)
Both missing	741	(5%)
Total	15,013	(100%)
Birth record		
Mother and father	8,426	(56%)
Father only	4,458	(30%)
Mother only	839	(6%)
Both missing	1,290	(9%)
Total	15,013	(100%)

occupational information. This suggests that there may be under-reporting of valid maternal occupations on the birth record. The high degree of missingness for maternal occupational information on the birth record is in line with the findings from the aforementioned studies from the United States (see Carucci and Prasad, 1979; Brender *et al.*, 2008).

There may be a higher degree of paternal missingness on the survey compared with the birth records, as 15 per cent of MCS children were born to parents who were not in a co-residential partnership and non-resident parents did not take part in the survey (Kiernan, 2006). For resident parents, there was also a higher degree of missingness for partner interviews (mainly undertaken by fathers) compared to main interviews (Dex and Joshi, 2004). Father's information may be more likely to be included on the birth record, as this can be used to gain parental rights, and can also be used as evidence of paternity in claims for child maintenance payments.[10] There may also be social stigma attached to not including a child's fathers' details on the birth record (see for example Maldonado, 2011). Overall, there are far stronger incentives for a father's details to be entered on a child's birth record, than for a father to take part in the MCS survey.

Agreement between the Survey and the Birth Records

We now investigate the agreement between the SOC2000 codes reported in the birth records and the

survey. Overall 36 per cent of maternal occupational information and 37 per cent of paternal occupational information is the same in the two data sources (Table 5). When we consider only those cases where valid occupational information is available in both data sources, 59 per cent of maternal occupations match and 54 per cent of paternal occupations match (Table 6).

The per cent agreement between sources is a measure of consistency. We also present estimates of Cohen's Kappa (Table 7), a measure of inter-rater reliability (Cohen, 1960). Although interpretations of the magnitude of Kappa should be treated with caution (see Bakeman *et al.*, 1997), Landis and Koch (1977) suggest that kappa values over 0.61 should be considered as substantial, and Fleiss *et al.* (2013) suggest that values over 0.75 should be considered as excellent. Table 7 shows the Kappa statistic for agreement between SOC2000 codes and NS-SEC. The Kappa values are calculated for those cases without missing occupational information, and show a moderate, but not overwhelming level of reliability between sources.

Error in Practice and Error in Principle

There are disagreements between the SOC2000 codes reported in the birth records and the survey, and we theorize that these disagreements take two forms. The first we term 'error in principle', and the second we term 'error in practice'. An error in principle occurs when the SOC2000 codes do not match but this does not impact the position of the individual when the occupation is coded to a socio-economic measure (e.g. NS-SEC). For example, a secondary school teacher (SOC2314) who is recorded as a primary school teacher (SOC2315) would have a different SOC2000 code but both occupational codes would be included in NS-SEC 2 (lower managerial, administrative, and professional occupations). In 'principle', this is an error but in 'practice' it would have no effect in an analysis that used NS-SEC as an explanatory variable.

By contrast, an error in practice occurs when the SOC2000 codes do not match and also lead to a discrepancy in the socio-economic position which would be allocated to an individual. For example, a dispensing optician (SOC3216) who is recorded as an ophthalmic optician (SOC2214) would be coded to NS-SEC 2 (lower managerial, administrative, and professional occupations) instead of NS-SEC 3 (intermediate occupations). In 'practice', this disagreement could have an effect on an analysis that used NS-SEC as an explanatory variable. We reiterate that our analysis allows us to consider the consistency between the two data sources and not whether either data set is error free; however, the consideration of 'error in principle' and 'error in practice' provides additional insight into the nature of the disagreement between the survey and administrative data.

The degree of 'error in practice' and 'error in principle' will depend on the occupation-based measure that is derived from the detailed occupational codes (e.g.

Table 6. Agreement between SOC2000 if there is valid occupational information on the survey and the birth record

	Mother		Father	
	n	(%)	*n*	(%)
SOC2000 matches	5,370	(59%)	5,528	(54%)
SOC2000 does not match	3,798	(41%)	4,713	(46%)
Total	9,168	(100%)	10,241	(100%)

Table 7. The percent agreement and Kappa for SOC2000 codes and NS-SEC in the MCS survey and birth records

	% agreement	Kappa statistic
Mother		
SOC	60%	0.59
NSSEC	75%	0.68
Father		
SOC	54%	0.54
NSSEC	67%	0.62

Table 5. The percent agreement between SOC2000 codes in the MCS survey and birth records

	Mother		Father	
	n	(%)	*n*	(%)
SOC2000 matches	5,370	(36%)	5,528	(37%)
SOC2000 does not match	3,798	(25%)	4,713	(31%)
SOC2000 missing on both	1,411	(9%)	1,457	(10%)
SOC2000 missing on birth record only	4,337	(29%)	672	(4%)
SOC2000 missing on survey only	97	(1%)	2,643	(18%)
Total	15,013	(100%)	15,013	(100%)

SOC2000). For example, there would be less 'error in practice' if using a three-category social class scheme (e.g. the three-class version of NS-SEC) compared with the more common eight-category NS-SEC scheme. When SOC2000 is used to construct finer-grained measures such as a social stratification scales (e.g. CAMSIS or SIOPS, see Treiman, 1977; Prandy, 1999), then it is likely that more 'errors in practice' will occur than when a categorical socio-economic measure with a limited number categories is derived.

We demonstrate these two forms of 'error' using the eight-class version of the NS-SEC. Table 8 demonstrates the degree of 'error in practice' and 'error in principle' for cases where there is a valid SOC2000 code on both the survey and birth record. An 'error in practice' occurs for 60 per cent of cases where mothers' occupations do not match, and 71 per cent of cases where fathers' occupations do not match. If we consider all cases with valid SOC2000 codes on the survey and birth record regardless of whether they match, we can determine a total rate of 'error in practice'. There is a total 'error in practice' rate of 25 per cent for mothers and 33 per cent for fathers. These 'error in practice' rates indicate that there is a notable, and consequential, degree of disagreement between these two data sources.

