University of Glasgow

Adam Smith
Business School

WORKING PAPER SERIES

Is the social origin pay gap bigger than we thought? Identifying and acknowledging workers with undefined social origins in survey data

Michael Vallely, Jeanette Findlay and Kristinn Hermannsson

Paper No. 2023-06
February 2024

Is the social origin pay gap bigger than we thought? Identifying and acknowledging workers with undefined social origins in survey data

Michael Vallely, Jeanette Findlay & Kristinn Hermannsson

23rd October 2022

Abstract

This paper investigates whether recent empirical studies have underestimated the social origin pay gap by omitting respondents with undefined social origins. Specifically, individuals that were not assigned a social origin because the identity of their parental household was not clear, nobody was earning in the household, or the occupational identity of the main wage-earner could not be identified. Data from the UK Quarterly Labour Force Survey is analysed to establish the prevalence of undefined social origins and the extent to which the socioeconomic characteristics of those with undefined social origins are different from those who can be identified using the SOC Classification. We then examine how omitting these groups affects estimates of social origin pay gaps. The results show that 11% of the working age population are from undefined social origins and that the labour market outcomes of these people are on average much worse than those with defined social origins. Results show that omitting these respondents underestimates the range of the social origin pay gap and the number of people affected. This highlights that there is a further effect of parental association with the labour market or not clearly belonging to a household which profoundly affect the life outcomes of a substantial share of the working age population.

Keywords: item non-response, labour market outcomes, pay gaps, social origin

Journal of Economic Literature Classification Numbers: E24, J3, J7

1 Introduction

A recent wave of empirical work has identified the existence of social origin pay gaps that are unexplained, i.e. they persist even when observable characteristics such as education and a range of labour market observables have been controlled for. These estimates have been obtained by applying established analytical approaches to a variety of datasets for the UK, US and other high-income countries. Prime facie, this phenomenon is analogous to other pay gaps and has been referred to as the "class ceiling" (Laurison & Friedman, 2016), referencing similarities to the gender pay gap. However, asking about social origin in a survey is arguably more complicated than asking about gender or racial identities, notwithstanding that these can also be challenging issues on which survey respondents define or are defined. As we shall see, the specifics of how social origin is derived can have a marked impact on results. Scrutiny of social origin data from the UK Quarterly Labour Force Survey (LFS) reveals that in practice substantive groups from non-traditional backgrounds and households with less structured occupational profiles are not identified in the Standard Occupational Classification[1]. Overall, the social origin of 11% of working age respondents is undefined, corresponding to approximately 4.7 million individuals. Conceptually, this is consistent with the view that occupation-based classification of social origin is a circumscribed instrument for capturing the diverse ways through which social class intersects with labour market disadvantage in the 21st century. Empirically, this suggests that the prevalence and intensity of the social origin pay gap is underestimated.

1.1 Earnings gaps

Research into the social origin pay gap draws on methods and insights from research on social mobility, returns to education and gender and minority pay-gaps. Following Mincer (1974) wage equations fitted on cross-sectional data from around the world reveal that on average, more qualified individuals are better off in terms of employment and earnings than less qualified individuals (e.g. Psacharopoulos & Patrinos, 2004; Walker & Zhu, 2008, 2011, 2013; Montenegro & Patrinos, 2014; Conlon & Patrignani, 2013). For this reason, investing in education has been seen not only to improve economic competitiveness (Krueger & Lindahl, 2001; Hermannsson et al., 2014; LSE Growth Commission, 2013; OECD, 2012) but also to aid social mobility (see Duta & Ianelli, 2018 for a critical discussion).

Although qualifications are a key predictor of earnings, other empirical insights suggest income inequality is more complicated. Studies of occupational mobility show a persistence across generations in occupational attainment (e.g. Bukodi &

[1] The Standard Occupational Classification (SOC) is a coding framework used in the UK to classify occupations, enabling comparisons of occupations across different datasets.

Goldthorpe, 2011) and educational attainment (Shavit, 2007). In recent years, availability of social origin data in social surveys has revealed an even more insidious form of intergenerational income persistence, where the previous generation's occupational status influences earnings, even when controlling for education, experience and occupational attainment. This effect has been observed in different types of data, such as a graduate follow up survey (Crawford & Vignoles, 2014), a cohort study (Crawford & van der Erve, 2015), large scale administrative data (Britton et al, 2016) and the Labour Force Survey (Friedman et al, 2017; Friedman & Laurison, 2017, 2019; Laurison & Friedman, 2016). Similar effects have been observed for Scandinavian countries, Spain and the US (Bernardi & Gil-Hernandez, 2021; Masketasa, 2011; Hallsten, 2013; Hersbein & Bartik, 2016).

The analogy of the "class ceiling" (Laurison & Friedman 2016, Friedman & Laurison 2020), rests on similarities with the gender pay gap, which also persists despite observable features being controlled for (Blinder, 1973; Blau & Kahn, 2017; Chevalier, 2007; Arulampalam et al, 2007; Fortin e al, 2017), as is often highlighted through use of decomposition techniques (e.g. Manning & Robinson, 2004; Fortin et al 2011). This approach has been extended to other sub-groups, such as ethnic minorities (Brynin & Güveli, 2012; Longhi & Brynin, 2017; Rafferty, 2012), disabled people (Berthoud, 2008), LGBT people (Bridges & Mann, 2019) and those living in rural locations (Culliney, 2017).

When it comes to collecting survey data, a crucial difference between social origin compared to sex, ethnicity or sexual orientation is that the latter are all features of the respondent as a person, whereas social origin is derived from response to a series of questions recollecting the status of a previous generation. Prime facie, the more questions that are required to derive a variable, the more likely it becomes that the variable cannot be constructed, as data could be missing for any of several underlying questions. If non-response to any of the questions is systemic then the resulting variable is likely to be biased (for overview of issues and mitigation strategies see: Jelke et al, 2011; Särndal & Lundström 2005; Groves et al, 2002; Groves & Couper, 2012).

1.2 Socioeconomic classification in survey data

To identify the socioeconomic status of a survey respondent's household, the Office for National Statistics (ONS) in the UK deploys the National Statistics Socioeconomic Classification (ONS 2009, p. 102). This approach was developed in sociological research (e.g. Goldthorpe 1980, 1987, 1997; Erikson & Goldthorpe, 1992) and also underpins the European Socio-Economic Classification (Rose & Harrison, 2007, 2014). The Labour Force Survey user guide explains that "the decision to adopt the Goldthorpe classification as the basis for the NS-SEC was made because it is widely used and accepted internationally", (ONS, 2009, p. 102). Prior to this, the ONS had commissioned the Economic and Social Research Council (ESRC) to conduct a review of social classifications (for an overview of findings see

Rose & Pevalin, 2003). The NS-SEC is an occupationally based classification. First, the person that is judged to best define the household position, the Household Reference Person, is identified[2]. An NS-SEC category is derived from a series of questions about employment status and occupation, which are coded to the Standard Occupational Classification 2010 (SOC2010).

The NS-SEC is underpinned by the argument that occupational conditions shape social conditions (Connelly et al, 2016; Rose & Pevalin, 2001, 2003). The development and origins of the scheme is summarised by Rose & Pevalin (2001). Each NS-SEC class is created by analysing employment relations data to identify combinations of occupational groups and employment status sharing similar employment relations. This is then mapped against an occupational classification scheme. As Connelly et al (2016) point out in their review of occupation-based social classifications, the empirical and conceptual merits of different approaches are debated. An enduring problem of occupational indicators is "the complexity of making comparison over time when the underlying structure of the labour market has changed" (Connelly et al, 2016, p. 9). Moreover, as Lambert and Bihagen (2014) show in their simulation exercise, results are sensitive to both the indicator used and the level of disaggregation for which it is derived. These are well established criticisms of the approach, which researchers need to be conscious of. However, specific additional challenges arise when occupational status is derived from the previous generation as in social origin indicators and, in turn, when these are related to earnings data, as in the social origin pay gap. Moreover, researchers have acknowledged the limitations of using parental occupation as a proxy for social class (e.g. Friedman and Laurison, 2019).

1.3 Who are the people with undefined social origins?

Since 2014, the LFS has included data capturing additional dimensions of social status, namely social origin, as proxied by the occupational status of the previous generation. In a series of questions, respondents are asked about their household composition when they were 14, who was the main earner in the household and what was their occupation. Occupational information is again coded according to the Standard Occupational Classification 2010 (SOC 2010).

From the point of view of conducting social surveys, the benefit of occupational classifications is that they can be operationalised through a handful of questions. In practice, however, the approach does not produce comprehensive data as social origin cannot be identified for a substantial minority of respondents (around 11% in the LFS as we will see in Section 2.1). This can be problematic if those with

[2] According to the ONS the Household Reference Person (HRP) is identified as the person responsible for owning or renting or who is otherwise responsible for the accommodation. In the case of joint householders, the person with the highest income takes precedence and becomes the HRP. Where incomes are equal, the oldest person is taken as the HRP. For details see: https://www.ons.gov.uk/methodology/classificationsandstandards/otherclassifications/thenationalstatisticssocioeconomicclassificationnssecrebasedonsoc2010#history-and-origins

undefined social origins are a non-random sub-population. A priori, this is likely to be the case for at least two reasons.

First, as social origin relies on recall of household composition and occupational status of parents when respondent was 14, social origin is undefined for individuals not living with parents at that age. This becomes salient when that data is used to analyse labour market disadvantage. Individuals who do not live with their family during their adolescence are more likely to come from non-traditional or fragmented households. This occurs for a multitude of reasons and may include individuals living in care, individuals whose parents died before they were 14 or those whose parent(s) were imprisoned. Evidence suggests that family instability can adversely affect children in many ways (Fomby and Cherlin, 2007). A range of studies has evidenced that living in care has an enduring impact on several socio-economic outcomes including reduced educational attainment, increased homelessness and unemployment and lower income and socio-economic status (Bywaters et al, 2016; Gypen et al, 2017; Harker et al, 2004; Jackson and Sachdev, 2001; Viner and Taylor, 2005). Evidence from the 1970 British Cohort Study used to examine outcomes for individuals at ages 16 and 30 found that when compared with individuals in foster care, residential care was associated with several poorer outcomes including mental health, life satisfaction and self-efficacy (Dregan and Gulliford, 2012).

