SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

Survey Practice Forum (Online) • 1 October 2024

Developments in occupational coding in online self-completion surveys

Lisa Calderwood (Centre for Longitudinal Studies, UCL)
Sebastian Kocar (Institute for Social Science Research, University of Queensland)

University of Essex | University of Southampton | UKRI Economic and Social Research Council

NCRM National Centre for Research Methods | Office for National Statistics | UCL | Warwick The University of Warwick | Manchester 1824 The University of Manchester | City University of London Est 1894 | National Centre for Social Research | LSE The London School of Economics and Political Science | Ulster University | Unil UNIL | Université de Lausanne | Ipsos | Kantar Public

Outline

- Context and general overview.
- Evidence review.
- Occupational coding in UK surveys.
- An application: Look-up approach in the Next Steps study.
- Conclusions and recommendations.

Research Strand 5 – Complex measurements in online social surveys

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

i) **Industry and occupation**
ii) Cognitive assessments
iii) Consents for data linkage / re-contact etc
iv) Retrospective data collection

Core team

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Work package lead: Lisa Calderwood (Centre for Longitudinal Studies (CLS), UCL)

Researchers: Sebastian Kocar (University of Queensland), Cristian Domarchi (University of Southampton), Marc Asensio (CLS)

Core team: Matt Brown (CLS), Curtis Jessop (NatCen), Jo D'Ardenne (NatCen), Olga Maslovskaya (Southampton), Orlaith Fraser (ONS), Andrew Phelps (ONS), Laura Wilson (ONS), Joe Sakshaug (Warwick)

Context and general overview

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- **Occupation** is a key measure in many social surveys
- Important marked of socio-economic status
- Significant impact on income, health, lifestyle and other domains of life
- Occupation details typically coded to standard code-frames – e.g. SOC in the case of the UK

Context and general overview

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- Challenging and complex to measure:
  - Occupations can be as diverse as the people participating in surveys, and different individuals might describe the same job in different ways (Simson et al., 2023).
  - The range of questions required for accurate occupation coding might be extensive (Belloni et al., 2016) and can vary greatly between occupations.
  - Respondents (and/or interviewers) may provide insufficient or invalid information
  - Occupation coding is typically conducted after data collection and is based solely on the provided answers without the ability to request more information from respondents (Simson et al., 2023).

Context and general overview

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- The “traditional” approach involves interviewers asking open questions to collect job title and a description of duties which are then manually coded by specialist office-based coders (Lyberg & Dean, 1992)
- Interviewers can help ensure the respondent provides the information required to ensure successful coding (Conrad et al., 2016)
- In **self-completion surveys**, the absence of interviewers can have a negative impact on the quality of the collected data for coding (Conrad et al., 2016).

Evidence review

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- Generally speaking, we distinguish between the following approaches to occupation coding based on when the data are coded, who/what codes the data, and what software/algorithm is used:
  - **Coding during the survey** (also known as self-coding, e.g. look-up) and **post-survey coding** (answers to open-ended questions are recorded and office coded after data collection)
  - **Manual coding** (arguably the "gold standard") and **automated coding** (using programming solutions)
  - **Manual coding:** software assisted office coding (e.g. CASCOT) and coding using only classification (like SOC)
  - **Automated coding:** rule based, machine learning models, language processing models (AI)

Evidence review

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- Different occupation coding approaches lead to different coding rates and agreement rates (as indicators of occupation data quality); compared to collection of open-descriptions of jobs and manual office coding:
  - **Self-coding rates are expected to be lower** (even more in self-completion modes using a look-up), which requires manual coding for those with no occupation codes selected
  - **Agreement between self-coding and office coding** is expected to be lower than between two office coders
  - **Finding balance between coding rates and agreement rates** is challenging for automated coding - targeting high accuracy rates can have a negative impact on coding/production rates and vice versa
- Less notable differences in the content of open-ended questions asked to collect occupation data (especially in the UK), but based on the approach

Occupation coding in UK surveys

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- **Review of practices in the UK:** official/government statistics (ONS), cross-sectional surveys (ESS), longitudinal surveys (cohort and household panel studies)
- **Observable trends**: self-completion, self-coding (look-up), automated coding (e.g. machine learning), methodological research to develop occupation coding approaches
- **Examples of notable changes over time** (relevant for occupation coding): surveys moving online (web panels, COVID), censuses conducted online-first, development of coding tools to assist coding, development of tools/algorithms for self-coding, increasing costs of survey data collection

Occupation coding in UK surveys

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Occupation data are collected, also using self-completion, in:

- **Censuses**: 2021 Census (England and Wales), 2021 Census (Northern Ireland) and 2022 Census (Scotland).
- **Other official statistics surveys:** Labour Force Survey → Transformed LFS
- **Cross-sectional surveys:** European Social Survey (ESS), NatCen opinion panel surveys (e.g. for His Majesty's Revenue & Customs, Skills and Employment survey), surveys conducted by Verian (e.g. Community Life Survey, Participation Survey)
- **Panel/cohort/longitudinal studies:** Understanding Society, National Child Development Study, British Cohort Study, Millenium Cohort Study, Children of the 2020s, Early Life Cohort Feasibility Study

Occupation coding in UK surveys

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Main identified occupation coding practices:

- **Open-ended job descriptions and manual office coding:** 2021 Census of England and Wales (assisted by the ONS coding tool), European Social Survey (ISCO classification), Understanding Society (office coded assisted by CASCOT, Verian), longstanding birth cohort studies (UCL)
- **Open-ended job descriptions and automated office coding:** Labour Market Survey data (SIC), ONS tested large language models, also suitable for SOC
- **Look-up:** Labour Force Survey (look up assists interviewers), Next Steps Age 25 and Age 32 Sweeps (also self-completion, double-coding approach in Age 32 Sweep), pilot stage of the BCS70 study Age 46 Survey (coding by nurse interviewers)
- **Closed-ended:** NatCen panel (closed 26-category SOC question, sometimes following open-ended question, e.g. in Skills and employment survey), Verian surveys

Occupation coding in UK surveys

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Key issues identified:

- **Labour Force Survey:** cases with low confidence levels needed to be manually coded
- **BCS70 Age 46 pilot:** coding during the survey by nurse interviewers using the look-up was less accurate than office coding
- **European Social Survey:** the proportion of cases that could not be coded due to missing data averaged 15% across all countries using self-completion, compared to 2% in the previous face-to-face round
- **Census 2021 development research by Verian:** respondents not being aware of the ability to overwrite the suggestions using a look-up, long lists of the occupation options proposed by the function, duplicate entries, primacy effect
- As well as…

Occupation coding in UK surveys

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Methodological research

- **ONS:**
  - **Labour Force Survey/Opinions and Lifestyle Survey 2010**: Comparison of the automated coding frame with expert manual office coding found discrepancies for 60% of jobs at the 4-digit level.
  - **Opinions and Lifestyle Survey 2010**: Respondents self-coded their answers to occupation questions, and their results were compared with interviewer codes, with an agreement rate of 68%.
  - **Census 2011**: ONS explored the proportion of households in which respondents were not prepared to provide proxy data for other household members' occupations, but the results have not been published.

Occupation coding in UK surveys

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Methodological research

- **Next Steps Age 25 Survey**
- Mixed mode approach (web>tel>F2F)
- Look-up approach to occupation coding trialled – office coding where no code selected.
- 82% assigned a code (90% in web and telephone – c. 20% lower in F2F
- Coding occupational descriptions performed comparably across modes.
- Web respondents spent more time using look-up and provided longer answers (but without a positive effect on coding rates).
- Higher coding rates by profile:
  - Respondents: white study members, attended university, living with a partner
  - Interviewers: female, younger, more experienced