Tables 9 and 10 show the cross-tabulation of the two NS-SEC measures, one coded using occupations on the birth record and the other from the survey (for mothers and fathers). The shaded cells show the percentage of mothers or fathers who would be coded to the same NS-SEC category in both data sources. For mothers, only 53 per cent of those identified as belonging to NS-SEC 5 using the birth records, for example, were also coded to this category using the survey data. Twenty-two per cent of these mothers were coded to NS-SEC 6 using the survey data. For fathers, 61 per cent of those coded to NS-SEC 3 using the birth records, for example, were also coded to this category using the survey data. Using the survey data, 12 per cent of these fathers would be in NS-SEC 2.

Question 2: Are parental characteristics associated with patterns of agreement and missingness of occupational information on the survey and the birth records?

We now investigate what factors are associated with patterns of consistency and missingness in the

Table 8. Error in practice and error in practice in the coding of SOC2000 to NS-SEC

	Cases where the two SOC2000 codes do not match		All cases with a valid SOC2000 in the survey and birth record (matches and mis-matches)	
	Mother *n* (%)	Father *n* (%)	Mother *n* (%)	Father *n* (%)
Error in practice (NS-SEC mismatch)	2,291 (60%)	3,336 (71%)	2,291 (25%)	3,336 (33%)
Error in principle (NS-SEC unaffected)	1,507 (40%)	1,347 (29%)	1,507 (16%)	1,347 (13%)
Total	3,798	4,713	9,168	10,241

Table 9. Mother's NS-SEC coded from occupations in the birth record and occupations in the survey

		Survey NS-SEC								
		1.1 %	1.2 %	2 %	3 %	4 %	5 %	6 %	7 %	*n* (%)
Birth Record NS-SEC	1.1 Large employers and higher managerial	56	3	24	10	1	0	5	1	301 (100)
	1.2 Higher professionals	3	78	10	5	0	0	2	0	526 (100)
	2 Lower managerial and professional	4	2	75	10	3	0	5	1	2,299 (100)
	3 Intermediate	1	3	7	76	2	0	9	3	2,675 (100)
	4 Small employers and own account	1	1	8	5	66	1	14	5	333 (100)
	5 Lower supervisory and technical	0	1	3	11	2	53	22	7	87 (100)
	6 Semi-routine	0	0	3	8	1	1	78	9	1,986 (100)
	7 Routine	0	0	1	2	3	1	17	76	961 (100)

Note: NS-SEC is coded using the simplified method. The base *n* is the number of cases where there is a valid SOC2000 on the survey and the birth record (total $n = 9{,}168$).

Table 10. Father's NS-SEC coded from occupations in the birth record and occupations in the survey

		Survey NS-SEC								
		1.1 %	1.2 %	2 %	3 %	4 %	5 %	6 %	7 %	Total *n* (%)
Birth Record NS-SEC	1.1 Large employers and higher managerial	63	5	22	2	3	1	2	2	735 (100)
	1.2 Higher professionals	7	73	9	4	1	2	2	1	1,065 (100)
	2 Lower managerial and professional	10	4	62	6	5	3	6	4	1,835 (100)
	3 Intermediate	5	9	12	61	1	5	5	2	1,046 (100)
	4 Small employers and own account	1	1	4	1	76	4	7	7	1,216 (100)
	5 Lower supervisory and technical	2	3	5	3	3	70	7	8	1,102 (100)
	6 Semi-routine	2	2	5	4	4	7	65	13	1,585 (100)
	7 Routine	1	0	2	2	4	5	16	69	1,657 (100)

Note: NS-SEC is coded using the simplified method. The base *n* is the number of cases where there is a valid SOC2000 on the survey and the birth record (total *n* = 10,241).

occupational information between the two data sources. In the first stage of this analysis, we estimate a series of logistic regression models to investigate missingness; then we estimate multinomial logistic regression models to investigate different patterns of missingness and agreement in the two data sources. In the regression analyses, we use additional information about the parents taken from the MCS survey[11] as our explanatory variables (see Table 11).

Estimating regression models to investigate the extent to which missingness and agreement are associated with socio-demographic factors have proved to be effective (see for example Plewis, 2007). When analysing administrative social science data, there are often a limited number of explanatory variables, however, which could be used in techniques such as multiple imputation (Connelly *et al.*, 2016). Through this enquiry we seek to deepen our understanding of the nature of occupational data available on UK birth records by assessing the extent to which the patterns of missingness in these data are associated with parental characteristics, and therefore to identify any key biases in the data source. Previous studies of non-response in social surveys have documented that those who do not respond are likely to be younger, less educated, and from ethnic minorities (Dex *et al.*, 2008). In previous studies investigating the agreement of occupations on birth records and an interview in the United States, younger mothers (<25 years), mothers with lower levels of education, and those of Black ethnicity were more likely to have mismatched occupations between data sources (Brender *et al.*, 2008).