Second, occupational classifications have been criticised for being overly rigid and imposing a static view of occupational classes, which represents the economic structure at its inception but misses the dynamics of economic relations (Connelly et al, 2016, Rose & Pevalin, 2001). Moreover, an implicit assumption is that occupational status is clear, that there is an understood occupational identity. However, this may not always be the case, especially in more precarious and informal employment where odd jobs may be combined into a more fragmented livelihood. This becomes even more problematic when identifying social origin because the less clear the occupational identify was for the first generation, the less likely the second generation is to recall the occupation of the main wage earner in their household when they were growing up.

1.4 Objectives

The premise of this article is that those who do not fit the occupational classification are among those that we should be most interested in knowing about in order to understand the impact of social origin on labour market outcomes. To test this, several objectives have to be achieved.

The first objective is to scrutinise the process through which social origin information is derived and identify sub-groups of respondents by the technical reason why their social origin is undefined. A second objective is to evaluate ex post whether undefined social origin is non-random by comparing observable traits of respondents with defined and undefined social origins. A third objective is to evaluate ex-post whether undefined social origins are associated with labour market disadvantage. A fourth objective is to assess whether omission of respondents with undefined social origins has led to biased estimates of social origin pay gaps; and in that event establish the likely direction and magnitude of the bias.

2 Comparing those with defined and undefined social origins using the Quarterly Labour Force Survey

In this section we review how the social origin variable used in the Quarterly Labour Force Survey is derived and explore whether there is likely to be systemic non-response to this variable. We use the UK Quarterly Labour Force Survey between 2014 when information for social origin was first included through to 2021 which is the latest data available. The LFS is the largest employment survey in the UK and provides nationally representative data[3]. The benefits of such datasets have been emphasised by other scholars (Charlwood et al, 2014). We use the third quarter as this is when the social origin questions are administered. Where the sample is extrapolated to obtain population-level estimates, this is done for a single year 2019 based on population weights provided in the LFS. 2019 was chosen as the most recent year prior to the Covid-19 pandemic. For some analyses we pool data for all the years in order to reduce influence of sampling variation. We also run analyses separately for each year in order to examine the sensitivity of results to specific waves, which we find to be immaterial (see Appendix). As the focus is on the working age population, we omit all respondents that are not of working age, i.e. under 16 or over 70. Moreover, as the LFS is administered to the same respondents over five consecutive quarters, a number of respondents will be observed twice when waves are pooled. We omit respondents that have information brought forward from a previous wave, i.e. when the social origin questions were not asked and would therefore be coded as 'does not apply'. For 2019 this results in an analytical sample of 46,533 (see Table 1), which corresponds to a working age population of 43,155,629. For details of the pooled sample see Appendix.

[3]For methodological background of the LFS please see technical guidance from the Office of National Statistics: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/labourforcesurveylfsqmi#methodology-background

Table 1 Analytical sample and population estimate

Occupation of main wage earner when respondent was 14 years old (Major)	No. of observations in sample	% of (unweighted) sample	Estimated population in 2019	Estimated % of working age population (weighted)
Does not apply	4,667	10.0	4,549,117	10.5
No answer	236	0.5	231,874	0.5
Managers, directors and senior officials	5,602	12.0	5,385,378	12.5
Professional occupations	7,318	15.7	7,102,452	16.5
Associate professional and technical occupations	3,787	8.1	3,585,071	8.3
Administrative and secretarial occupations	2,298	4.9	2,164,163	5.0
Skilled trades occupations	9,895	21.3	8,654,481	20.1
Caring, leisure and other service occupations	1,542	3.3	1,515,458	3.5
Sales and customer service occupations	1,556	3.3	1,479,788	3.4
Process, plant and machine operatives	5,326	11.5	4,683,004	10.9
Routine occupations	4,306	9.3	3,804,843	8.8
Total	**46,533**	**100.00**	**43,155,629**	**100.00**

2.1 Social origin in the LFS

From 2014 onwards, the LFS provides a variable for social origin, identifying the occupational classification of the previous generation in line with the SOC 2010 occupational classification[4]. The social origin variable (SMSOC101) identifies the occupation of the main wage earner when the respondents was 14 years old. However, the question is not administered unless a satisfactory answer has been obtained for two underlying questions. A summary of these three variables and how they can each contribute to social origin being undefined is provided in Figure 1 below.

First, respondents are asked about their household composition when they were 14 years old (SMHCOMP). Social origin will not be identified unless a respondent was either living with parent(s) or living with other family members at this age. Consequently, respondents who were not living with their family when they were 14 drop out at this stage and therefore the main social origin question (SMSOC101) does not apply to them.

[4] It should be noted that the ONS does not derive the NS-SEC categories of the previous generations, only their occupational classification, but in the past researchers have applied a coding rubric to map the occupational classification onto NS-SEC, see e.g. Laurison & Friedman (2016). For further details see the LFS User Guide, Vol. 5, Section 5.1.

If respondents were living with one or both parents or other family members, they are then asked to identify the main wage earner when they were 14 years old. If nobody in the household was earning at that time the social origin question will not be administered.

If a respondent identifies a main wage earner when they were 14, they are then asked what the occupation of the main wage earner in their household was (SMSOC101). The response to this question, if given, is matched to a SOC code. Social origin can be undefined at this stage if an answer is not provided or if the response cannot be classified. In wave 3 of the LFS 2019 data, information on social origin is not available for 10.5% of respondents, the majority of which are coded as 'does not apply'. This is a substantial share of the UK working age population, approximately equivalent to the combined working age populations of Scotland and Northern Ireland or the Northwest of England[5].

Figure 1 How social origin (SMSOC101) is derived and conditions for social origins to be classified. Variable names in brackets.

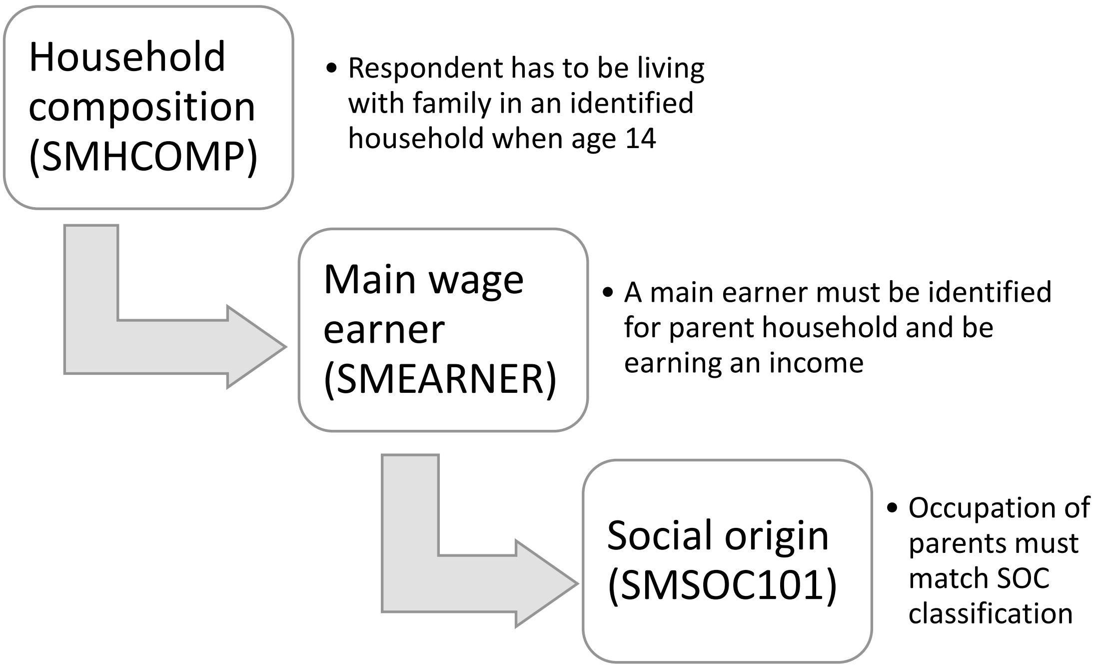


Table 2 further disaggregates respondents whose social origins are not defined and reveals at what stage in the survey process their social origins became undefined. The largest group are respondents where no-one was earning when they were age 14, accounting for 49.3% of those with undefined social origins and 5.2% of the sample. The second largest group contains those for whom the occupational identity of the previous generation could not be classified, i.e. the question was answered but the response could not be classified as a SOC code for

[5] In 2019 the working age population of Scotland and Northern Ireland accounted for 8.3% and 3.3%, respectively and therefore stands at 11.6%. Another comparison is with the Northwest of England, encompassing Greater Manchester, Merseyside and the rest of the Northwest, which accounted for 10% of the UKs working age population in 2019.

the occupation (30.1% of those with undefined social origins and 3.2% of respondents). Jointly these two reasons account for nearly 80% of all undefined social origins. Furthermore, 15.7% of those with undefined social origins were not living with family or their household composition was unclear. The least important category in this regard are respondents explicitly not answering the question which are only 0.5% of respondents and 4.8% of all undefined social origins.