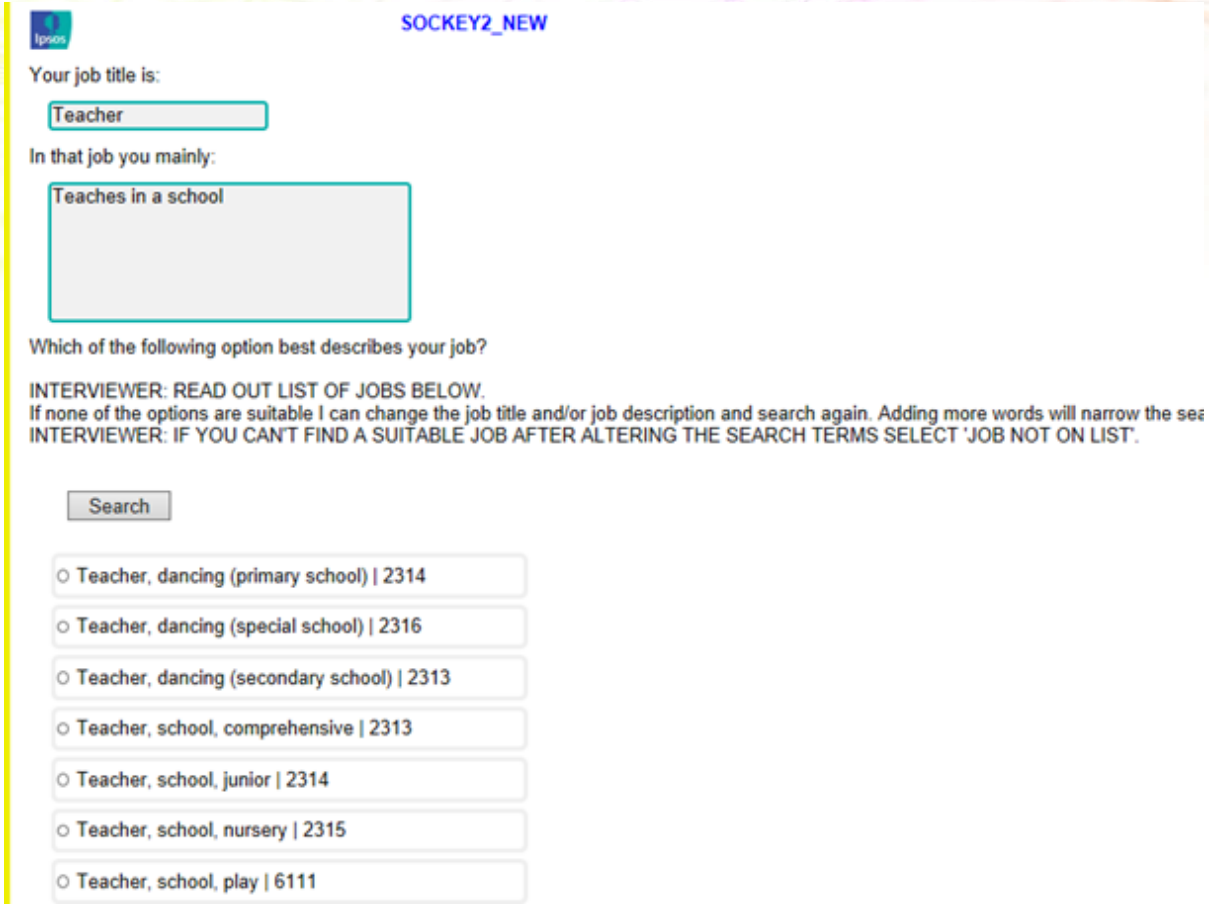
Trialling a look-up approach in the Next Steps Age 32 Survey

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- The following method was used to collect occupation data with the look-up approach:
  1) Respondents were asked: "*What is your job title?*" (max. 50 characters) and "*Please tell us keywords which describe what you do in your job*" (max. 200 characters).
  2) A look-up trigram search function used these answers to generate a list of possible occupations with SOC codes at the 4-digit level. The list was presented to the respondent or read out loud by the interviewer.
  3) Job titles and keywords could be edited to generate a new list of occupations, if required.
  4) The respondent (or the interviewer after discussion with the respondent) selected the code that best described the job. "*Job not on the list*" could also be chosen as the final answer.
  5) Open text description also collected

The look-up

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION



Evaluation

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- Additional information was collected to assess data quality:

1) Open-ended job description also collected for double office coding from all respondents

2) Respondent ratings of suitability of look-up code: "*How well do you think the option you selected actually describes the job that you do?"* (answer options: "very well", "fairly well", "not very well", "not at all well").

3) The consecutive number of the occupation (occupation code) from the list of answers generated by the look-up

An application: Next Steps Age 32

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- Research questions:
- *What proportion of respondents could successfully select a code – and did this vary by mode/device? How did respondents rate the accuracy of their selected code?*
- *What is the consistency between look-up codes and office codes – and did this vary by mode/device? Is the level of consistency different to the level of consistency between two office coders?*
- *What factors affect successful look-up coding rates?*
- *What factors affect consistency between look-up and office codes?*

Results

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

‘Successful’ coding rates:

	Look-up coding rate	Office coding rate
Web (n=4,626)	81.5%	99.4%
F2F (including Video) (n=583)	88.2%	99.5%
Phone (n=114)	93.0%	100.0%
Total (n=5,323)	**82.5%**	**99.4%**

Results

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Self-reported accuracy of selected look-up code:

	Very well	Fairly well	Not very/not at all well
Web	44.8%	49.3%	5.9%
F2F (including Video)	62.3%	36.0%	1.7%
Phone	58.5%	36.8%	4.7%
Total	**47.4%**	**47.3%**	**5.3%**

Results

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- **Consistency of coding:**