The results of the logistic regression models of missingness are summarized in Table 12. Separate models are estimated for mothers, fathers, the survey, and the birth record. In relation to age, families with older mothers are less likely to have missing information on the survey and the birth record for both mothers and fathers. It is plausible that younger mothers and fathers

Table 11. Descriptive statistics for the additional variables taken from the MCS survey

Parents' characteristics (from survey)	*n*	(%)
Mother's age at delivery		
Under 19	1,318	(8%)
20–29	6,877	(43%)
30–39	6,323	(47%)
Over 40	309	(2%)
Mother's ethnicity		
White	12,653	(91%)
Mixed	167	(1%)
Indian	351	(2%)
Pakistani	647	(2%)
Bangladeshi	253	(1%)
Black Caribbean	214	(1%)
Black African	271	(1%)
Other (Including Chinese)	271	(1%)
Parents' Highest Academic Qualification		
Higher Degree	222	(2%)
Undergraduate Degree	1,351	(11%)
Diploma	823	(6%)
A Levels	1,119	(8%)
GCSE (Grades A–C)	5,049	(36%)
GCSE (Grades D–G)	2,130	(15%)
None	4,133	(23%)
n	14,827	

Note: This sample is formed of cases which contain complete information on the four additional variables. In this sample, 185 cases are dropped, as they contain missing information on two variables, and one case is dropped, as it contains missing information on three variables. The final analytical sample for the regression analyses is 14,827. The data are adjusted to reflect the MCS survey design.

are more likely to not have a prior occupation. There are complex patterns of association between mother's ethnicity and missingness. For mothers' occupational information on the birth record, cases with mothers from Pakistani and Bangladeshi backgrounds are more likely to be missing than other groups; however, this pattern is less clear for the survey data. Again, for fathers there are no clear relationships with ethnicity, although cases with mothers from Black African and Black Caribbean backgrounds are more likely to have missing fathers' occupational information on the survey and the birth record than those from White backgrounds. In terms of education, for mothers' occupational information, there is a fairly clear pattern of parents with higher education levels being less likely to be missing on both the survey and birth record. There is a clear pattern of missingness related to education level for fathers' occupations. Overall there are no notable differences in the patterns of occupational information missingness on the survey and the birth records, and there are clear patterns of non-random missingness in both data sources.

Tables 13 and 14 show the multinomial logistic regression models for mothers and fathers, respectively. The distribution of mothers and fathers in the five outcome categories is summarized in Table 5. Only 1 per cent of mothers ($n = 97$) are in the 'missing on survey only' category; this estimate should therefore be treated with caution. Families with older mothers and those from the White ethnic group are less likely to have any combination of missing occupational information for mothers or fathers, compared to having matching occupational information. There are less clear patterns for the association with education level. There is no obvious consistent pattern of association between the variables considered here and whether occupations match or do not match on the birth record and survey, for either mothers or fathers.

Genuine Occupational Change

A weakness of the comparison between these two data sources is that the survey interview occurred around 9 months after the birth registration. Therefore, it is

Table 12. Logistic regression models. Outcome 1 = SOC2000 missing; 0 = SOC200 not missing

Parents' characteristics (from survey)	Mother survey		Mother birth record		Father survey		Father birth record	
	Log odds	SE	Log odds	SE	Log odds	SE	Log odds	SE
Mother's age at delivery								
Under 19	1.47***	(0.12)	0.90***	(0.09)	1.40***	(0.08)	1.26***	(0.08)
20–29	Ref.		Ref.		Ref.		Ref.	
30–39	−0.76***	(0.10)	−0.35***	(0.05)	−0.66***	(0.05)	−0.63***	(0.07)
Over 40	−1.61***	(0.30)	−0.43*	(0.17)	−0.66***	(0.18)	−0.98***	(0.29)
Mother's ethnicity								
White	Ref.		Ref.		Ref.		Ref.	
Mixed	1.37***	(0.31)	0.65**	(0.21)	0.95***	(0.23)	1.03***	(0.18)
Indian	1.79***	(0.23)	0.55**	(0.17)	−0.18	(0.24)	−0.91**	(0.29)
Pakistani	2.62***	(0.16)	1.89***	(0.14)	0.24	(0.14)	−0.09	(0.19)
Bangladeshi	3.26***	(0.22)	2.01***	(0.21)	−0.04	(0.21)	−0.68**	(0.22)
Black Caribbean	0.41	(0.28)	0.17	(0.15)	1.66***	(0.21)	1.12***	(0.14)
Black African	2.44***	(0.24)	0.57***	(0.14)	1.67***	(0.18)	1.47***	(0.18)
Other (including Chinese)	2.32***	(0.23)	1.15***	(0.17)	−0.07	(0.21)	0.36	(0.25)
Parents' highest academic qualification								
Higher degree	0.23	(0.46)	−0.76***	(0.21)	0.14	(0.21)	−0.05	(0.25)
Undergraduate degree	−0.44	(0.25)	−0.73***	(0.09)	−0.48***	(0.11)	−0.34*	(0.14)
Diploma	−0.32	(0.26)	−0.48***	(0.12)	0.11	(0.11)	−0.20	(0.14)
A Levels	−0.13	(0.28)	−0.21*	(0.10)	−0.02	(0.09)	−0.00	(0.12)
GCSE (Grades A–C)	Ref.		Ref.		Ref.		Ref.	
GCSE (Grades D–G)	0.31*	(0.13)	0.34***	(0.07)	−0.21**	(0.06)	−0.04	(0.09)
None	1.85***	(0.11)	1.04***	(0.06)	0.15*	(0.06)	0.48***	(0.07)
Constant	−3.92***	(0.13)	−0.91***	(0.06)	−1.13***	(0.05)	−2.03***	(0.06)
n	14,827		14,827		14,827		14,827	

Note: The data are adjusted to reflect the MCS survey design.

Table 13. Multinomial logistic regression of comparisons of SOC2000 between the survey and birth record for mothers. The base category is: SOC2000 matches