Table 2 Undefined social origins: disaggregation of missing data fields for the occupation of main earner when respondent was 14 years old (SMSOC101)

<table>
<tr><th colspan="2">Response category</th><th>No. of obs in sample</th><th>% of unweighted sample</th><th>% of respondents with undefined social origins</th><th>Est working age population in 2019</th><th>Est % of working age population with undefined social origins</th></tr>
<tr><td colspan="2">No answer</td><td>236</td><td>0.5</td><td>4.8</td><td>231,874</td><td>4.8</td></tr>
<tr><td rowspan="3">Not classified</td><td>Not living with family or household composition at age 14 unclear</td><td>772</td><td>1.7</td><td>15.7</td><td>769,890</td><td>16.1</td></tr>
<tr><td>No-one was earning in household when respondent was 14 or not clear who was main earner</td><td>2,417</td><td>5.2</td><td>49.3</td><td>2,309,828</td><td>48.3</td></tr>
<tr><td>Occupation not identified</td><td>1,478</td><td>3.2</td><td>30.1</td><td>1,469,399</td><td>30.7</td></tr>
<tr><td colspan="2">Total</td><td>4,903</td><td>10.5</td><td>100.0</td><td>4,780,991</td><td>100.0</td></tr>
</table>

2.2 Does missingness appear random?

In order to evaluate ex-post whether undefined social origins in the LFS appear random, Table 3 compares selected observed features of those with defined and undefined social origins respectively. This comparison reveals differences, which are statistically significant with the exception of gender composition. Those with undefined social origins tend to be younger by about three years on average; almost half as likely to belong to a visible minority; more likely to have responded to the survey via proxy; more likely to have no qualifications; less likely to hold a degree; more likely to be on benefits; less likely to be married; more likely to be in rented accommodation; less likely to be in work; less likely to have reached higher occupational destinations; and receive 28% lower hourly pay than respondents with defined social origins. In summary those with undefined social origins are demographically and socially different from those whose social origin we can define – they are disadvantaged in terms of several life outcomes, such as educational attainment, housing tenure, occupational attainment and earnings.

Based on our scrutiny of how the social origin question is derived and comparison of observed features of those with defined and undefined social origins, it is clear that undefined social origins are not a coincidence.

Table 3 Undefined social origins: Comparison of selected observed attributes between those with defined social origins and those with undefined social origins.

	Social origin				% difference	t value	
	defined		undefined				
	n	mean	n	mean			
age in years	41,630	45.0	4,903	42.2	-7%	12.4	***
male	41,630	47.6%	4,903	46.1%	-3%	1.9	*
visible minority	41,630	10.0%	4,903	17.9%	44%	-17.1	***
disability	41,630	21.1%	4,903	28.4%	26%	-11.6	***
proxy response	41,630	32.9%	4,903	43.4%	24%	-14.7	***
no qualifications	41,630	8.2%	4,903	17.3%	53%	-21.0	***
degree holder	41,630	19.9%	4,903	13.1%	-52%	11.4	***
post-graduate degree holder	41,630	11.7%	4,903	6.1%	-92%	11.7	***
receiving benefits	41,630	32.0%	4,903	39.2%	18%	-10.2	***
married	41,630	59.9%	4,903	45.8%	-31%	19.1	***
living in rented accommodation	41,630	26.9%	4,903	50.6%	47%	-34.9	***
in work	41,630	71.0%	4,903	64.5%	-10%	9.3	***
occupational destination: NS-SEC 1	41,630	14.5%	4,903	8.9%	-63%	10.7	***
occupational destination: NS-SEC 2	41,630	24.2%	4,903	16.6%	-46%	12.0	***
hourly pay in £	8,935	16.4	692	12.8	-28%	2.9	***

*** p<0.01, ** p<0.05, * p<0.1

Table 4 expands this comparison by benchmarking each of the groups with undefined social origins against those whose social origins are defined. First, we look at those who do not provide an answer to the social origin question. It is probable that this form of non-response is largely random as estimated differences are small and insignificant, with the notable exception that those belonging to visible minorities are substantially overrepresented in the 'no answer' group. For the other three groups, there are substantial and significant differences in their labour market outcomes, with undefined social origins associated with 16-19% earnings gap, much lower representation in higher occupational destinations and lower likelihood of being in work for those from non-traditional households or where no-one in the household was earning. Further inspection reveals that the three groups are demographically different from those with defined social origins, being slightly younger on average and much more likely to belong to a visible minority. They are less likely to hold a degree and more likely to have no qualifications. They are further disadvantaged through weaker housing tenure and those from non-traditional and non-earning households are more likely to be in receipt of benefits.

Given these multiple forms of disadvantage associated with undefined social origin, can we simply treat those with undefined social origins as if they were from routine social origins? This is explored in Table 5 by comparing those with undefined social origins to those from social origins in routine occupations (i.e. SOC9). Overall, this comparison reveals a mixed picture. When compared on earnings, respondents with undefined social origin are similar to those from routine origins and differences are insignificant. However, the two groups are significantly different in terms of their demographic makeup. The three undefined social origin groups are younger and more likely to belong to a visible ethnic minority. These groups are also different in terms of educational attainment, with the undefined social origin groups having more polarised outcomes. Two of the undefined origin groups (non-traditional household and non-earning household) are more likely to have no qualifications than the SOC9 group but all three undefined groups are more likely to hold degrees compared to those from routine origins i.e. those at the bottom of the occupational scale.

Table 4 Undefined social origins: Comparison of selected observed attributes between those with defined social origins and those with undefined social origins in 2019, separately identifying 6 different sub-groups based on what caused their social origin to be undefined.

Attribute	Social origin												
					undefined								
	classified (n=41,630)	No answer (n=236)			Household composition at age 14 unclear (n=772)			No-one was earning in household when respondent was 14 (n=2,417)			Occupation not identified (n=1,478)		
	mean	mean	% difference	t value	mean	% difference	t value	mean	% difference	t value	mean	% difference	t value
age in years	45.0	43.3	**-4%**	1.8 *	45.1	**0%**	-0.4	39.8	**-12%**	16.6 ***	44.4	**-1%**	1.5
male	48%	43%	**-9%**	1.4	49%	**2%**	-0.6	44%	**-9%**	3.9 ***	49%	**4%**	-1.5
visible minority	10%	15%	**48%**	-2.5 **	20%	**96%**	-8.8 ***	17%	**65%**	-10.2 ***	20%	**101%**	-12.6 ***
disability	21%	23%	**9%**	-0.7	31%	**45%**	-6.4 ***	33%	**57%**	-14.1 ***	20%	**-5%**	1.0
proxy response	33%	38%	**15%**	-1.6	36%	**9%**	-1.7 *	39%	**18%**	-6.0 ***	56%	**69%**	-18.4 ***
no qualifications	8%	10%	**24%**	-1.1	18%	**113%**	-9.3 ***	19%	**127%**	-17.7 ***	16%	**96%**	-10.8 ***
degree holder	20%	20%	**-2%**	0.2	13%	**-34%**	4.6 ***	11%	**-44%**	10.7 ***	15%	**-23%**	4.3 ***
post-graduate degree	12%	9%	**-24%**	1.3	9%	**-26%**	2.7 ***	5%	**-55%**	9.6 ***	6%	**-50%**	7.0 ***
receiving benefits	32%	27%	**-15%**	1.6	41%	**27%**	-5.1 ***	44%	**37%**	-12.1 ***	33%	**3%**	-0.7
married	60%	54%	**-10%**	1.9 *	43%	**-29%**	9.8 ***	42%	**-31%**	17.8 ***	53%	**-12%**	5.5 ***
renting	27%	35%	**29%**	-2.7 ***	56%	**106%**	-17.8 ***	56%	**109%**	-31.4 ***	41%	**54%**	-12.3 ***
in work	71%	65%	**-8%**	2.0 *	61%	**-14%**	5.8 ***	61%	**-14%**	10.7 ***	72%	**2%**	-1.1
NS-SEC 1 destination	15%	12%	**-15%**	1.0	9%	**-39%**	4.5 ***	7%	**-51%**	10.3 ***	12%	**-21%**	3.3 ***
NS-SEC 2 destination	24%	22%	**-9%**	0.8	18%	**-26%**	4.1 ***	14%	**-42%**	11.5 ***	19%	**-21%**	4.5 ***
hourly pay (£)	16.4	15.1	**-8%**	0.3	12.8	**-22%**	0.9	12.4	**-25%**	2.6 **	13.3	**-19%**	1.3

Table 5 Undefined social origins: Comparison of selected observed attributes between those with social origin in routine occupations (SOC9) and those with undefined social origins, separately identifying 6 different sub-groups based on what undefined their social origin.

Attribute	Social origin												
					undefined								
	classified as SOC 9 (n=4,306)	No answer (n=236)			Household composition at age 14 unclear (n=772)			No-one was earning in household when respondent was 14 (n=2,417)			Occupation not identified (n=1,478)		
	mean	mean	% difference	t value	mean	% difference	t value	mean	% difference	t value	mean	% difference	t value
age in years	47.9	43.3	**-10%**	4.6 ***	45.1	**-6%**	4.7 ***	39.8	**-17%**	21.5 ***	44.4	**-7%**	8.0 ***
male	47%	43%	**-9%**	1.2	49%	**3%**	-0.7	44%	**-8%**	3.0 ***	49%	**4%**	-1.5
visible minority	10%	15%	**47%**	-2.3 **	20%	**94%**	-7.6 ***	17%	**63%**	-7.7 ***	20%	**99%**	-10.1 ***
disability	27%	23%	**-15%**	1.4	31%	**13%**	-2.2 **	33%	**23%**	-5.6 ***	20%	**-25%**	5.2 ***
proxy response	32%	38%	**19%**	-1.9 *	36%	**12%**	-2.2 **	39%	**22%**	-5.9 ***	56%	**75%**	-16.8 ***
no qualifications	16%	10%	**-37%**	2.5 **	18%	**7%**	-0.8	19%	**14%**	-2.4 **	16%	**-1%**	0.2
degree holder	11%	20%	**82%**	-4.2 ***	13%	**23%**	-2.1 **	11%	**4%**	-0.5	15%	**43%**	-4.8 ***
post-graduate degree	5%	9%	**68%**	-2.4 **	9%	**62%**	-3.6 ***	5%	**0%**	0.0	6%	**9%**	-0.7
receiving benefits	40%	27%	**-31%**	3.8 ***	41%	**3%**	-0.6	44%	**11%**	-3.5 ***	33%	**-17%**	4.6 ***
married	56%	54%	**-4%**	0.6	43%	**-24%**	6.9 ***	42%	**-26%**	11.3 ***	53%	**-6%**	2.0 **
renting	37%	35%	**-7%**	0.9	56%	**48%**	-9.6 ***	56%	**50%**	-15.1 ***	41%	**11%**	-2.7 ***
in work	64%	65%	**2%**	-0.3	61%	**-4%**	1.5	61%	**-5%**	2.8 ***	72%	**13%**	-5.7 ***
NS-SEC 1 destination	8%	12%	**46%**	-2.1 **	9%	**5%**	-0.4	7%	**-15%**	2.0 *	12%	**37%**	-3.6 ***
NS-SEC 2 destination	18%	22%	**24%**	-1.6	18%	**1%**	0.0	14%	**-21%**	4.1 ***	19%	**8%**	-1.1
hourly pay (£)	15.4	15.1	**-2%**	0.1	12.8	**-17%**	0.3	12.4	**-20%**	0.9	13.3	**-14%**	0.4

