SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

Survey Practice Guide 11: Occupation coding in self-completion surveys

Helena Koerber[1], Lisa Calderwood[2], Matt Brown[2], Sebastian Kocar[3]

[1] London School of Economics and Political Science, [2] University College London, [3] University of Queensland

November 2026

www.surveyfutures.net

Survey Futures is an Economic and Social Research Council (ESRC)-funded initiative (grant ES/X014150/1) aimed at bringing about a step change in survey research to ensure that high quality social survey research can continue in the UK. The initiative brings together social survey researchers, methodologists, commissioners and other stakeholders from across academia, government, private and not-for-profit sectors. Activities include an extensive programme of research, a training and capacity-building (TCB) stream, and dissemination and promotion of good practice. The research programme aims to assess the quality implications of the most important design choices relevant to future UK surveys, with a focus on inclusivity and representativeness, while the TCB stream aims to provide understanding of capacity and skills needs in the survey sector (both interviewers and research professionals), to identify promising ways to improve both, and to take steps towards making those improvements. Survey Futures is directed by Professor Peter Lynn, University of Essex, and is a collaboration of twelve organisations, benefitting from additional support from the Office for National Statistics and the ESRC National Centre for Research Methods. Further information can be found at www.surveyfutures.net.

The suggested form of citation for this guide is:

Koerber, H., Calderwood, L., Brown, M. & Kocar, S. (*2026*) 'Survey Practice Guide: Occupation coding in self-completion surveys, *Survey Futures Survey Practice Guide* no. 11. Colchester, UK: University of Essex. Available at https://surveyfutures.net/practice-guides/.

Contents

1. Introduction 4

2. Methods 4

3. Summary of Methods for Occupation Data Collection and Coding 6

4. Key Recommendations 9

5. References 10

1. Introduction

Occupation is an important measure collected in many social surveys. Occupations are typically classified to standard code frames such as the Standard Occupational Classification System (SOC) in the case of the UK. Traditionally, occupation data has been collected through open-ended questions on job title and duties, which are later coded by office coders.

Measuring occupation is inherently challenging with respondents potentially describing the same job in different ways or providing insufficient information for accurate coding. These difficulties are often magnified in online surveys, where participants complete questions without interviewer support.

As social research increasingly moves towards online and mixed-mode designs, identifying approaches that enable accurate, consistent and efficient occupation coding in web surveys has become essential.

This survey practice guide draws on a comprehensive evidence review (Kocar et al., 2026) which examined how occupation data are collected and coded in UK self-completion surveys. The guide presents a structured comparison of commonly used approaches to collect occupation data in online surveys, outlining the strengths, limitations and practical considerations of each. The guide also offers general recommendations to support method selection, based on data quality needs and operational constraints. Readers who are seeking more detail, including the full evidence base, examples from specific surveys, and a more in-depth discussion of coding approaches, are encouraged to consult the evidence review alongside the practice guide.

2. Methods

2.1 Post-interview coding of open text

Manual or clerical coding of open text

In the traditional approach, respondents provide free-text descriptions of their job title, main duties and other relevant details. Trained office coders then review these responses and assign codes using standard occupational classification systems such as SOC (in the case of the UK) after the survey. Coding may be supported by software tools. A widely used example is CASCOT, which analyses free-text responses and suggests likely SOC codes, along with confidence scores. Coders can review these suggestions, accept the highest-confidence code, or manually select the most appropriate code if needed which can improve both efficiency and coding consistency. In this guide, manual office coding refers both to fully manual coding and to computer-assisted or semi-automated coding in which human coders remain responsible for assigning the final occupation code.

Manual post-interview coding remains common in major UK surveys, including Understanding Society, the 1970 British Cohort Study (BCS70), the European Social Survey (ESS) and the National Child Development Study (NCDS). For online surveys, some studies have adapted question wording to prompt respondents to provide richer descriptions or added soft checks (for example, minimum character requirements) to help improve data quality.

Automated post-interview coding of open text

In this approach, respondents still provide open-text answers, but occupation codes are assigned automatically by rule-based systems or statistical/machine-learning models. This method is designed to reduce costs and speed up processing. Rule-based systems rely on expert-defined logic, while machine-learning models are trained on pre-coded datasets and use predictive algorithms to assign codes.

The Office for National Statistics (ONS) has implemented automated coding in the 2021 Census, the Labour Force Survey and the Transformed Labour Force Survey (TLFS). In the 2021 Census, around 65% of occupation responses were coded automatically, with the remainder requiring manual coding.