Parents' characteristics (from survey)	Does not match		Missing on both		Missing on birth record only		Missing on survey only	
	Versus matches		Versus matches		Versus matches		Versus matches	
	Log odds	SE	Log odds	SE	Log odds	SE	Log odds	SE
Mother's age at delivery								
Under 19	0.19	(0.12)	1.99***	(0.14)	0.79***	(0.10)	0.93**	(0.31)
20–29	Ref.		Ref.		Ref.		Ref.	
30–39	−0.10	(0.06)	−0.92***	(0.11)	−0.35***	(0.06)	−0.93**	(0.31)
Over 40	−0.23	(0.18)	−1.87***	(0.34)	−0.42*	(0.18)	−1.48	(0.97)
Mother's ethnicity								
White	Ref.		Ref.		Ref.		Ref.	
Mixed	−0.04	(0.26)	1.57***	(0.37)	0.54*	(0.25)	2.01***	(0.53)
Indian	0.07	(0.19)	2.02***	(0.27)	0.33	(0.20)	1.52**	(0.56)
Pakistani	−0.35	(0.25)	3.39***	(0.23)	1.31***	(0.18)	1.26*	(0.51)
Bangladeshi	−0.74	(0.44)	3.80***	(0.30)	1.06**	(0.38)	2.19***	(0.56)
Black Caribbean	−0.15	(0.23)	0.34	(0.29)	0.11	(0.17)	0.98	(0.50)
Black African	−0.04	(0.22)	2.47***	(0.27)	0.01	(0.22)	2.22***	(0.59)
Other (including Chinese)	0.08	(0.25)	2.77***	(0.29)	0.94***	(0.25)	2.77***	(0.51)
Parents' highest academic qualification								
Higher degree	0.20	(0.16)	−0.19	(0.52)	−0.70**	(0.23)	1.22	(0.98)
Undergraduate degree	−0.02	(0.08)	−0.67*	(0.27)	−0.74***	(0.10)	−0.38	(0.72)
Diploma	−0.12	(0.10)	−0.46	(0.29)	−0.53***	(0.13)	−0.83	(0.70)
A Levels	0.07	(0.09)	−0.36	(0.33)	−0.16	(0.11)	0.71	(0.53)
GCSE (Grades A–C)	Ref.		Ref.		Ref.		Ref.	
GCSE (Grades D–G)	−0.00	(0.07)	0.43**	(0.14)	0.34***	(0.08)	0.52	(0.38)
None	0.02	(0.07)	2.31***	(0.11)	0.87***	(0.07)	1.54***	(0.34)
Child's age at interview (days)	0.00**	(0.00)	−0.01*	(0.00)	−0.00	(0.00)	−0.01	(0.01)
Constant	−1.61***	(0.47)	−1.02	(0.99)	0.40	(0.56)	−1.35	(2.58)
n	14,827							

Note: The data are adjusted to reflect the MCS survey design.

possible that the mothers and fathers genuinely changed occupations between these two data collections. In the multinomial logistic regression models presented in Tables 13 and 14, we include the child's age at the survey interview in days, to provide a measure of the length of time between the birth of the child and the survey interview. The mean age of the child at the time of the survey interview was 295 days (s.d. = 14), around 9.5 months of age (min = 244 days, max = 382 days). We have no details of when, within the 21-day (for Scotland) or 42-day (for England and Wales) stipulated period, the child's birth was registered. The child's age at the survey interview represents the maximum possible time difference between the recording of the parents' occupations on the birth record and the recording of their occupation in the survey. Including the age of the child at interview in the regression models allows us to investigate whether there is more change in occupations observed when more time has passed between the child's birth and the social survey interview. The multinomial logistic regressions indicate that there were only very small effects for the age of the child at interview on the likelihood of a mismatch between the data resources. We investigate this issue further by comparing the degree of occupational mismatch observed between the birth records and the survey, with the degree of occupational change that might be expected for adults in this stage of the life course.

Longhi and Brynin (2010) investigate the degree of occupational change over a year in the population using the British Household Panel Survey (BHPS). Taking into account changes in occupational codes and also reported changes in jobs, Longhi and Brynin find that around 8.1 per cent of men and women of working age change occupations in a year. We duplicate Longhi and Brynin's methodology using the BHPS data (SN5151, Institute

Table 14. Multinomial logistic regression of comparisons of SOC2000 between the survey and birth record for fathers. The base category is: SOC2000 matches

Parents' characteristics (from survey)	Does not match		Missing on both		Missing on birth record only		Missing on survey only	
	Versus matches		Versus matches		Versus matches		Versus matches	
	Log odds	SE	Log odds	SE	Log odds	SE	Log odds	SE
Mother's age at delivery								
Under 19	0.27*	(0.12)	1.96***	(0.12)	1.20***	(0.18)	1.38***	(0.12)
20–29	Ref.		Ref.		Ref.		Ref.	
30–39	−0.12*	(0.05)	−1.17***	(0.11)	−0.25*	(0.11)	−0.55***	(0.06)
Over 40	−0.35*	(0.15)	−1.45***	(0.37)	−0.85*	(0.42)	−0.61**	(0.21)
Mother's ethnicity								
White	Ref.		Ref.		Ref.		Ref.	
Mixed	0.03	(0.23)	1.47***	(0.29)	0.73*	(0.36)	0.73*	(0.29)
Indian	0.19	(0.15)	−0.87	(0.47)	−0.78*	(0.38)	0.07	(0.27)
Pakistani	0.47*	(0.19)	−0.20	(0.28)	0.83***	(0.20)	0.81***	(0.17)
Bangladeshi	−0.10	(0.19)	−1.23***	(0.31)	−0.06	(0.33)	0.25	(0.25)
Black Caribbean	0.05	(0.27)	2.09***	(0.25)	−0.07	(0.47)	1.48***	(0.30)
Black African	0.40	(0.25)	2.59***	(0.27)	0.62	(0.39)	1.53***	(0.27)
Other (including Chinese)	−0.02	(0.20)	0.24	(0.31)	0.42	(0.29)	−0.20	(0.29)
Parents' highest academic qualification								
Higher degree	0.10	(0.20)	−0.09	(0.38)	0.18	(0.39)	0.25	(0.26)
Undergraduate degree	0.07	(0.07)	−0.85***	(0.23)	0.07	(0.19)	−0.35**	(0.13)
Diploma	−0.12	(0.10)	−0.29	(0.19)	−0.10	(0.22)	0.16	(0.13)
A Levels	0.16	(0.09)	−0.00	(0.15)	0.18	(0.22)	0.08	(0.11)
GCSE (Grades A–C)	Ref.		Ref.		Ref.		Ref.	
GCSE (Grades D–G)	−0.03	(0.07)	−0.18	(0.11)	0.02	(0.14)	−0.24***	(0.07)
None	0.05	(0.06)	0.50***	(0.08)	0.54***	(0.13)	0.06	(0.08)
Child's age at interview (days)	−0.00	(0.00)	0.01*	(0.00)	0.01*	(0.00)	0.01***	(0.00)
Constant	0.75	(0.50)	−3.48***	(0.82)	−4.29***	(0.94)	−3.90***	(0.67)
n	14,827							