3 Social origin pay gap revisited

In this section we examine how omitting those with undefined social origins influences estimates of the social origin pay gap. This follows established practice where an earnings function is estimated based on pooled cross-sectional data for the years 2014-21. We estimate a cross-sectional wage equation, where the dependent variable is the log of hourly wages. This is regressed on the category of social origin, including undefined origins ($\beta_i S_i$) and respondents from SOC 1 (Managers, Directors and Senior Officials) origins are omitted as a reference category. The specification includes a quadratic term for age ($\gamma_1 X + \gamma_2 X^2$) and a range of controls ($\theta_k C_k$), which we extend incrementally with each specification of the model. The analysis includes controls for sex, disability, ethnicity, year of survey, qualifications, degree classification, country of birth, location of workplace, part-time work, firm size, sector of employment and occupational status.

$$\ln(w) = \alpha + \beta_i S_j + \gamma_1 X + \gamma_2 X^2 + \delta_j S_j + \theta_k C_k + \varepsilon$$

Table 6 reveals estimates for these progressively more elaborate wage equations. The first model only controls for demographic features and can be thought of as capturing the raw social origin pay gap. Results are in line with previous analyses of the social origin pay gap, in that that all social origins are disadvantaged vis-á-vis managerial origins. For those with defined social origins, the biggest pay gap is observed for those from SOC 9 origins, 26.3%, followed by those from SOC 8 origins at 25.9%. Examing those with undefined social origins, the most disadvanted group are those those who were not living with family (or household composition could not be identified) at 29.9%, followed by respondents from households where no earner was identified in household at 27.5% and households where the occupation of the main earner could not be identified (26.1%). Overall, these three groups of respondents for which social origin could not be identified, are affected by raw pay gap of a similar or larger magnitude as those from SOC 8 and SOC 9 origins. A non-neglible raw pay gap of 15.1% is observed for those who did not answer the social origin question. This is of a similar magnitude to that observed for those from intermediate occupational origins.

Our second model includes controls for qualifications and therefore captures social origin pay gaps within attainment groups, i.e. the gap that remains despite an individual's educational attainment. It is important to highlight that for the most disadvantaged groups estimated pay gaps are approximately halved vis-à-vis model 1, reinforcing how important educational inequality is as a driver of earnings inequality. For those with defined social origins, the most disadvantaged group are those from SOC 8 origins, facing just over 12% earnings gap on average, closely followed by those from SOC 9 origins. The groups with undefined social origins face disadvantage of similar or larger magnitude. Those whose family at age 14 could not be identified are now associated with a larger pay gap than SOC 8 or SOC 9 at 17.7% and the same applies to those for whom no earner was identified in the parental household at 14.1%. Respondents with non-identified occupational origins are associated with a similar pay gap as SOC 8 and SOC 9 at 12%.

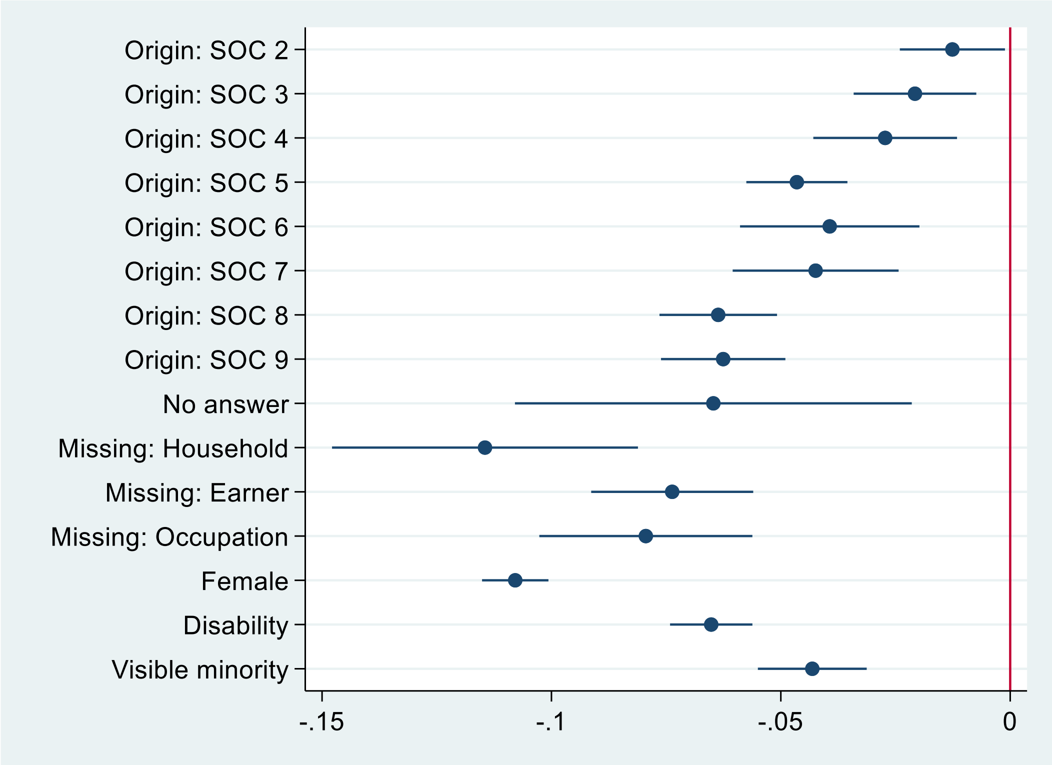
Subsequent models reveal increasingly conditioned forms of the social origin pay gap, as working in particular regions, working part-time, working for smaller firms and in low pay sectors can all affect earnings negatively. Finally, model 9 controls for occupational destination, thereby revealing the unexplained social origin pay gap that remains even when educational attainment and occupational status are accounted for. In this restricted setup just over 6% earnings gap remains for those from SOC 8 and SOC 9 origins. Of those with undefined social origins, the largest pay gap is observed for those whose parental household at age 14 could not be identified at 11.4%, followed by respondents for whom parental occupation could not be identified and at 7.9% and those from parental households where an earner could not be identified at 7.4%. All these point estimates are larger than those for the most disadvantaged groups with undefined social origins. Those who did not answer the social origin question are affected by an unexplained pay similar to those from SOC 8 and 9 social origins.

Table 6 Cross-sectional wage equations 2014-21. Dependent variable: natural logarithm of hourly wages in £. Reference category: Higher Managerial and Professional origins (SOC 1).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
SOC 2 Professionals	0.054***	-0.013*	-0.013**	-0.014**	-0.014**	-0.012*	-0.014**	-0.013**	-0.013**
SOC 3 Associate professional	-0.029***	-0.030***	-0.030***	-0.030***	-0.030***	-0.030***	-0.031***	-0.031***	-0.021***
SOC 4 Administrative and secretarial	-0.052***	-0.041***	-0.040***	-0.039***	-0.040***	-0.040***	-0.044***	-0.041***	-0.027***
SOC 5 Skilled trades	-0.161***	-0.079***	-0.078***	-0.076***	-0.074***	-0.073***	-0.074***	-0.070***	-0.047***
SOC 6 Caring and leisure	-0.190***	-0.089***	-0.087***	-0.089***	-0.086***	-0.081***	-0.079***	-0.073***	-0.039***
SOC 7 Sales and customer service	-0.168***	-0.083***	-0.082***	-0.079***	-0.079***	-0.075***	-0.075***	-0.067***	-0.042***
SOC 8 Process, plant and machine operatives	-0.257***	-0.123***	-0.122***	-0.119***	-0.114***	-0.113***	-0.112***	-0.107***	-0.064***
SOC 9 Routine occupations	-0.263***	-0.122***	-0.121***	-0.118***	-0.115***	-0.115***	-0.115***	-0.109***	-0.063***
No answer	-0.151***	-0.108***	-0.109***	-0.099***	-0.117***	-0.110***	-0.102***	-0.103***	-0.065***
Missing: Household	-0.299***	-0.177***	-0.177***	-0.169***	-0.166***	-0.166***	-0.157***	-0.150***	-0.114***
Missing: Earner	-0.275***	-0.141***	-0.140***	-0.138***	-0.137***	-0.132***	-0.129***	-0.123***	-0.074***
Missing: Occupation	-0.261***	-0.120***	-0.120***	-0.116***	-0.115***	-0.115***	-0.114***	-0.110***	-0.079***
Age	0.087***	0.070***	0.071***	0.071***	0.071***	0.067***	0.063***	0.060***	0.048***
Age^2	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.000***
Female	-0.182***	-0.199***	-0.200***	-0.200***	-0.198***	-0.148***	-0.149***	-0.122***	-0.108***
Disability	-0.121***	-0.092***	-0.092***	-0.094***	-0.093***	-0.085***	-0.086***	-0.082***	-0.065***
Non-white ethnicity	-0.032***	-0.077***	-0.070***	-0.060***	-0.071***	-0.066***	-0.073***	-0.069***	-0.043***
Survey year	√	√	√	√	√	√	√	√	√
Qualifications		√	√	√	√	√	√	√	√
Degree class 1st or 2.1			√	√	√	√	√	√	√
Country of birth				√	√	√	√	√	√
Region of workplace					√	√	√	√	√
Part-time						√	√	√	√
Firm size							√	√	√
Sector of employment								√	√
Occupational status									√
Constant	2.550***	0.807***	0.699***	0.687***	0.693***	0.775***	0.887***	0.887***	0.821***
Observations	79,234	79,234	79,234	79,234	79,234	79,234	79,234	79,234	79,234
R-squared	0.06	0.19	0.32	0.32	0.32	0.33	0.34	0.36	0.38

*** p<0.01, ** p<0.05, * p<0.1

Figure 2 Unexplained social origin pay gaps 2014-2021, in relation to gender, ethnicity and disability pay gaps.



