Office for National Statistics

Article

The impact of miscoding of occupational data in Office for National Statistics social surveys, UK

Following a coding error identified in Standard Occupational Classification 2020 across social surveys, an analysis was conducted to identify affected codes.

Contact:
Martina Helme
socialsurveys@ons.gov.uk
+44 1633 580181

Release date:
26 September 2022

Next release:
To be announced

Table of contents

1 . Overview of the miscoding of occupational data

- The Office for National Statistics announced on 18 July 2022 that an issue with the collection of some occupational data was identified, affecting a number of our surveys.
- The issue was caused by the implementation of the updated Standard Occupational Classification (SOC) from SOC10 to SOC20 and is limited to occupation variables and associated derived variables.
- Analysis at SOC Major Group (one-digit) level found that the coding error only had a marginal effect on results; this aligns with statements in our prior announcements that headline labour market measures are affected very little by this issue.
- Our research shows that around half of four-digit SOC codes are likely affected, although until the full recoding is complete we cannot quantify the full impact; a list of how each of the SOC codes is likely to be affected has been produced with this article.
- Over the coming few months, we will apply recoding of occupation variables in the Labour Force Survey (LFS) and Annual Population Survey (APS) collected from January 2021 to September 2022, which we aim to publish alongside associated Labour Market publications by the end of spring 2023.

2 . Background to miscoding of occupational data

The Office for National Statistics (ONS) announced on 18 July 2022 that we identified an issue with the collection of some occupational data affecting a number of our surveys, for more information see our Occupational data in ONS surveys statement. This issue was caused by the implementation of the updated Standard Occupational Classification (SOC) from SOC10 to SOC20, whereby responses to surveys were being miscoded to the wrong occupations. This error is limited to occupation variables and associated derived variables, such as Socio-Economic Classification (NS-SEC). This does not affect other variables or key headline labour market measures.

As stated in our Update on occupational data in ONS surveys on 15 August 2022, following a detailed review of data collected by the Labour Force Survey (LFS) this paper presents:

- an assessment of which occupations have been affected by the error
- the process in determining which ones were affected
- an outline of our approach to resolving the error

3 . Approach to analysing the impact

In order to evaluate which occupations were affected by the issue identified with survey collection, we conducted two strands of investigative analysis.

Analysis of the occupational coding index

We analysed the index used in the process of coding occupational data to identify which codes were affected by the coding issue and to what extent. Each entry in the Standard Occupational Classification (SOC) 2020 coding index contains information necessary to apply a code to an occupation, which includes three key pieces of information: the job title, the industrial qualifying terms, and additional information. As the issue focused on the handling of this additional information, we categorised the coding index into groups based upon the additional information in the coding frame. This allows interviewers to identify the most appropriate code based on the full occupational description provided by survey respondents.

Taking a specific example, two index entries have an identical index value of "accountant, financial, coal mine", with the additional information for one entry being "qualified". These two entries would allocate a respondent to two different occupations in two different parts of the SOC20 categorisation, making the term "qualified" an important distinguishing factor. Code groups with similar features were then filtered to remove any pairings where all the entries in the group contained the same four-digit SOC code, and therefore any errors in this case would have no impact. What remained are the occupation codes where the handling of additional information may have affected the coding of people into the correct occupations. Therefore SOC codes that appeared in at least one of the resulting groups were identified as "potentially being affected" by this issue, while SOC codes that did not appear in any group were classed as "unaffected".

To assess the extent that each four-digit SOC20 code was affected by this issue, all codes were matched against their equivalent SOC10 codes. Comparing the SOC20 code collected by the Labour Force Survey (LFS) January to March (JM) 2021 and the SOC10 code from the October to December (OD) 2020 data, a list was produced of how many times each SOC code appeared in each dataset. Cases that appear in both OD20 and JM21 - meaning a person was interviewed in both quarters - were then checked against each other to estimate the number of people who were either likely to have been categorised correctly, or those who were at risk of miscoding.

The outcome of this analysis on the coding index was then compared with the time series of occupational data on the Annual Population Survey (APS). This was to identify where the coding issue significantly contributed to the variation observed since the introduction of SOC20 codes in 2021. Based upon this, any occupation where there is a large number of cases (more than 50%) collected by the survey where the handling of qualifying information may potentially have caused a miscoding, have been classed as "high impact". Occupations where some but not most cases (between 5% and 50%) were affected have been classed as "moderate impact", and the remaining occupations (less than 5% affected) have been classed as "low impact". This informs us of the extent that an occupation grouping may have been impacted by miscoding, but until the full recoding exercise is completed we will not know the real extent of the impact.