More recently, large language models (LLMs) have been explored for occupational coding, showing promising coding accuracy and consistency, although performance remains lower at highly detailed classification levels and can be sensitive to variations in wording.

2.2 Self-coding during the interview

Look-up tools

This semi-automated method asks respondents to enter a job title or keywords describing their duties into a look-up tool which returns a ranked list of potential occupation codes based on similarity to the code frame. Respondents select the code that best fits their job. If they cannot find a suitable code, they are usually asked to provide open-text descriptions which are then manually coded.

An interactive look-up tool has been used in Next Steps and piloted in the European Social Survey, BCS70 and ONS's Opinions and Lifestyle Survey.

AI approaches

Recent developments have also explored the use of AI to support occupational coding in real time during survey completion. AI systems can analyse respondents' open-text answers as they are entered and automatically generate targeted follow-up questions when additional information is required to assign an occupation code. Initial evidence suggests that these systems can achieve reliability comparable to trained office coders while reducing coding costs and producing more timely occupational data, although they remain at an early stage of development and require further evaluation.

Closed questions

In this approach, respondents select their occupation from a predefined list of broad occupational groups, often aligned with top-level SOC categories (e.g., 1-digit or 2-digit codes). It has been used in studies such as the Community Life Survey, Participation Survey, Public Attitudes Tracker Survey of Attitudes to the Environment, and several commissioned surveys on the NatCen Opinion Panel.

3. Summary of Methods for Occupation Data Collection and Coding

Method	Summary	Data Quality		Wider considerations	Recommendations
		Pros	Cons		
Post-interview coding: manual/ clerical office coding	• Respondents provide open-text descriptions of their job (e.g. job title, duties) • Responses are coded by trained office coders to a standardised classification scheme after the survey • Often supported by computer-assisted tools (e.g. CASCOT)	• High accuracy and coding success • Suitable for complex or nuanced job descriptions • Fairly high levels of inter-coder reliability • Comparatively high coding consistency across time • Provides an important benchmark for evaluating newer coding approaches	• Labour intensive and costly • Time consuming, which can lead to delayed data availability • Requires training and experienced coders • Can be inconsistent over time	• Requires extensive coder training (and retention of skilled coders) • Updates to classification schemes (e.g., SOC2020) can disrupt coding consistency	• Use for surveys requiring high precision or high granularity (for more than 1 digit level) • Supplement with additional questions (e.g., qualifications, duties) to improve response quality and hence coding quality • Programme checks to query unusually short responses • Periodically benchmark coder performance using inter-coder reliability checks
Post-interview coding: automated coding of open-text descriptions: Rule-based and traditional machine learning	• Open-text responses are automatically coded after data collection • Rule-based systems use predefined rules, keywords or matching algorithms • Statistical/ Machine learning approaches are trained on previously coded data and predict the most likely occupation code	• Faster than manual coding • Reduces manual coding effort and potentially cost • Scalable for large datasets • Can automatically code high-confidence cases	• Accuracy can be lower than expert coding, particular at detailed classification levels • Some responses may remain uncoded or require manual review • Performance depends on the quality of rules or training data • A large amount of training data is required • Ambiguous or unusual occupation descriptions can be difficult to classify	• Initial setup can be expensive and technically complex (algorithm development) • Requires continuous training of algorithms with updated data • Different approaches (e.g., exact/fuzzy matching, dependency parsing) require ongoing maintenance • Machine learning models need a lot of training data • Trade off between assigning more cases and only assigning codes when confidence is high	• Validate automated coding against expert office coding • Use automated coding to support rather than fully replace expert where detailed, high-quality codes are required • Use a hybrid approach in which uncertain or low confidence cases are reviewed by human coders • Maintain appropriate quality assurance and monitoring