The estimates for the unexplained pay gap are summarised graphically in Figure 2 which shows point estimates and their 95% confidence interval. In order to place the magnitude of these pay gaps in context, the coefficients for the gender, disability and ethnic pay gaps are also plotted. The point estimate for those whose parental household could not be identified is slightly larger than for the gender pay gap; where social origin could not be derived due to unidentified earner or occupation the effect is smaller than for the gender pay gap, but slightly larger than for ethnicity or disability pay gaps. Those who did not respond to the question are associated with an unexplained pay gap of a similar magnitude as those who are disabled. It is also clear from the picture that the confidence intervals on the point estimates for the undefined groups are large. Therefore, it needs to be borne in mind when interpreting these findings that the specifics of any ranking of earnings gaps will be affected by sampling variation. At a glance, it can be observed from Figure 2 that the pay gap for the No Answer, Missing: Earner and Missing: Occupation groups are statistically similar to those observed for SOC 8 and SOC 9 origins and the disabled, but larger than the ethnic pay gap. The Missing: Household group is affected by an earnings gap that is statistically similar to the gender pay gap but larger than those associated with SOC 8 and SOC 9 origins. Moreover, these estimates represent averages for an eight year period, from 2014 through 2021. Whilst there is some variation between years, the pattern of disadvantage observed is not sensitive to choosing a particular year (see Appendix).

3.1 Has omitting undefined social origins led to biased estimates of the social origin pay gap?

The question that remains is whether omission of respondents with undefined social origins has led to biased estimates of social origin pay gaps. Drawing on both the descriptive statistics and the regression results it is clear that average hourly earnings are lower when those with undefined social origins are included and therefore, omitting those respondent leads to biased results in a general sense. Evaluating whether the specific concept of the social origin pay gap has been underestimated in previous work requires a bit more elaboration. The estimates for the pay gaps of those with defined social origins (as produced in Table 6) are not very sensitive to whether those with undefined social origins are included as an additional category or simply omitted (see Table 7).

However, the estimated pay gaps of those with undefined social origins were for three out of four groups greater than those for SOC 8 and 9 origins and therefore, omitting these observations clearly underestimates the potential range of social origin pay gaps. However, a complication arises in that the magnitude of the impacts is inherently sensitive to the definitions of the groups being compared. For instance, if this article were focussing on ethnic pay gaps, a disaggregation of our simple visible minority variable would likely to lead to a wider range of pay gaps as the extent of disadvantage affecting different ethnic groups varies (see e.g. Brynin & Güveli, 2012). The hypothetical question we would ideally like to answer is, if the social origins of the undefined groups were somehow to be discovered and they could be re-categorised into their respective SOC groups, would the estimated social origin pay gaps be materially different than when they were omitted?

By definition, a precise answer to that question cannot be obtained as the social origins cannot be revealed. However, the comparisons illustrated in Table 5 suggest those with undefined social origins share many characteristics with those from defined routine social origins and would therefore disproportionately swell those categories. A simple test would therefore be to recode those with missing social origins into routine social origins. In the absence of better information, we experiment with recoding them all into SOC 9. This has the disadvantage of potentially overestimating the impacts by concentrating all respondents in one category. An alternative approach would be to apply an imputation method to re-classify those of undefined social origin. This is an expansive topic in its own right and well beyond the scope of this article to explore the wide range of potential imputation methods available. Instead, we apply the SOC 9 recode as a preliminary exploration and then compare this with our original specification from Table 6 and a model where unidentified observations are omitted.

Table 7 Unexplained social origin pay gaps 2014-2021. Comparing estimates based on treatment of respondents with undefined social origins.

	(1)	(2)	(3)	(4)
SOC 2 Professionals	-0.015**	-0.015**	-0.015**	-0.015**
	-0.007	-0.007	-0.007	-0.007
SOC 3 Associate professional	-0.019***	-0.019**	-0.019***	-0.019***
	-0.007	-0.007	-0.007	-0.007
SOC 4 Administrative and secretarial	-0.030***	-0.030***	-0.030***	-0.030***
	-0.01	-0.01	-0.01	-0.01
SOC 5 Skilled trades	-0.049***	-0.049***	-0.049***	-0.049***
	-0.006	-0.006	-0.006	-0.006
SOC 6 Caring and leisure	-0.042***	-0.040***	-0.042***	-0.041***
	-0.011	-0.011	-0.011	-0.011
SOC 7 Sales and customer service	-0.042***	-0.041***	-0.042***	-0.042***
	-0.01	-0.01	-0.01	-0.01
SOC 8 Process, plant and machine operatives	-0.061***	-0.061***	-0.061***	-0.061***
	-0.007	-0.007	-0.007	-0.007
SOC 9 Routine occupations	-0.064***	-0.064***	-0.073***	-0.072***
	-0.007	-0.007	-0.007	-0.006
No answer	-0.065***	-0.065***	-0.065***	--
	-0.024	-0.024	-0.024	--
Missing: Household	-0.107***	--	--	--
	-0.022	--	--	--
Missing: Earner	-0.079***	--	--	--
	-0.009	--	--	--
Missing: Occupation	-0.081***	--	--	--
	--	--	--	--
Female	-0.100***	-0.099***	-0.100***	-0.100***
	-0.004	-0.004	-0.004	-0.004
Disability	-0.068***	-0.069***	-0.068***	-0.068***
	-0.005	-0.005	-0.005	-0.005
Non-white ethnicity	-0.046***	-0.047***	-0.046***	-0.046***
	-0.007	-0.007	-0.007	-0.007
Age	√	√	√	√
Survey year	√	√	√	√
Qualifications	√	√	√	√
Degree class 1st or 2.1	√	√	√	√
Country of birth	√	√	√	√
Region of workplace	√	√	√	√
Part-time	√	√	√	√
Firm size	√	√	√	√
Sector of employment	√	√	√	√
Occupational status	√	√	√	√
Observations	79,234	73,729	79,234	79,234
R-squared	0.458	0.457	0.458	0.458

Standard errors in parentheses
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

Table 7 compares alternative approaches for handling undefined social origins in a wage equation estimating social origin pay gaps. The specification of all models is that for unexplained pay gaps but, additional coefficients are redacted to preserve space and standard errors are reported below the estimated coefficients. The first model reproduces Model 9 of Table 6, and includes undefined social origins as separate categories. The second model omits all respondents with undefined social origins. Model 3 reclassifies the three unidentified categories for which social origins could not be derived as SOC 9. Finally, the 4th model reclassifies all those with undefined social origins, including those who refused to answer the questions as SOC 9. The results for Model 2 reveal that omitting those with undefined social origins has only a small impact on coefficients for defined social origins. However, a larger effect is observed in Models 3 and 4 when undefined social origins are re-classified into SOC 9. The magnitude of this effect is substantial, equivalent to just under a percentage point's earnings gap. That is similar to the distance between the pay gaps observed for SOC 2 and SOC 3 origins, however, perhaps smaller than expected given the scale of negative impacts, particularly for the Missing: Household category. As we saw in Tables 2 and 3, whilst the number of observations for the undefined groups is approximately similar to those with SOC 9 origins (4,903 and 4,306 respectively), only about 1 in 7 of those are undefined because household information was missing. Moreover, and crucially, as we observed in Tables 4 and 5, the undefined groups are less likely to be in employment, so will be relatively under-represented in any analyses based on wages.

Table 8 Post-tests of regression coefficients (Yellow = 90% level of significance, salmon = 95% and green = 99%)

