

Article

Revision of miscoded occupational data in the ONS Labour Force Survey, UK: January 2021 to September 2022

Following a coding error identified in the collection of Standard Occupational Classification 2020 (SOC20) data for the Office for National Statistics (ONS) Labour Force Survey, the occupational data is being revised.

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

Release date:
11 July 2023

Next release:
To be announced

Table of contents

1. [Main points](#)
2. [Background to miscoding of occupational data](#)
3. [Approach to revising the occupational data on the Labour Force Survey \(LFS\) and Annual Population Survey \(APS\)](#)
4. [Anticipated changes in the revised occupational data](#)
5. [Future developments](#)
6. [Related links](#)
7. [Cite this article](#)

1 . Main points

- The Office for National Statistics (ONS) 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.
- The ONS published an article on 26 September 2022 informing users of the impact the identified coding error had on occupational data, highlighting which occupational classifications were likely to be most affected.
- Following the identification and analysis of the coding error, a recoding exercise was undertaken to correct the error in the occupational data on the ONS Labour Force Survey (LFS) and Annual Population Survey (APS) using a combination of methods; all relevant datasets have subsequently been revised and scheduled for re-release in July and August 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](#). This issue was caused by the implementation of the updated Standard Occupational Classification (SOC) from SOC10 to [SOC20](#), where 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 important headline labour market measures.

Following a detailed review of data collected by the Labour Force Survey (LFS), we presented an assessment on 26 September 2022 of which occupations had been affected by the error, the process in determining which ones were affected, and an outline of our approach to resolving the error. For full details, please see our [Impact of miscoding of occupational data in Office for National Statistics social surveys, UK article](#).

3 . Approach to revising the occupational data on the Labour Force Survey (LFS) and Annual Population Survey (APS)

Along with the article released on 26 September 2022, we published an [associated dataset](#) including a full list of all 412 Standard Occupational Classification 2020 (SOC2020) codes and their respective estimated level of impact from the coding issue we identified. The 209 four-digit occupational classifications marked with “high impact” (50.7%) have been subjected to recoding.

The recoding of the occupational data from the Labour Force Survey (LFS) and Annual Population Survey (APS) involved two distinct stages. Firstly, an automated coding algorithm was applied, followed by clerical recoding applied to selected cases only. The selection of codes subjected to clerical recoding was based on whether a case met the required confidence threshold assigned by the automated coding algorithm.

The process of assigning a SOC code differs slightly depending on the method that is used. The automated coding algorithm reads the occupational title, occupational description, industry information and the highest qualification, and subsequently assigns a SOC code without human intervention. The automated coding tool then returns a score (that is, a confidence level) and cases failing to meet a certain threshold were referred to clerical recoding.

While the machine algorithm primarily uses the occupation description (resulting from the question “what do you mainly do in your job?”) to determine the appropriate code, clerical coders and interviewers mainly use the occupation title (resulting from the question “what is your job title?”).

The main difference between the coding process conducted by interviewers and the one used by clerical coders is accessibility of the coding frame. Interviewers operate on an interface where they enter search terms based on the occupational information provided by the respondent. The interviewer can then select a code from the most relevant options displayed. The clerical coders have the entire coding frame accessible to them at any one time.

4 . Anticipated changes in the revised occupational data

The use of different coding methods to recode Standard Occupational Classification (SOC) codes affected by the coding error mean that some discontinuities may be introduced to the occupational data.

The analysis presented in this update is based on comparisons of the averages of weighted counts for main job taken from the Labour Force Survey (LFS) data from 2019 to 2020 (SOC2010 mapped to SOC2020), the original data for main jobs for LFS from January to March 2021, and the revised data for main jobs for LFS from January to March 2021.

Table 1: Impact of revision for main occupation for January to March 2021 (JM21)

Impact of revision	Number of SOC codes
Revised JM21 estimate is closer to 2019 to 2020 average than original JM21 estimates	176
Revised JM21 estimate is the same or marginally further out from the 2019 to 2020 average than original JM21 estimate	65
Revised JM21 estimate is further out from the 2019 to 2020 average than original JM21 estimates	171

Source: Labour Force Survey from the Office for National Statistics

For 176 (43%) out of the 412 four-digit SOC codes, the revised total is closer to the 2019 to 2020 average than the original total. Of those 412 SOC codes, 108 (26%) were identified as highly impacted by the coding error.

Table 2: Examples of improved revised SOC codes

SOC unit (4-digit) level code	2019 to 2020 average (1000s)	Jan to Mar 2021 original (1000s)	Jan to Mar 2021 revised (1000s)	Percentage change between 2019 to 2020 average and Jan to Mar 2021 original	Percentage change between 2019 to 2020 average and Jan to Mar 2021 revised
3231 Higher level teaching assistant	70	238	38	238%	-46%
7113 Telephone sales person	32	68	30	109%	-6%
9131 Industrial cleaning process occupations	26	197	19	643%	-30%

Source: Labour Force Survey from the Office for National Statistics

For 65 (16%) of the four-digit SOC codes, the difference from the 2019 to 2020 average in January to March 2021 increased up to five percentage points. For 17 (4%) of these, the difference from the 2019 to 2020 average is the same between the original and the revised totals.