Post-interview coding: AI-based coding (transformers and LLMs)	• Open-text responses are coded using transformer-based models or large language models (LLMs) • These models can capture contextual meaning and relationships between occupation descriptions and classification codes • Codes can be generated automatically, with uncertain cases potentially referred for human review	• Promising potential to improve efficiency, scalability, and consistency • Better able than traditional machine learning approaches to capture contextual meaning in job descriptions • Can reduce manual coding effort • May improve identification of less common or difficult occupations	• Performance depends heavily on methodological decisions, including: - prompt design; - embedding model selection; - retrieval and ranking strategies • Results can vary across model families and configurations • Human oversight remains important for uncertain cases and quality assurance	• Organisations with strict data governance requirements may need on-premise solutions. • AI systems introduce new measurement decisions that may affect comparability across studies and over time • Transparency and documentation of model choices become essential for reproducibility • Benchmarking and calibration procedures are necessary before deployment	• Use confidence scores to implement a hybrid workflow: - automatically accept high-confidence cases; - review low-confidence cases manually • Carefully document all modelling choices and prompts • Test multiple models rather than assuming one model is universally best • Maintain human quality control procedures
Self-coding during interview (look-up approach)	• Respondents enter keywords describing their job title and/or main duties • The system searches a look-up table and displays a shortlist of potential occupation codes • Respondents select the option that best matches their job; if no match is found, they provide open text for later manual coding	• Can achieve high coding rates • Reduces the amount of post-survey coding required • Provides occupation codes immediately • Respondents generally report that selected codes describe their jobs well	• Agreement with expert office coding is only moderate • Coding consistency over time is lower than for office coding • Respondents may struggle to identify the correct option • Performance can depend on how closely respondents' wording matches the terminology used in the look-up system	• Respondent-facing terminology is important • Formal occupational classifications may not reflect how people naturally describe their work • User-testing is important to ensure the interface and response options are understandable to participants • Where detailed or highly accurate coding is required, a look-up tool should not necessarily replace office coding	• Further develop look-up tools so that terminology and response options are more intuitive and reflect how respondents naturally describe their work • Combine with manual coding for uncodable responses • Conduct user-testing to maximise usability of interface • Ask participants to rate the accuracy of their selected code • Collect open description for manual coding if accuracy rated poorly or no code assigned