Model 1	Model 2	Model 3	Model 4
(1) SOC8 - SOC9 = 0			
F(1, 79159) = 0.20	F(1, 73657) = 0.18	F(1, 79162) = 3.67	F(1, 79163) = 3.56
Prob > F = 0.6556	Prob > F = 0.6712	Prob > F = 0.0554	Prob > F = 0.0590
(1) SOC7 - SOC9 = 0			
F(1, 79159) = 4.78	F(1, 73657) = 4.88	F(1, 79162) = 10.12	F(1, 79163) = 10.01
Prob > F = 0.0288	Prob > F = 0.0272	Prob > F = 0.0015	Prob > F = 0.0016
(1) SOC6 - SOC9 = 0			
F(1, 79159) = 4.57	F(1, 73657) = 5.01	F(1, 79162) = 9.58	F(1, 79163) = 9.48
Prob > F = 0.0325	Prob > F = 0.0252	Prob > F = 0.0020	Prob > F = 0.0021
(1) SOC5 - SOC9 = 0			
F(1, 79159) = 6.38	F(1, 73657) = 5.94	F(1, 79162) = 20.60	F(1, 79163) = 20.51
Prob > F = 0.0116	Prob > F = 0.0148	Prob > F = 0.0000	Prob > F = 0.0000
(1) SOC4 - SOC9 = 0			
F(1, 79159) = 12.46	F(1, 73657) = 12.18	F(1, 79162) = 21.94	F(1, 79163) = 21.81
Prob > F = 0.0004	Prob > F = 0.0005	Prob > F = 0.0000	Prob > F = 0.0000
(1) SOC3 - SOC9 = 0			
F(1, 79159) = 34.68	F(1, 73657) = 34.09	F(1, 79162) = 59.66	F(1, 79163) = 59.72
Prob > F = 0.0000	Prob > F = 0.0000	Prob > F = 0.0000	Prob > F = 0.0000
(1) SOC2 - SOC9 = 0			
F(1, 79159) = 50.55	F(1, 73657) = 49.22	F(1, 79162) = 88.88	F(1, 79163) = 89.31
Prob > F = 0.0000	Prob > F = 0.0000	Prob > F = 0.0000	Prob > F = 0.0000
(1) - SOC9 + female = 0			
F(1, 79159) = 18.34	F(1, 73657) = 17.92	F(1, 79162) = 12.52	F(1, 79163) = 12.86
Prob > F = 0.0000	Prob > F = 0.0000	Prob > F = 0.0004	Prob > F = 0.0003
(1) - SOC9 + vismin = 0			
F(1, 79159) = 3.45	F(1, 73657) = 2.71	F(1, 79162) = 8.12	F(1, 79163) = 8.03
Prob > F = 0.0634	Prob > F = 0.0996	Prob > F = 0.0044	Prob > F = 0.0046
(1) - SOC9 + disability = 0			
F(1, 79159) = 0.15	F(1, 73657) = 0.40	F(1, 79162) = 0.35	F(1, 79163) = 0.32
Prob > F = 0.6943	Prob > F = 0.5282	Prob > F = 0.5547	Prob > F = 0.5740

A more formal comparison of the models is provided in Table 8, which tests whether the distance between the SOC 9 pay gap and other pay gaps in each model is statistically significant. A useful feature for comparison across the models is that the estimates for the SOC 8 pay gap coefficients are stable across the models. In Models 1 and 2 the difference between the SOC 8 and SOC 9 coefficients is not statistically significant. However, in Models 3, when undefined socials origins (and, in Model 4, those that do not answer the social origin questions) are recoded in SOC 9 the difference between SOC 8 and SOC 9 becomes statistically significant at 90% level of confidence. Some of the other differences are reinforced, with the statistical significance of difference between SOC9 and SOC 7, SOC 6 and SOC 5, respectively, moving from 95% to 99% level of significance. The same is observed for the difference between the SOC 9 and ethnic pay gaps. In this particular setup, omitting those with undefined social origins has underestimated pay gaps by approximately nine percentage points (see Table 7). To put this in context, this underestimate is equivalent to a fifth (20%) of the ethnic pay gap. On the other hand, the exclusion of this part of the data makes no significant difference to the disability pay gap.

4 Discussion: Implications for practice and theory

This article sets out our investigation in four stages. First, an examination of how the social origin variable is derived - this established that respondents' social origins are undefined due to specific attributes of the previous generation's household. Second, a comparison of observable features of those with defined and undefined social origins - this revealed that missingness of social origins is non-random. Third, further analysis of the characteristics of these groups found that undefined social origins are associated with economic and social disadvantage across a range of indicators, including education, occupational destination and earnings. Fourth, estimation of class pay gaps are shown to be significantly and substantially underestimated when data for those with undefined social origins are omitted. Overall, the analyses demonstrate that those who do not fit the occupational classification are among those that we should be most interested in knowing about in order to understand the impact of social origin on labour market outcomes. Moreover, these respondents represent approximately five million individuals of working age, living in the UK. This group is not typical, but, on average, younger, more likely to be of colour and more likely to be disabled than the population at large. They share similar material outcomes as those from routine origins but are demographically different — and as we have demonstrated, do not fit well into an occupation based social class schema.

Respondents with undefined social origins present labour market researchers with an empirical and a conceptual problem. How should empirical issues be addressed? First, it is imperative that respondents with undefined social origins should not be dropped – symbolically this an egregious act, it forfeits information about a large sub-population and is likely to lead to biased parameter estimates. Our preferred solution is simply to include these groups as separate categories. This is the simplest solution. It lends these respondents a voice and the results are straightforward to interpret. In large scale social surveys, it may be possible to re-classify observations by drawing on other observed features. This is an area for future research, whether through statistical imputation approaches or through artificial intelligence classification algorithms. These would complicate analyses through additional steps and require assumptions to which results would inevitably be sensitive. In order for researchers to pursue such approaches, it needs to be clear that obtaining simulated but comprehensive social origin data within a specific occupational framework provides sufficient analytical benefits to justify the additional complications. Whilst the focus of this article has been on the UK LFS, similar levels of missingness of parental occupation have been found in other data sets e.g. the 1958 National Child Development Study (Betthaeuser and Bourne, 2016). Moreover, our preliminary analysis of data from the Higher Education Statistics Agency for students in higher education in 2018-19 shows that NS-SEC codes of parents was missing for 18.4% of the sample. We observed a similar level of missingness for parental occupation in Understanding Society - the UK Household Longitudinal Study. Further research could examine the item non-response of social origin in other UK datasets.

A less straightforward issue to conclude is how this affects occupation-based social classifications conceptually? The findings presented in this article chime with well-established criticism of occupation-based classifications as being overly rigid or too static to capture the dynamics of a fluid social reality. However, it needs to be borne in mind that analysing the role of social origin in the labour market benefits from social origin indicators being available as part of key labour market statistics. The relative simplicity of occupation-based classifications makes them suitable for large scale application in surveys and therefore more easily deployed as part of the national statistics programme administered by the ONS. For the purposes of empirical labour market research that aims to generalise about a population, any proxy for social class must pass the test of being straightforward to gather data for at scale. Therefore, on balance, labour market research is far richer using these frameworks, whilst acknowledging their limitations, than doing without simple social origin proxies in surveys. Moreover, as demonstrated earlier, a thorough understanding of how this data is collected and under what circumstances respondents “drop out” of the classification can be used to meaningfully interpret findings for those with undefined social origins.

Not being identified in an occupation-based classification is associated with specific forms of disadvantage. This group displays characteristics of a more diverse society that is perhaps not aligned with historical notions of the industrial working class. This is, in itself, a much longer discussion, but what does it mean for the specific concept of the social origin pay gap? There is no doubt that the availability of social origin data in national statistics has been an overwhelmingly positive step – bridging research on earnings and social mobility respectively. This has highlighted and created awareness of the insidious nature of class-based disadvantage. However, we argue that if anything the social origin pay gap as estimated in the wave of research that has emerged since social origin was first included in the LFS in 2014 is a conservative estimate of this material disadvantage. We have demonstrated that omitting those that don't fit the classification leads to an underestimate. Moreover, built into the unexplained social origin pay gap are selection issues that are likely to lead to a further channel by which underestimation takes place. For instance, as shown in our descriptive analyses, labour market attachment varies, so there's likely to be a survivor bias in those from the most disadvantaged backgrounds that make it into work.

5 Conclusion

In conclusion, this article re-examines how researchers have applied the social origin variable (SMSOC101) in the UK Quarterly Labour Force Survey to estimate wage equations and argues that for social origin, item non-response is non-random. We show this by disentangling the way the social origin variable is derived and disaggregating the non-response groups as far as possible. We highlight the characteristics of those to whom the social origin question does not apply and show that overall, this group reports less favourable outcomes in relation to education, occupational attainment and show a higher pay-gap in comparison to those who do report social origin.

We estimate wage equations and show that a subgroup of respondents who did not report social origin have a lower wage coefficient than those from SOC 9 origins i.e. those whose parents were employed in routine occupations. This suggests that the social origin wage gap is larger than previously estimated. Furthermore, the wage coefficients for this group are statistically significant even after considering a range of demographics, educational attainment and labour market observables. Therefore, we argue that previous empirical studies which have omitted respondents with undefined social origins have underestimated the social origin pay gap and the number of individuals affected.

Our results contribute to a theoretical criticism of the internal logic in the SOC schema, in that those who do not fit into this occupational framework are omitted from studies on social origin pay gaps and possibly previous studies on social mobility. This highlights how respondents from non-traditional backgrounds are not captured in the way social origin is operationalised. The results from this article indicate that the excluded group (comprising 11% of the working age population) is non-random and has several characteristics which indicate disadvantage in education, housing tenure and employment.

However, there are relatively simple ways in which the framework could be operationalised and data from social surveys could be used to address this. As shown here, those with undefined social origins could be acknowledged as a separate group and included in labour market research. However, there are differences within this group and further analysis would still be required to better understand the underlying drivers of disadvantage for the various sub-groups. Longer term, surveys should be enhanced to probe more deeply into non-traditional backgrounds, e.g. what is behind not living with parents? Is it care experience or some other experience associated with disadvantage that social policy could be mobilised to address?

Our findings show, in line with previous research, that occupational backgrounds (employment relations) have an important intergenerational impact on labour market outcomes. What our analyses adds, is that there is a further effect of parental association with the labour market or not clearly belonging to a household which profoundly affect the life outcomes of a substantial share of the working age population. The latter is particularly important given that there is already compelling evidence of labour market, educational and socio-economic effects of having been part of the care system in childhood (Bywaters et al, 2016; Gypen et al, 2017; Harker et al, 2004; Jackson and Sachdev, 2001; Viner and Taylor, 2005).

Overall, our results reinforce the urgency to better understand and address socio-economic inequalities in the UK labour market.