Mode	Agreement the look-up & office coder				Agreement 1st office coder & 2nd office coder			
	1-digit level	2-digit level	3-digit level	4-digit level	1-digit level	2-digit level	3-digit level	4-digit level
Web (n=3,768)	78.3%	74.1%	69.9%	62.1%	94.3%	93.1%	92.2%	89.2%
F2F (including video) (n=514)	77.6%	73.5%	68.7%	62.8%	95.2%	94.5%	93.5%	91.6%
Phone (n=106)	81.1%	79.3%	75.5%	69.8%	93.0%	92.1%	91.2%	88.6%
Total (n=4,388)	**78.3%**	**74.2%**	**69.9%**	**62.3%**	**94.4%**	**93.3%**	**92.3%**	**89.4%**

Results

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- **Factors affecting successful look-up coding:**

- Mode/Device – no impact (after including other predictors/controls)
- Length of inputs – Longer job titles (↑), Longer keywords (↓), no entries (↓↓)
- Time – Time spent entering job title (↓), Time spent entering key words (n.s.)
- Editing look-up entries (↓↓)
- Whether job changed since previous sweep (n.s.)
- Occupation group – Managers/Directors (↓) Admin/Secretarial (↓), All other groups (n.s.)
- Socio-demographics: resident of other UK States than England (↑), married individuals (↑), South Asian ethnicity (↓)

Results

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- **Factors affecting consistency between look-up and office-coding:**
- Mode/Device – no impact
- Length of inputs – Longer job titles (n.s.), Longer keywords (↓), Open-text descriptions (n.s.)
- Time – Time spent entering job title (n.s.), Time spent entering key words (n.s.)
- Editing look-up entries (n.s.)
- Whether job changed since previous sweep (n.s.)
- Occupation group – Managers/Directors (↓), Associate Professionals (↓), Admin/Secretarial (↓), All other groups (n.s.)
- Look-up answer selected – further down list of options (↓↓)
- Respondent rated suitability of look-up code – very well (↑↑), not very/not at all well (↓↓)
- Socio-demographics: Graduate degree (↓)

Discussion and conclusions

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- Majority of respondents in all modes successfully selected a code
  - F2F participants more likely to select a code than web participants
- Majority of respondents reported that look-up code described their job well
- Evidence that look-up approach has promise as a cost-effective approach and could substantially reduce the need for manual coding.
- Further improvement of the look-up function are possible based on our findings (e.g. by using paradata in real time, additional prompts and instructions)
- However, level of consistency between look-up code and manual office codes are low – and significantly lower than consistency between two office coders – potentially raising questions about quality of look-up based codes.

Upcoming research

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- **Census Non-Response Link Study:**
  - Study conducted by the ONS.
  - Matches households (and the individuals living within them) who participated in the 2021 Census for England and Wales, and Labour Force Survey (LFS) conducted around the same time.
  - Plan to use this matching to analyse occupation data collected through different modes:
    - Census 2021 collected 4-digit SOC occupation data via online self-completion.
    - LFS data were collected via computer-assisted telephone interviewing (CATI), with interviewers assigning the 4-digit SOC codes.

Upcoming research

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- **Transformed Labour Force Survey (TLFS):**
  - TLFS is primarily administered via computer-assisted web interviewing, with some telephone interviews.
  - Automated office coding is carried out with occupation data that are collected through free text questions
  - Plan to evaluate the accuracy of this automated coding by comparing to manually coded data
  - Evaluate impact of changes to questionnaire design and respondent guidance

Upcoming research

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- **Next Steps Measurement Lab**
  - Two-wave experiment - participants randomly allocated to 3 different modes(web, video and F2F) at each wave (2 weeks apart)
  - Occupation details collected in both waves – look-up and open-text
  - Further analysis of the impact of mode on collection of occupation data using two methods

Upcoming research

SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

- **NatCen Opinion Panel Surveys**
  - Use experimental designs to examine impact of additional participant prompts and additional closed question asking participants to self-classify themselves to 1-digit SOC groups on successful coding rates

SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

Survey Practice Forum (Online) • 1 October 2024

Developments in occupational coding in online self-completion surveys

Lisa Calderwood (Centre for Longitudinal Studies, UCL)
Sebastian Kocar (Social Science Research, University of Queensland)

University of Essex | University of Southampton | UKRI Economic and Social Research Council

NCRM National Centre for Research Methods | Office for National Statistics | UCL | Warwick The University of Warwick | Manchester 1824 The University of Manchester | City University of London Est 1894 | National Centre for Social Research | LSE The London School of Economics and Political Science | Ulster University | Unil UNIL | Université de Lausanne | Ipsos | Kantar Public