Net change between Major Groups

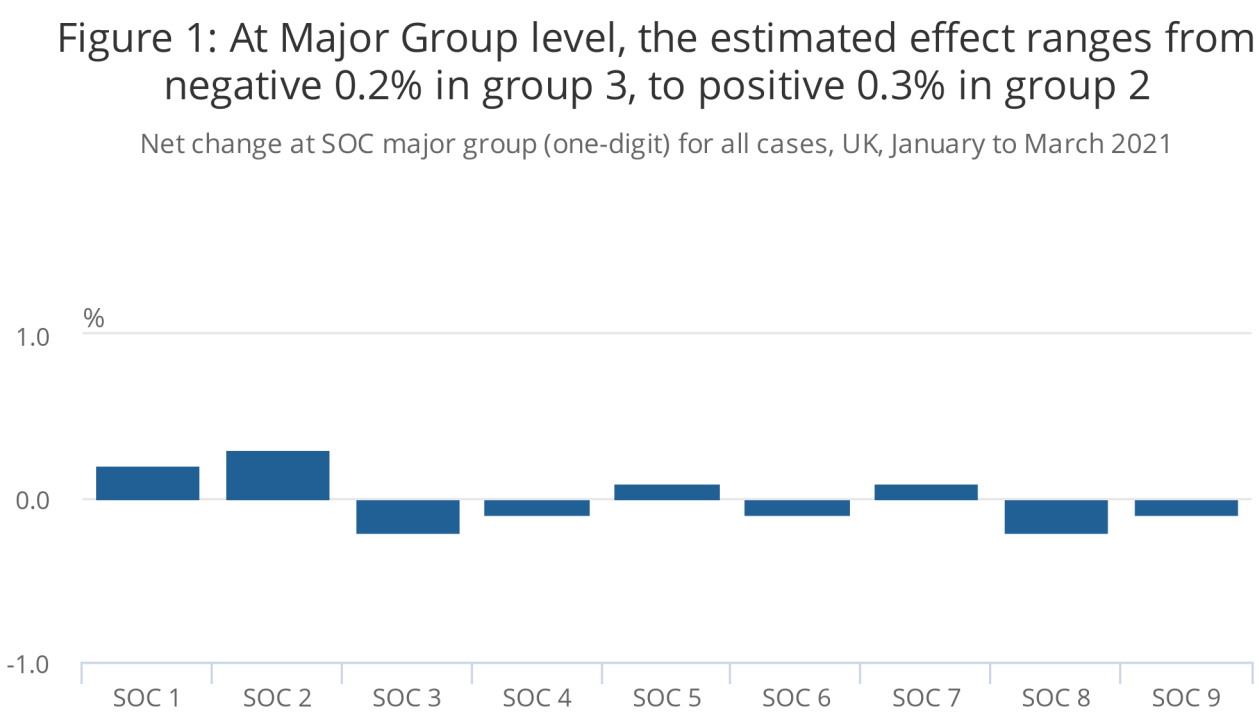
We also analysed the impact at SOC Major Group (one-digit) level using the LFS data for OD20 and JM21. To achieve this, first the data are filtered to include only people who were interviewed in both quarters, who also were in employment and had a valid occupation code for their job. Analysing people who were in the same job for the previous 12 months, we could then extract people who had stayed within the same occupation Major Group (for example, they were in the same job role in both quarters), and those who had legitimately changed between Major Groups as a result of structural changes between SOC10 and SOC20. What remained were the people who had stayed with the same employer but changed between Major Groups likely as a result of miscoding. With this, we could estimate the net effect of the error at the one-digit occupation level.

4 . Main findings from analysing the impact

Analysis at Standard Occupational Classification (SOC) Major Group (one-digit) level was conducted. It found that for respondents in both the October to December 2020 (OD20) and the January to March 2021 (JM21) quarter of the Labour Force Survey (LFS), who had remained in the same job for at least 12 months, the effect of correcting the error to those occupations identified as potentially highly impacted would cause only marginal net change at the Major Group level. As shown in Figure 1, the effect is estimated to range from negative 0.2% in group 3 (associate professional occupations) to positive 0.3% in group 2 (professional occupations). This aligns with statements in our prior announcements that the main headline labour market measures are not significantly affected by this coding issue.

Figure 1: At Major Group level, the estimated effect ranges from negative 0.2% in group 3, to positive 0.3% in group 2

Net change at SOC major group (one-digit) for all cases, UK, January to March 2021



Source: Office for National Statistics – Labour Force Survey

Analysis of the demographic composition of workers coded into the potentially highly affected occupation codes indicates there should be no overall bias caused by the error. The percentage of workers in those highly affected groups, compared with all workers as a whole follows almost exactly the same pattern when analysing by age, sex, ethnicity and location. By way of example, we present the percentages of workers in the highly impacted occupations in Table 1 broken down by region or country, where we see it follows almost the exact same pattern as for all workers.