6 Appendix

Table 9 Variable description and descriptive features of pooled sample (2014-21)

Variable name	Variable description	N	Mean	Std. Dev.
loghourpay	Natural logatithm of hourly pay	79,234	2.54	0.592
SOC1	Binary variable = 1 if social origin is SOC 1	393,254	11%	--
SOC2	Binary variable = 1 if social origin is SOC 2	393,254	15%	--
SOC3	Binary variable = 1 if social origin is SOC 3	393,254	8%	--
SOC4	Binary variable = 1 if social origin is SOC 4	393,254	5%	--
SOC5	Binary variable = 1 if social origin is SOC 5	393,254	21%	--
SOC6	Binary variable = 1 if social origin is SOC 6	393,254	3%	--
SOC7	Binary variable = 1 if social origin is SOC 7	393,254	4%	--
SOC8	Binary variable = 1 if social origin is SOC 8	393,254	12%	--
SOC9	Binary variable = 1 if social origin is SOC 9	393,254	10%	--
RegUnidentifiedNoA~r	Binary variable = 1 if no answer to social origin question	393,254	1%	--
RegUnidentifiedHou~d	Binary variable = 1 if social origin undefined because parent household when respondent was age 14 could not be identified	393,254	2%	--
RegUnidentifiedEar~r	Binary variable = 1 if social origin undefined because no earner identified in parent household when respondent was age 14.	393,254	5%	--
RegUnidentifiedOcc~n	Binary variable = 1 if social origin undefined because occupation could not be identified for main earner in parent household when respondent was age 14.	393,254	0.0	--
AGE	Age in years	393,254	45.1	15.2
age2	Age in years squared	393,254	2,265	1,336
female	Binary variable = 1 if respondent is female	393,254	53%	--
vismin	Binary variable = 1 if respondent belongs to a visible minority.	393,254	10%	--
disability	Binary variable = 1 if respondent is classified as disabled according to the Equality Act	393,254	21%	--
qualmissing	Highest qualification attained: no response to question	393,254	0%	--
qualnoresponse	Highest qualification attained not identified	393,254	0%	--
qualaca3plus	Highest qualification attained: academi postgraduate	393,254	10%	--
qualaca3	Highest qualification attained: academic graduate	393,254	19%	--
qualvoc3	Highest qualification attained: vocational graduate	393,254	0%	--
qualaca3sub	Highest qualification attained: academic sub-degree	393,254	5%	--
qualvoc3sub	Highest qualification attained: vocational sub-degree	393,254	5%	--
qualaca2plus	Highest qualification attained: academic post-secondary	393,254	8%	--
qualvoc2plus	Highest qualification attained: vocational post-secondary	393,254	14%	--
qualaca2sub	Highest qualification attained: academic lower secondary	393,254	16%	--
qualvoc2sub	Highest qualification attained: vocational lower secondary	393,254	5%	--
qualother	Highest qualification attained: other	393,254	9%	--
qualnoqual	Highest qualification attained: none	393,254	9%	--
degclass1	Degree class: 1st	393,254	3%	--
degclass21	Degree class: 2.1	393,254	10%	--
birthdum1	Country of birth: not-identified	393,254	0%	--
birthdum2	Country of birth: no answer	393,254	0%	--
birthdumIndia	Country of birth: India	393,254	1%	--
birthdumROI	Country of birth: Republic of Ireland	393,254	0%	--
birthdumPakistan	Country of birth: Pakistan	393,254	1%	--
birthdumPoland	Country of birth: Poland	393,254	1%	--
birthdumEngland	Country of birth: England	393,254	69%	--
birthdumNI	Country of birth: Northern Ireland	393,254	6%	--
birthdumScotland	Country of birth: Scotland	393,254	8%	--
birthdumWales	Country of birth: Wales	393,254	4%	--
birthdumUK	Country of birth: UK (nor further specified)	393,254	0%	--
birthdumOther	Country of birth: Other	393,254	11%	--
locDoesNotApply	Location of workplace: not identified	337,091	31%	--
locNoAnswer	Location of workplace: no answer	393,254	5%	--
locNorthEast	Location of workplace: North East England	393,254	86%	--
locNorthWest	Location of workplace: North West England	393,254	87%	--
locYorksAndHumber	Location of workplace: Yorkshire and Humber	393,254	6%	--
locEastMidlands	Location of workplace: East Midlands	393,254	5%	--
locWestMidlands	Location of workplace: West Midlands	393,254	6%	--

locEastofEngland	Location of workplace: East of England	393,254	6%	--
locLondon	Location of workplace: London	393,254	87%	--
locSouthEast	Location of workplace: South East of England	393,254	8%	--
locSouthWest	Location of workplace: South West of England	393,254	7%	--
locWales	Location of workplace: Wales	393,254	4%	--
locScotland	Location of workplace: Scotland	393,254	5%	--
locNorthernIreland	Location of workplace: Northern Ireland	393,254	4%	--
locWorkplaceoutsid~K	Location of workplace: Outside UK	393,254	1%	--
parttime	Respondent works part time (fewer than 35 hours a week)	393,254	55%	--
firmsizedum1	Firm size: not identified	393,254	40%	--
firmsizedum2	Firm size: no answer	393,254	1%	--
firmsizedum3	Firm size: 1-10 employees	393,254	12%	--
firmsizedum4	Firm size: 11-19 employees	393,254	5%	--
firmsizedum5	Firm size: 20-24 employees	393,254	3%	--
firmsizedum6	Firm size: Don't know but fewer than 25	393,254	1%	--
firmsizedum7	Firm size: 25-49 employees	393,254	8%	--
firmsizedum8	Firm size: 50-249 employees	393,254	14%	--
firmsizedum9	Firm size: 250-499 employees	393,254	4%	--
firmsizedum10	Firm size: Dont know but between 50 and 499	393,254	2%	--
firmsizedum11	Firm size: more than 500 employees	393,254	11%	--
sectorunknown	Sector of work: unidentified	393,254	32%	--
sectormissing	Sector of work: no answer	393,254	0%	--
sectorA	Sector of work: agriculture, forestry and fishing	393,254	1%	--
sectorBDE	Sector of work: energy and water	393,254	1%	--
sectorC	Sector of work: manufacturing	393,254	7%	--
sectorF	Sector of work: construction	393,254	5%	--
sectorGI	Sector of work: distribution, hotels and restaurants	393,254	12%	--
sectorHJ	Sector of work: transport and communications	393,254	6%	--
sectorKLMN	Sector of work: banking and finance	393,254	12%	--
sectorOPQ	Sector of work: public administration, education and health	393,254	22%	--
sectorRSTU	Sector of work: unidentified	393,254	4%	--
jobdum1	Occupational classification of work: not identified	393,254	0%	--
jobdum2	Occupational classification of work: higher managerial and professional	393,254	14%	--
jobdum3	Occupational classification of work: lower managerial and professional	393,254	23%	--
jobdum4	Occupational classification of work: intermediate	393,254	12%	--
jobdum5	Occupational classification of work: small employers and own account workers	393,254	8%	--
jobdum6	Occupational classification of work: lower supervisory and technical occupations	393,254	6%	--
jobdum7	Occupational classification of work: semi-routine occupations	393,254	11%	--
jobdum8	Occupational classification of work: routine occupations	393,254	8%	--
jobdum9	Occupational classification of work: never worked or long-term unemployed	393,254	19%	--

Table 10 Wage equations by year 2014-21 - regression coefficients for unexplained social origin pay gaps

	2014	2015	2016	2017	2018	2019	2020	2021
SOC 2 Professionals	0.003	-0.015	-0.007	-0.01	-0.019	-0.005	-0.036*	-0.035*
SOC 3 Associate professional	-0.045**	-0.012	-0.009	-0.015	-0.019	-0.03	-0.006	-0.029
SOC 4 Administrative and secretarial	-0.006	0.008	-0.038	-0.021	-0.033	-0.070**	-0.048	-0.042*
SOC 5 Skilled trades	-0.034**	-0.039**	-0.049***	-0.030*	-0.043**	-0.054***	-0.057***	-0.085***
SOC 6 Caring and leisure	-0.049*	0.007	-0.045*	-0.023	-0.048	-0.048*	-0.070**	-0.060**
SOC 7 Sales and customer service	-0.047*	-0.035	-0.005	-0.067**	-0.046*	-0.027	-0.04	-0.072***
SOC 8 Process, plant and machine operatives	-0.049***	-0.062***	-0.048***	-0.056***	-0.072***	-0.059***	-0.059***	-0.087***
SOC 9 Routine occupations	-0.058***	-0.043**	-0.078***	-0.044**	-0.070***	-0.075***	-0.063***	-0.083***
No answer	-0.096**	0	-0.138	-0.029	0.1	-0.049	-0.102	-0.09
Not living with family, etc.	-0.267***	-0.095*	-0.066*	-0.066*	-0.053	-0.182***	-0.096	-0.092
No earner identified in household	-0.075***	-0.073**	-0.042*	-0.061***	-0.039	-0.091***	-0.114***	-0.135***
Occupation not identified	-0.021	-0.084***	-0.092***	-0.061**	-0.084***	-0.116***	-0.053	-0.143***
Age	0.048***	0.054***	0.048***	0.052***	0.048***	0.046***	0.043***	0.050***
Age^2	-0.000***	-0.001***	-0.000***	-0.001***	-0.000***	-0.000***	-0.000***	-0.001***
Female	-0.116***	-0.097***	-0.121***	-0.104***	-0.116***	-0.090***	-0.067***	-0.090***
Disability	-0.069***	-0.069***	-0.069***	-0.055***	-0.067***	-0.061***	-0.082***	-0.067***
Non-white ethnicity	-0.055***	-0.075***	-0.054***	-0.043***	-0.047***	-0.043**	-0.046**	-0.0170
Qualifications	√	√	√	√	√	√	√	√
Degree class 1st or 2.1	√	√	√	√	√	√	√	√
Country of birth	√	√	√	√	√	√	√	√
Region of workplace	√	√	√	√	√	√	√	√
Part-time	√	√	√	√	√	√	√	√
Firm size	√	√	√	√	√	√	√	√
Sector of employment	√	√	√	√	√	√	√	√
Occupational status								
Constant	1.492***	1.294***	1.456***	1.416***	1.576***	1.550***	1.654***	1.644***
Observations	10741	10579	9975	10355	9491	9627	8759	9707
R-squared	0.499	0.484	0.486	0.435	0.439	0.438	0.406	0.421

*** p<0.01, ** p<0.05, * p<0.1

7 References

Arulampalam, W., Booth, A. L., & Bryan, M. L. (2007). Is there a glass ceiling over Europe? Exploring the gender pay gap across the wage distribution. ILR Review, 60(2), 163-186.