Note: The data are adjusted to reflect the MCS survey design.

for Social and Economic Research, 2010). We examine changes in jobs and occupations between Waves 10 and 11 of the BHPS which coincide with the period of the birth of the MCS cohort members. We look at the occupational change of women and men within 2 standard deviations of the mean age of the MCS mothers and fathers, and also the occupational change of only those woman and men who had a baby between sweeps 10 and 11 of the survey (see Table 15).

For those men within the same age range as the MCS parents, the change in occupations is approximately 4 per cent. For those who had a baby, less than 1 per cent of mothers changed occupations, whereas a greater percentage of men (9 per cent) changed occupations over this period (*n.b.* sample sizes become very small for this subsample and should be treated with suitable caution). The amount of change found here is far less than the degree of discrepancy observed between the occupations reported on birth records and the survey (see Table 6). It is unlikely that the high level of disagreement between the two data sources is due to genuine changes in occupations over the 9-month period between the registration of the birth and survey interview.

Table 15. Percentage of women and men who change occupation between sweeps 10 and 11 of the BHPS

	n	(%)
Women		
Age 16–40	1,307	(4%)
Age 16–40 and new baby	61	(<1%)
Men		
Age 20–44	1,515	(4%)
Age 20–44 and new baby	90	(9%)

Note: Based on the methodology described in Longhi and Brynin (2010) which defined occupational change as a change in occupational code and also a change in job between survey sweeps. The data are adjusted to reflect the BHPS survey design.

Question 3: What potential impact do disagreements have on empirical sociological analyses?

Socio-economic inequalities in test scores are strong and well reported, and children from less advantaged groups perform less well on these tests (see for example Feinstein, 2003; Sullivan *et al.*, 2013; Dickerson and Popli, 2016). Performance on cognitive tests in childhood is important because it is widely found to be associated with later educational attainment, and with occupational positions in adulthood (see Mascie-Taylor and Gibson, 1978; Jencks, 1979; Jensen, 1998; MacKintosh, 1998; Tittle and Rotolo, 2000; Sternberg *et al.*, 2001; Bartels *et al.*, 2002; Nettle, 2003; Schmidt and Hunter, 2004; Deary *et al.*, 2007; Connelly, 2012). We conduct a concise sensitivity analysis to compare the substantive conclusions which would be drawn in an analysis of social class inequalities in cognitive test scores if parental occupation-based measures were used from the survey or the birth records.

We consider two cognitive tests taken at different sweeps of the MCS. At age 5, we use the 'Naming Vocabulary' test, and at age 11, we use the 'Verbal Similarities' test. These are both subscales of the British Ability Scales, second edition (Elliott *et al.*, 1996). We use standardized test scores that are adjusted for the child's age, and the range of items which they have completed (see Connelly, 2013).

We run eight separate ordinary least squares (OLS) models with the cognitive test at age 5 or 11 years as the outcome. Each model contains an NS-SEC measure based on either the mother or father's occupational information derived from either the survey or the birth record. The models also control for the child's gender. To allow for comparison, the analytical sample comprises those sample members who completed the cognitive test and have occupational information available in both the survey and birth record. Due to attrition, the MCS sample size decreases between the first sweep and the third (age 5) and fifth (age 11) sweeps of the survey (see Platt, 2014). The final analytical sample sizes for the models are 8552 and 7600 for models comparing fathers' and mothers' measures, respectively, at age 5, and 7614 and 6710 for models comparing fathers' and mothers' measures, respectively, age 11. The coefficients and 95 per cent quasi-variance-based comparison intervals for the NS-SEC variables are presented in Figures 1 and 2 (full models are available in the supplementary materials). This sensitivity analysis indicates that the same substantive conclusions would be reached for these samples irrespective of whether the occupations reported in the birth records or survey are used. This is an encouraging finding; however, it should be noted that these models compare only those cohort members with mothers' and fathers' information available in both data resources and do not take into account the other patterns of missing data described above.

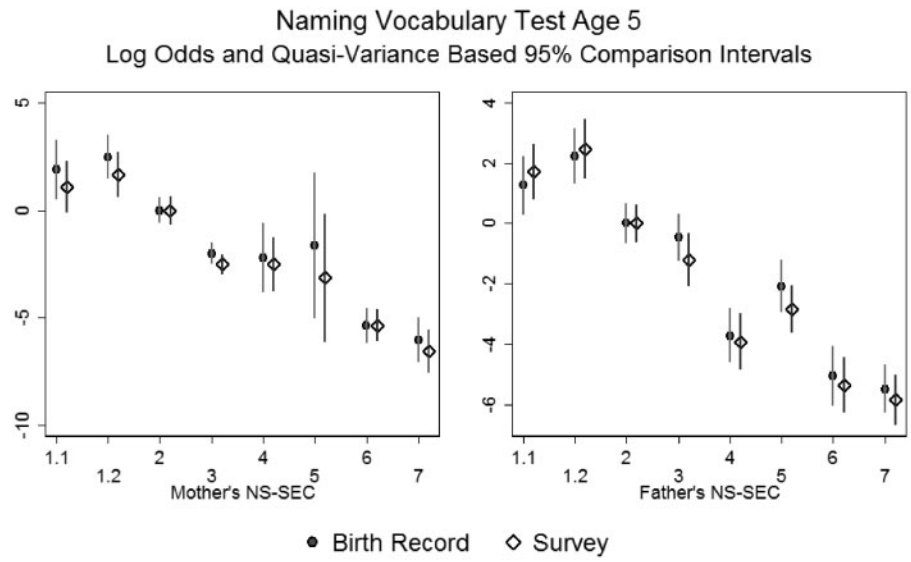

Figure 1. OLS regression models of naming vocabulary test scores at age 5.
Note: UK Millennium Cohort Study (SN4683 & SN5614). Models also contain gender. Models are adjusted for survey design. Models are run separately with mothers' and fathers' variables and include families with valid occupational information available for both the survey and birth record, *n* (fathers) = 8522, *n* (mothers) = 7600.