Table 3: Examples of unchanged SOC codes

SOC unit (4-digit) level code	2019 to 2020 average (1000s)	Jan to Mar 2021 original (1000s)	Jan to Mar 2021 revised (1000s)	Percentage change between 2019 to 2020 average and Jan to Mar 2021 original	Percentage change between 2019 to 2020 average and Jan to Mar 2021 revised
3534 Financial account managers	169	145	146	-14%	-14%
5231 Vehicle technicians, mechanics, and electricians	187	135	135	-28%	-28%
8222 Forklift truck drivers	78	59	58	-24%	-25%

Source: Labour Force Survey from the Office for National Statistics

In 171 instances (42%), the revised total saw an increase of more than five percentage points in the level of deviation between the January to March 2021 total and the 2019 to 2020 average, when comparing the original with the revised total. This is true for 76 (18%) of codes that were marked as highly impacted by the coding error, and 95 (23%) of codes that saw a low to moderate impact from the coding error. However, the extent to which the revised total is further from the historical average than the original total varies.

Table 4: Examples of new SOC code changes

SOC unit (4-digit) level code	2019 to 2020 average (1000s)	Jan to Mar 2021 original (1000s)	Jan to Mar 2021 revised (1000s)	Percentage change between 2019 to 2020 average and Jan to Mar 2021 original	Percentage change between 2019 to 2020 average and Jan to Mar 2021 revised
2211 General medical practitioners	169	183	197	8%	16%
3554 Marketing associate professionals	217	208	169	-4%	-22%
8214 Delivery drivers and couriers	214	249	283	16%	32%

Source: Labour Force Survey from the Office for National Statistics

Analysis has also been conducted to assess the impact on Annual Population Survey (APS) data. The impact on APS data is in line with the impact on LFS data, with minimal impact at the one-digit level, but greater volatility at the four-digit level. The four-digit codes most affected in the LFS were also the most affected in the APS.

Because of the combination of coding methodologies applied and the differences with which these operate, it is not unexpected to see changes in the revised totals that deviate from the historical time series, which was entirely coded by interviewers.

Users should therefore expect to see slight discontinuities for some four-digit level codes for time series of occupational data, including the period from December 2020 to January 2021 (representing the switch from SOC2010 to SOC2020 and switch from interviewer-led to partially recoded data) and September 2022 to October 2022 (representing the switch from partially recoded to interviewer-led coded data).

5 . Future developments

The issue in the coding frame resulting in this coding error has been resolved and new occupational data collected from October 2022 is therefore not affected by the coding error anymore.

With all Standard Occupational Classification (SOC) codes that were identified as highly impacted by the coding error having been recoded, the focus then turned to re-processing the Labour Force Survey (LFS) and Annual Population Survey (APS) microdata. This involved revising relevant derived variables and re-running the income weight, which uses SOC codes in population calibrations. All relevant aggregate tables for the LFS and APS have subsequently been updated.

Outputs from other Office for National Statistics (ONS) social surveys will not be revised as these are largely based on major SOC groups, for which the impact is only marginal. The ONS will be assessing the difference that the new LFS estimates make in the weighting for the Annual Survey of Hours and Earnings (ASHE) and will report on findings.

The following LFS and APS micro-datasets will be re-released over the next two months, including periods from January 2021 to December 2022. This is because, while the period from October to December 2022 was not affected by the coding error, the datasets for the relevant period still need to be revised, as data brought forward from the previous period for imputation purposes will need to be updated

The following micro-datasets will be re-released on 11 July 2023:

- LFS person
- LFS two quarter longitudinal
- APS person

The following micro-dataset will be re-released on 19 July 2023:

- APS 3 year pooled

The following micro-datasets will be re-released on 15 August 2023:

- LFS household
- LFS five quarter longitudinal
- APS household
- APS 2 year longitudinal

Revised SOC variables

The full list of revised SOC variables are as follows:

- NSECM20 -NSECMJ20
- PIWT22 (LFS)
- PIWTA22 (APS)
- SC2010A
- SC2010L
- SC2010LMJ
- SC2010LMN
- SC2010M
- SC2010MMJ -SC2010MMN
- SC2010R
- SC2010S
- SC2010SMJ
- SC2010SMN
- SC20LMJ
- SC20LMN
- SC20MMJ
- SC20MMN
- SC20SMJ
- SC20SMN
- SMSOC201
- SMSOC203
- SMSOC204
- SOC20A
- SOC20L
- SOC20M
- SOC20O
- SOC20R
- SOC20S

6 . Related links

[Occupational data in ONS surveys](#)

Announcement | Released 18 July 2022

Details of the issues identified by the Office for National Statistics (ONS) with the collection of some occupational data in a number of our surveys.

[Update on occupational data in ONS surveys](#)

Announcement | Released 15 August 2022

The Office for National Statistics issued a statement explaining we are continuing to investigate an issue with the collection of occupational data in our surveys, and that an article would be published in September 2022 containing an overview of the level of impact of this issue.

[The impact of miscoding of occupational data in Office for National Statistics social surveys, UK](#)

Article | Released 26 September 2022

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

[Standard Occupational Classification \(SOC\)](#)

Methodological information

The Standard Occupational Classification (SOC) is a common classification of occupational information for the UK.

7 . Cite this article

Office for National Statistics (ONS), released 11 July 2023, ONS website, article, [Revision of miscoded occupational data in the ONS Labour Force Survey, UK: January 2021 to September 2022](#)