Table 1: Comparison at country and regional level, UK, 2021

Country or region	Percentage of working population in area in 2021 (%)	Percentage of working population in highly affected SOC codes in 2021 (%)	Difference between percentages
North East	3.7	3.7	0.1
North West	10.5	10.7	0.2
Yorkshire and The Humber	8.0	8.0	-0.1
East Midlands	7.1	7.1	0.0
West Midlands	8.6	8.6	0.0
East of England	9.6	9.3	-0.2
London	14.8	14.9	0.1
South East	14.1	14.2	0.1
South West	8.4	8.3	-0.1
Wales	4.5	4.5	0.0
Scotland	8.1	8.1	0.0
Northern Ireland	2.6	2.6	-0.1

Source: Office for National Statistics – Annual Population Survey

There is, however, more significant impact with far more granular break downs. Our analysis indicates that SOC codes at the four-digit level were not all affected to the same extent, with the primary determinant being the significance of the qualifying "additional information" from the coding index used in assigning a correct SOC code.

As shown in Table 2, out of the 412 individual SOC20 (four-digit) unit groups, we estimate that 113 (27.4%) codes saw a low impact, 90 (21.8%) with moderate impact, and 209 (50.7%) with potentially high impact. Although these codes are categorised as high impact, this only means that there was a high potential for these occupations to be miscoded. It is likely that most respondents have been coded correctly, only where the qualifying "additional information" was relevant to a specific respondent would it have been miscoded. Even for those that were miscoded, they should have been categorised into a similar type of work. For example, being a different type of sales assistant but remaining within the same Major Group, or being in a certified profession rather than an uncertified worker. Therefore, the headline statistics remain largely unaffected.

Table 2: Impact of SOC coding issue at Unit Group (four-digit) level, UK

Level of impact	Number of SOC20 codes	Percentage (%) of all SOC20 codes
Low impact (<5% affected)	113	27.4
Moderate impact (5% to 50% affected)	90	21.8
High impact (>50% affected)	209	50.7

Source: Office for National Statistics – Labour Force Survey

It is only once the data have been fully recoded that the precise detail of the impact at this lower level of break down can be evaluated. For a table with the list of all 412 SOC codes and their respective level of impact, please see our accompanying dataset.

5 . Future developments

Having now identified the occupations potentially highly affected by the error, we will now work to apply recoding of occupations to the Labour Force Survey (LFS) and the Annual Population Survey (APS) collected since January 2021. Data collected before this time remain unaffected by this issue. This process will involve a combination of automated and clerical recoding, where the occupation codes subjected to recoding in the LFS and APS include those for main job, last job, second job, apprenticeships, redundancies, jobs held one year ago and those related to social mobility. All derived variables based on these occupation codes, such as Socio-Economic Classification (NS-SEC), will subsequently be updated.

These revisions will be carried out over the next six months and, allowing time to update associated Labour Market publications, we aim to publish revised LFS and APS data in spring 2023. Until the revised LFS and APS data have been published, we advise caution when using data based upon the occupation codes identified as being affected. For any analysis conducted based on the affected occupational data, we advise users revisit these once the revised data have been published in 2023.

This coding issue has only a negligible impact on published data from other social surveys. Results from the Annual Survey of Hours and Earnings (ASHE) use LFS occupation data at Major Group (one-digit) level as part of its weighting process. Once the LFS data have been revised we will then review the need to revise ASHE data, but any such impact should be minimal based on the findings from this analysis.

6 . The impact of miscoding of occupational data in Office for National Statistics social surveys, UK data

Impact at four-digit Standard Occupational Classification level
Dataset | Released 26 September 2022
This table contains a full list of all 412 Standard Occupational Classification 2020 (SOC20) codes and their respective estimated level of impact from this coding issue.

7 . Related links

Occupational data in ONS surveys
Statement | Released 18 July 2022
The Office for National Statistics (ONS) has identified an issue with the collection of some occupational data in a number of our surveys. The ONS urges users to be cautious in the interpretation of these detailed data as the issue is being further investigated and resolved.

Update on occupational data in ONS surveys
Statement | Released 15 August 2022
The Office for National Statistics has continued to investigate an issue with the collection of occupational data in our surveys. An article will be published in September containing an overview of the level of impact of this issue.

SOC 2020 Volume 2: the coding index and coding rules and conventions
Methodological information | Released 2020
The Standard Occupational Classification (SOC) is a common classification of occupational information for the UK.

8 . Cite this article

Office for National Statistics (ONS), released 26 September 2022, ONS website, Article, The impact of miscoding of occupational data in Office for National Statistics social surveys, UK