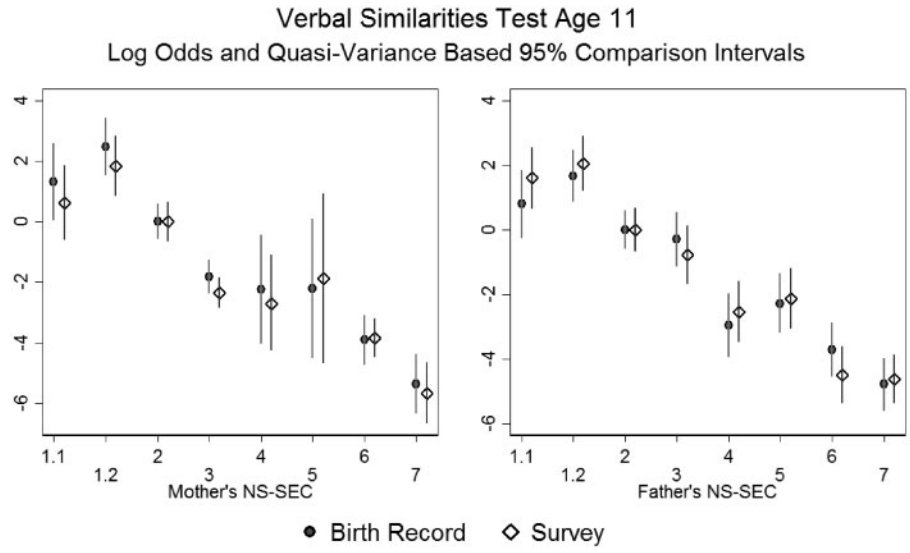

Figure 2. OLS regression models of naming verbal similarities test scores age 11.
Note: UK Millennium Cohort Study (SN4683 & SN5614). Models also contain gender. Models are adjusted for survey design. Models are run separately with mothers' and fathers' variables and include families with valid occupational information available for both the survey and birth record, *n* (fathers) = 7614, *n* (mothers) = 6710.

Conclusions

The new forms of administrative social science data that are emerging are likely to increase the scope and scale of empirical sociological inquiries. New infrastructural resources in the UK aim to be instrumental in improving access to administrative social science data. New sources of administrative social science data are emerging rapidly. In countries like the UK administrative data are haphazard and there are no national population registers against which to organize data. The empirical work undertaken in this article is original because it assesses the consistency of administrative birth records data using survey data collected from the same individuals. Measures derived from information on parental occupations are central to a wide spectrum of sociological analyses, and the occupational information on UK birth records will provide a central measure of social origins for sociological research on inequalities. A clear message from this work is that there are inconsistencies in the occupations reported in the birth records when compared with the information collected by professional interviewers shortly afterwards in a social survey. These findings are similar to US studies which have also examined occupational information on administrative birth records (see Carucci and Prasad, 1979; Brender *et al.*, 2008). This finding warns against the naïve or uncritical use of UK birth records data for sociological research.

It is fortuitous that data were available from the birth records for the participants in the MCS. Ordinarily researchers using administrative social science data will not have access to data that act as a comparative source. In these circumstances, researchers might reasonably be concerned about the quality of the administrative data. In one illustrative empirical example, we have shown that the inconsistencies in the birth records data have no appreciable influence on substantive conclusions. We strongly assert that this is not a necessarily general finding and must not be assumed *a priori*.

We advocate that further research is undertaken into the consistency and accuracy of UK birth records data. One potential strategy would be to compare parental occupational information within the birth records with official data collected for taxation (in the case of the UK, National Insurance information might provide a potential benchmark). Another potential strategy would be to compare parental occupational information on birth records with data collected from parents in a large-scale longitudinal study (for example, the UK Household Longitudinal Study). These analyses should also be extended to examine the quality of other sources of administrative social science data. A final comment is that in the changing climate of administrative social science data analysis, organizations engaged in collecting and curating information should be encouraged to place more emphasis on providing researchers with clear

information on the provenance of the data that they collect.

Notes

1 See: www.adrn.ac.uk.

2 Despite various official data collection and registration exercises that parents routinely have to engage with (for example, relating to children's health and enrolment at school), there is no single organized national activity that collects detailed information on parental occupations. The UK does not have national registers, identification numbers, or identification cards.

3 Northern Ireland is excluded from this analysis, as the Northern Irish data available to us were stored in an older standardized occupational classification than the other territories. More details of this analytical decision are provided in later sections of the article.

4 Standardized occupational codes organize job-related information (e.g. job titles) into a list of occupations. Examples include the UK Standard Occupational Classification (SOC) and the International Labour Organization's International Standard Classification of Occupations (ISCO).

5 The proportion of women misclassified as housewives is presented by occupation in this article. Of those mothers who report having a job in the interview, the per cent misclassified as housewives on the birth record varies from 0 per cent for those mothers working as 'health diagnosing and treating practitioners' (n = 14 in interview) to 65 per cent of those working in 'food preparation and serving occupations' (n = 14 in interview).

6 There was an overall achieved response rate of 68 per cent in the UK Millennium Cohort Study (Dex and Joshi, 2004).

7 Employment status defined whether an individual is an employer, self-employed, or employee; whether a supervisor; and the number of employees at their workplace. For more details of this measure see here: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/classifications/current-standard-classifications/soc2010/soc2010-volume-3-ns-sec–rebased-on-soc2010–user-manual/index.html.