Bernardi, F., & Gil-Hernández, C. J. (2021). The social-origins gap in labour market outcomes: Compensatory and boosting advantages using a micro-class approach. European Sociological Review, 37(1), 32-48.

Berthoud, R. (2008). Disability employment penalties in Britain. Work, employment and society, 22(1), 129-148.

Bridges, S., & Mann, S. (2019). Sexual Orientation, Legal Partnerships and Wages in Britain. Work, Employment and Society, 33(6), 1020-1038.

Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. Journal of economic literature, 55(3), 789-865.

Blinder, A. S. (1973). Wage discrimination: reduced form and structural estimates. Journal of Human resources, 436-455.

Brynin, M., & Güveli, A. (2012). Understanding the ethnic pay gap in Britain. Work, Employment & Society, 26(4), 574-587.

Bukodi, E., & Goldthorpe, J. (2011). Social class returns to higher education: chances of access to the professional and managerial salariat for men in three British birth cohorts. Longitudinal and Life Course Studies, 2(2), 185-201.

Bywaters, P., Bunting, L., Davidson, G., Hanratty, J., Mason, W., McCartan, C., & Steils, N. (2016). The relationship between poverty, child abuse and neglect: An evidence review. York: Joseph Rowntree Foundation.

Chevalier, A. (2007). Education, occupation and career expectations: determinants of the gender pay gap for UK graduates. Oxford Bulletin of Economics and Statistics, 69(6), 819-842.

Connelly, R., Gayle, V., & Lambert, P. S. (2016). A Review of occupation-based social classifications for social survey research. Methodological Innovations, Vol. 9, pp. 1-14.

Conlon, G. & Patrignani, P. (2013). The Impact of University Degrees on the Lifecycle of Earnings: Some Further Analysis. BIS Research Paper No. 45. London: Department for Business Innovation and Skills.

Crawford, C., & Erve, L. V. D. (2015). Does higher education level the playing field? Socio-economic differences in graduate earnings. Education Sciences, 5(4), 380-412.

Crawford, C and Vignoles, A. (2014) Heterogeneity in graduate earnings by socio-economic background, IFS working paper 14(30) Retrieved from the World Wide Web: http://www.ifs.org.uk/publications/7419

Culliney, M. (2017). Escaping the rural pay penalty: location, migration and the labour market. Work, employment and society, 31(3), 429-446.

Duta, A., & Iannelli, C. (2018). Social Class Inequalities in Graduates' Labour Market Outcomes: The Role of Spatial Job Opportunities. Social Sciences, 7(10), 201.

Dregan A, Gulliford MC (2012) Foster care, residential care and public care placement patterns are associated with adult life trajectories: Population-based cohort study, Social Psychiatry and Psychiatric Epidemiology, 47 (9): 1517-1526.

Erikson, R., Goldthorpe, J. H., & Goldthorpe, J. H. (1992). The constant flux: A study of class mobility in industrial societies. Oxford University Press.

Fomby, P., & Cherlin, A. J. (2007). Family instability and child well-being. American Sociological Review, 72(2): 181-204.

Fortin, N., Lemieux, T., & Firpo, S. (2011). Decomposition methods in economics. In Handbook of labor economics (Vol. 4, pp. 1-102). Elsevier.

Fortin, N. M., Bell, B., & Böhm, M. (2017). Top earnings inequality and the gender pay gap: Canada, Sweden, and the United Kingdom. Labour Economics, 47, 107-123.

Friedman, S., Laurison, D., & Macmillan, L. (2017). Social mobility, the class pay gap and intergenerational worklessness: new insights from the Labour Force Survey.

Friedman, S., & Laurison, D. (2017). Mind the gap: financial London and the regional class pay gap. The British journal of sociology, 68(3), 474-511.

Friedman, S., & Laurison, D. (2020). The class ceiling: Why it pays to be privileged. Oxford: Oxford University Press.

Goldthorpe, J.H. (with Llewellyn, C. & Payne, C.) (1980/1987). Social Mobility and Class Structure in Modern Britain, 1st edition. Oxford: Clarendon Press.

Goldthorpe, J. H. (1997). The ‘Goldthorpe’class schema: some observations on conceptual and operational issues in relation to the ESRC review of government social classifications. Constructing classes: towards a new social classification for the UK. Swindon: ESRC/ONS.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. (Eds.). (2002). Survey nonresponse (Vol. 23). New York: Wiley.

Groves, R. M., & Couper, M. P. (2012). Nonresponse in household interview surveys. John Wiley & Sons.

Gypen, L., Vanderfaeillie, J., De Maeyer, S., Belenger, L., & Van Holen, F. (2017). Outcomes of children who grew up in foster care: Systematic-review. Children and Youth Services Review, 76, 74-83.

Hällsten, M. (2013). The class‐origin wage gap: heterogeneity in education and variations across market segments. The British journal of sociology, 64(4), 662-690.

Harker, RM., Dobel-Ober, D., Akhurst, S., Berridge, D. Sinclair, R. (2004) Who takes care of education eighteen months on? A follow-up study of looked after children’ s perceptions of support for educational progress. Child and Family Social Work. 9 (3) 273-284

Hermannsson, K., Lecca, P., Lisenkova, K., McGregor, P. & Swales, K. (2014). The Regional Economic Impact of More Graduates in the Labour Market: a “Micro-to-Macro” Analysis for Scotland. Environment and Planning A, 46 (2), 471-487.

Hersbein,B., & Bartik, T. (2016) College grads earn less if they grew up poor, Research Note: W.E. Upjohn Institute for Employment Research. Retrieved from the World Wide Web: http://www.upjohn.org/sites/default/files/pdf/hershbein-bartik-degrees-of-poverty.pdf

Jackson, S., & Sachdev, D. (2001). Better education, better futures: Research, practice and the views of young people in public care. Ilford: Barnardo's.

Krueger, A. B., & Lindahl, M. (2001). Education for Growth: Why and For Whom? Journal of Economic Literature, 39, 1101-1136.

Lambert, P. S., & Bihagen, E. (2014). Using occupation-based social classifications. Work, employment and society, 28(3), 481-494.

Laurison, D., & Friedman, S. (2016). The class pay gap in higher professional and managerial occupations. American Sociological Review, 81(4), 668-695.

Longhi, S., & Brynin, M. (2017). The ethnicity pay gap. Equality and Human Rights Commission. Retrieved from the World Wide Web: https://www.equalityhumanrights.com/sites/default/files/research-report-108-the-ethnicity-pay-gap.pdf

Manning, A., & Robinson, H. (2004). Something in the way she moves: a fresh look at an old gap. Oxford economic papers, 56(2), 169-188.

LSE Growth Commission (2013). Investing for Prosperity: Skills, Infrastructure and Innovation. Report of the LSE Growth Commission. Retrieved from the World Wide Web: http://www.lse.ac.uk/researchAndExpertise/units/growthCommission/documents/pdf/LSEGC-Report.pdf

Mastekaasa, A. (2011). Social origins and labour market success—stability and change over Norwegian birth cohorts 1950–1969. European Sociological Review, 27(1), 1-15.

Mincer, J. (1974). Schooling, Experience, and Earnings. New York: National Bureau of Economic Research.

Montenegro, C. E., & Patrinos, H. A. (2014). Comparable estimates of returns to schooling around the world. Montenegro, Claudio E. and Patrinos, Harry A., Comparable Estimates of Returns to Schooling Around the World (September 1, 2014). World Bank Policy Research Working Paper No. 7020.

OECD (2012). Promoting Growth in All Regions: Lessons From Across the OECD. Paris: OECD. ISBN: 9789264174627'

ONS - Office for National Statistics (2009). Labour Force Survey User Guide – Volume 5: LFS Classifications. Retrieved from the World Wide Web: https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/labourforcesurveyuserguidance#labour-force-survey-lfs-user-guides

Psacharopoulos, G., & Patrinos, H. A. (2004). Human capital and rates of return. In Johnes and Johns (eds.) International handbook on the economics of education, pp. 1-57. Cheltenham: Edward Elgar.

Rafferty, A. (2012). Ethnic penalties in graduate level over-education, unemployment and wages: evidence from Britain. Work, Employment & Society, 26(6), 987-1006.

Rose, D., & Harrison, E. (2007). The European socio-economic classification: a new social class schema for comparative European research. European Societies, 9(3), 459-490.

Rose, D., & Harrison, E. (Eds.). (2014). Social class in Europe: An introduction to the European socio-economic classification (Vol. 10). Routledge.

Rose, D. & Pevalin, D. (2001). The national statistics socio-economic classification: unifying official and sociological approaches to the conceptualisation and measurement of social class in the united kingdom. ISER Working Papers. Number 2001-4

Rose, D., & Pevalin, D. J. (2003). A researcher's guide to the national statistics socio-economic classification. London: Sage.

Särndal, C. E., & Lundström, S. (2005). Estimation in surveys with nonresponse. John Wiley & Sons.

Shavit, Y. (2007). Stratification in higher education: A comparative study. Stanford University Press.

Viner, R. M., & Taylor, B. (2005). Adult health and social outcomes of children who have been in public care: population-based study. Pediatrics, 115(4), 894-899.

Walker, I., & Zhu, Y. (2008), The College Wage Premium and the Expansion of Higher Education in the UK, The Scandinavian Journal of Economics 110(4), 695-709.

Walker, I. and Zhu, Y. (2011) 'Differences by degree: evidence of the net financial rates of return to undergraduate study for England and Wales', Economics of Education Review, 30(6): 1177–186.

Walker, I., & Zhu, Y. (2013). The impact of university degrees on the lifecycle of earnings: some further analysis. BIS Research Paper No. 112. London: Department for Business Innovation and Skills.