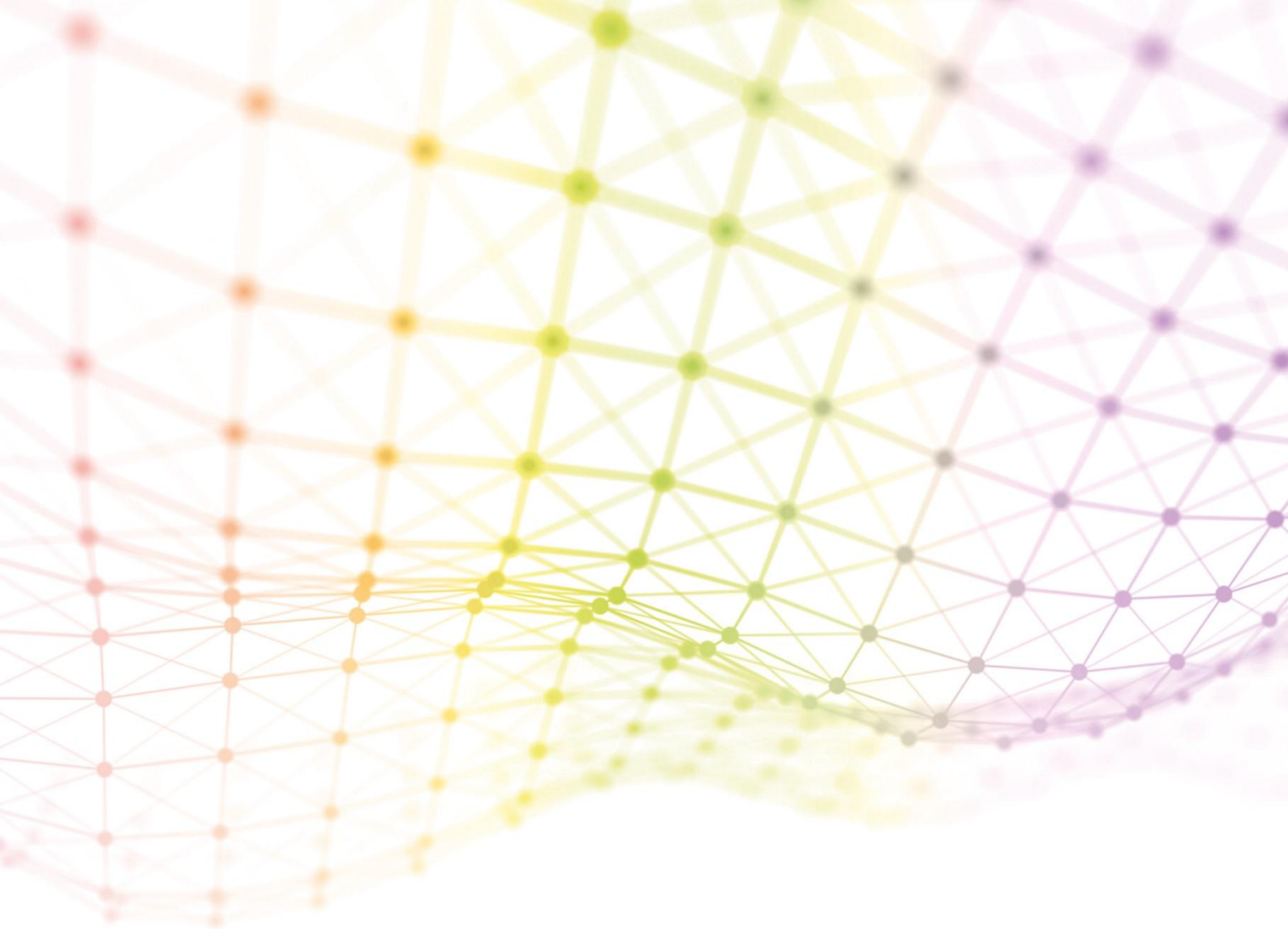
Automated coding during the survey	• Respondents provide open text input during the survey • LLMs analyse the response in real time and assign or suggest an occupation code • Where information is sufficient, the system can generate targeted follow-up questions about job task, industry, qualifications or other relevant details	• Eliminates separate post-fieldwork coding, reducing turnaround time and cost • Dynamic probing collects additional information precisely when needed, improving coding accuracy • Follow-up questions are tailored to the respondent's specific answer rather than using a fixed script • Probes can clarify: - industry; - job tasks; - qualifications; - supervisory responsibilities; - garbled or abbreviated job titles. • Respondent burden remains similar to traditional approaches despite adaptive questioning • The approach can improve data completeness by obtaining information that would otherwise be missing	• AI hallucination probes may introduce measurement error or respondent confusion • Some respondents may refuse AI processing or additional probing • Dynamic coding systems are technically more complex and require integration into survey platforms. They can also create time lags between a participant's response and the LLM-generated follow-up question, which may negatively affect the respondent experience	• Raises issues around informed consent and acceptability of AI use in surveys • Requires infrastructure capable of supporting real-time API calls and latency management • Data privacy considerations are particularly important if responses are transmitted to commercial AI providers • Adaptive questioning may improve respondent experience by asking only relevant questions	• Pilot and evaluate AI probing carefully • Consider self-hosted environments when handling sensitive survey data • Continue validating dynamic coding performance against expert manual office coding, respondent assessments of coding accuracy and longitudinal consistency measures • Design systems which ensure that probes use language which matches how participants naturally describe their work
Closed questions	• Respondents select their occupation from a fixed list of broad categories (e.g., 1-digit SOC)	• Low respondent burden • Low cost and quick to administer • Very high response rates • Suitable when only broad occupation groupings are required	• Provides less detailed occupational information • Agreement with office coding can only be moderate • Respondents may struggle to identify the category that best fits their job • Formal occupational categories may not reflect how respondents naturally think about or describe their work	• Usefulness depends on whether broad occupational groupings suffice for the research purpose • Response options should be intuitive and meaningful for respondents • Category design and terminology should be user-tested • Less suitable where detailed SOC coding is required	• Use for low-budget surveys where collecting detailed occupation data is not core objective

4. Key Recommendations

- Maintain manual post-survey office coding where detailed and high-quality occupation data are required. Manual office coding remains an important benchmark, but coder training, coding protocols and quality assurance are needed to minimise inconsistencies between coders and over time.
- Use automation and AI to support, rather than fully replace, expert coding at present. Machine-learning and LLM-based approaches can improve efficiency, consistency and scalability, but further validation is needed, particularly for detailed occupational classifications. Human oversight should remain for uncertain cases.
- Do not generally replace manual office coding with self-coding during the survey where accurate and detailed codes are required. Look-up tools can achieve high coding rates, but agreement with office coding and consistency over time remain only moderate.
- Use closed-list questions only where broad occupational groupings are sufficient. They can reduce respondent burden and coding costs but are less appropriate when detailed occupation information is needed.
- Make respondent-facing tools more intuitive. Further development of look-up and closed-list approaches should focus on terminology and response options that reflect how respondents naturally understand and describe their work.
- Continue to evaluate real-time AI coding and adaptive probing before widespread use. Particular attention should be paid to transparency, explainability, model stability, data governance, and acceptance by respondents.

5. References

Kocar, S., Brown, M., Calderwood, L. & Koerber, H. (2025). Occupation coding in self-completion surveys: Evidence review. *Survey Futures Report* 20. Available at: https://surveyfutures.net/reports/.



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

verian Formerly Kantar Public

www.surveyfutures.net