8 The data deposited in the UK Data Archive and available to us as researchers are a sub-set of all of the data which are collected in administrative birth records.

9 To investigate the differences in NS-SEC classification when the full (i.e. with employment status) and simplified (i.e. without employment status) derivation methods are used, we have coded the MCS mothers' and fathers' occupational information using both methods. For mothers, there was 86 per cent agreement between the two measures (K = 0.83, r = 0.96, P = 0.001). For fathers, there was 78 per cent agreement between the two measures (K = 0.75, r = 0.92, P = 0.001).

10 For Scotland see: https://www.citizensadvice.org.uk/scotland/relationships/birth-certificates-and-changing-your-name-s/birth-certificates-s/ [accessed 01/06/2016]. For England and Wales see: http://www.oneplusone.org.uk/content_topic/married-or-not/children/ [accessed 01/06/2016].

11 We use mother's information only for the age and ethnicity variables to reduce the amount of missingness.

Acknowledgement

We are grateful to the participants of the Millennium Cohort Study and the British Household Panel Study. We thank the Centre for Longitudinal Studies, UCL Institute of Education, and the Institute of Social and Economic Research, University of Essex for the use of these data. We also acknowledge the UK Data Archive and Economic and Social Data Service for making these data available to us. These organizations bear no responsibility for the analysis or interpretation of these data. We are grateful for helpful comments on an earlier draft of this manuscript from two anonymous referees as well as feedback from participants at the 2015 Social Stratification Seminar, University of Milan.

Supplementary Data

Supplementary data are available at *ESR* online.

Funding

This work was supported by the Economic and Social Research Council [Grant Number: ES/N011783/1].

References

Administrative Data Taskforce. (2012). *The UK Administrative Data Research Network: Improving Access for Research and Policy*. London: Economic and Social Research Council.

Bakeman, R. *et al*. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, **2**, 357–370.

Bartels, M. *et al*. (2002). Heritability of educational achievement in 12-year-olds and the overlap with cognitive ability. *Twin Research*, **5**, 544–553.

Brender, J. D., Suarez, L. and Langlois, P. H. (2008). Validity of parental work information on the birth certificate. *BMC Public Health*, **8**, 95–106.

Bukodi, E. and Goldthorpe, J. H. (2012). Decomposing 'social origins': the effects of parents' class, status, and education on the educational attainment of their children. *European Sociological Review*, **29**, 1024–1039.

Burrows, R. and Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data and Society*, **1**, 1–6.

Carucci, P. M. and Prasad, S. (1979). A comparison of mothers' occupations reported on live births certificates and on a survey questionnaire. *Public Health Reports*, **94**, 432–437.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**, 37–46.

Connelly, R. (2012). Social stratification and cognitive ability: an assessment of the influence of childhood ability test scores and family background on occupational position across the lifespan. In Lambert P. S., Connelly R., Blackburn M., and Gayle V. (Eds.), *Social Stratification: Trends and Processes*. Aldershot: Ashgate, pp. 101–113.

Connelly, R. (2013). *Millennium Cohort Study Data Note 2013/1: Interpreting Test Scores*. London: Centre for Longitudinal Studies.

Connelly, R., Gayle, V. and Lambert, P. (2016). Statistical modelling of key variables in social survey data analysis. *Methodological Innovations Online*, **9**, 1–17.

Connelly, R. and Platt, L. (2014). Cohort profile: UK millennium Cohort study (MCS). *International Journal of Epidemiology*, **43**, 1719–1725.

Connelly, R. *et al*. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, **59**, 1–12.

Deary, I. *et al*. (2007). Intelligence and educational achievement. *Intelligence*, **35**, 13–21.

Dex, S. and Joshi, H. (2004). *Millennium Cohort Study First Survey: A User's Guide to Initial Findings*. London: Centre for Longitudinal Studies, Institute of Education, University of London.

Dex, S., Rosenberg, R. and Hawkes, D. (2008). *Ethnic Minorities and Non-response in the Millennium Cohort Study*. London: Centre for Longitudinal Studies, Institute of Education, University of London.

Dickerson, A. and Popli, G. (2016). Persistent poverty and children's cognitive development: evidence from the UK Millennium Cohort Study. *Journal of the Royal Statistical Society: Series a (Statistics in Society)*, **179**, 534–558.

Elias, P. (2014). Administrative data. In Duşa A., Nelle D., Stock G., and Wagner G. (Eds.), *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*. Berlin: SCIVERO, pp. 47–48.

Elias, P., Halstead, K. and Prandy, K. (1993). *Computer Assisted Standard Occupational Classification*. London: HMSO.

Elliott, C. D., Smith, P. and McCulloch, K. (1996). *British Ability Scales Second Edition (BAS II). Technical Manual*. London: Nelson.

Feinstein, L. (2003). Inequality in early cognitive development of British children in the 1970 cohort. *Economica*, **70**, 73–97.

Fleiss, J. L., Levin, B. and Paik, M. C. (2013). *Statistical Methods for Rates and Proportions*. Hoboken: John Wiley & Sons.

Goerge, R. M. and Lee, B. J. (2001). Matching and cleaning administrative data. In Citro C. F., Moffitt R. A., and Van Ploeg M., (Eds.), *Studies of Higher Population: Data Collection and Research Issues*. Washington DC: National Academies Press, pp. 197–219.

Goerge, R. M. *et al*. (1992). Special-education experiences of foster children: an empirical study. *Child Welfare: Journal of Policy, Practice, and Program*, **71**, 419–437.

Graham, H. (2007). *Unequal Lives: Health and Socioeconomic Inequalities*. Maidenhead: Open University Press.

Grätz, M. (2015). When growing up without a parent does not hurt: Parental separation and the compensatory effect of social origin. *European Sociological Review*, **31**, 546–557.

Halfpenny, P. and Procter, R. (2015). *Innovations in Digital Research Methods*. London: Sage.

Hockley, C. *et al*. (2007). Millennium cohort study: birth registration and hospital episode statistics linkage. *Paediatric Perinatal Epidemiology* **22**, 99–109

Institute for Social and Economic Research (2010). British Household Panel Survey: Waves 1-18, 1991-2009 [data collection], 7th edn. UK Data Service. SN:5151. Colchester: University of Essex, UK Data Service.

Jencks, C. (1979). *Who Gets Ahead? The Determinants of Economic Success in America*. New York, NY: Basic Books.

Jensen, A. (1998). *The G Factor: The Science of Mental Ability*. Westport: Praeger.

Jones, R. (2004). *CASCOT (Computer Aided Structured Coding Tool)*. Warwick: University of Warwick.

Ketende, S. and Jones, E. (2011). *User Guide to Analysing MCS Data Using Stata*. London: Centre for Longitudinal Studies.

Kiernan, K. (2006). Non-residential fatherhood and child involvement: evidence from the Millennium Cohort Study. *Journal of Social Policy*, **35**, 651–669.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data and Society*, **1**, 2053951714528481.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Lipsky, M. (1979). *Street Level Bureaucracy*. New York, NY: Russell Sage Foundation.

Longhi, S. and Brynin, M. (2010). Occupational change in Britain and Germany. *Labour Economics*, **17**, 655–666.

MacKintosh, N. J. (1998). *IQ and Human Intelligence*. Oxford: Oxford University Press.

Maldonado, S. (2011). Illegitimate harm: law, stigma, and discrimination against nonmarital children. *Florida Law Review*, **63**, 345.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. *Debates in the digital humanities*, **2**, 460–475.

Mascie-Taylor, C. and Gibson, J. (1978). Social mobility and IQ components. *Journal of Biosocial Science*, **10**, 263–276.

Nettle, D. (2003). Intelligence and class mobility in the british population, *British Journal of Psychology*, **94**, 551–561.

Office for National Statistics (2000). *Standard Occupational Classification. Volume 1: Structure and Descriptions of Unit Groups*. London: Office for National Statistics.

Office for National Statistics (2010). *Standard Occupational Classification 2010*. Baingstoke: Palgrave MacMillan.

Parkin, F. (1971). *Class Inequality and Political Order: Social Stratification in Capitalist and Communist Societies*. New York, NY: Praeger.

Platt, L. (Ed.) (2014). *Millennium Cohort Study: Initial Findings from the Age 11 Survey*. London: Centre for Longitudinal Studies.

Plewis, I. (2007). Non-response in a birth cohort study: the case of the Millennium Cohort Study. *International Journal of Social Research Methodology*, **10**, 325–334.

Plewis, I. *et al*. (2004). *Millennium Cohort Study: Technical Report on Sampling*. London: Centre for Longitudinal Studies.

Prandy, K. (1999). The social interaction approach to the measurement and analysis of social stratification. *International Journal of Sociology and Social Policy*, **19**, 204–236.

Rose, D., Pevalin, D. J. and O'Reilly, K. (2005). *The National Statistics Socio-economic Classification: Origins, Development and Use*. Basingstoke: Palgrave Macmillan.

Schmidt, F. and Hunter, J. (2004). General mental ability in the world of work: occupational attainment and job performance. *Journal of Personality and Social Psychology*, **86**, 162–173.

Schroeder, R. (2014). Big data and the brave new world of social media research. *Big Data and Society*, **1**, 2053951714563194.

Shaw, G. *et al*. (1990). An assessment of error in parental occupation from the birth certificate. *American Journal of Epidemiology*, **131**, 1072–1079.

Sternberg, R., Grigorenko, E. and Bundy, D. (2001). The predictive value of IQ. *Merrill-Palmer Quarterly*, **47**, 1–41.

Sullivan, A., Ketende, S. and Joshi, H. (2013). Social class and inequalities in early cognitive scores. *Sociology*, **47**, 1187–1206.

Tittle, C. and Rotolo, T. (2000). IQ and stratification: an empirical evaluation of Herrnstein and Murray's social change argument. *Social Forces*, **79**, 1–28.

Treiman, D. J. (1977). *Occupational Prestige in Comparative Perspective*. New York, NY: Academic Press.

UCL Institute of Education (2008). Millennium Cohort Study, 2001–*2003: Birth Registration and Maternity Hospital Episode Data [computer file]*, 3rd edn. UK Data Service SN:5614. Colchester, Essex: UK Data Archive [distributor].

UCL Institute of Education (2012). *Millennium Cohort Study: First Survey, 2001–2003 [computer file]*, 11th edn. SN: 4683. Colchester, Essex: UK Data Archive [distributor].

United Nations (2007). *Register-based Statistics in the Nordic Countries. Review of Best Practices with Focus on Population and Social Statistics*. New York, NY: United Nations.

Woollard, M. (2014). Administrative data: problems and benefits. a perspective from the United Kingdom. In Duşa A., Nelle D., Stock G., and Wagner G. G. (Eds.), *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*. Berlin: SCIVERO, pp. 49–60.

Roxanne Connelly is Assistant Professor of Sociology at the University of Warwick, UK. Her work is focused in the areas of Social Stratification and the Sociology of Education. She is a specialist in quantitative research methods, particularly longitudinal data analysis. Her research has an interdisciplinary focus and integrates insights from Psychology and Sociology.

Vernon Gayle is Professor of Sociology and Social Statistics at the University of Edinburgh, UK. His main research interests include social stratification, the sociology of youth and the sociology of education. But he also has research interests in migration, subjective well being, populations, fertility, digital social research and the sociology of sport. His methodological interests include the analysis of large-scale and complex social science datasets, statistical modelling, longitudinal data, administrative social science data, missing data methods and social networks.