SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

Working Paper 25:

Evaluating a Respondent-Led Look-Up Tool for Occupation Coding: Evidence on Feasibility, Validity, and Reliability

Helena Koerber[1], Matt Brown[1], Lisa Calderwood[1]

[1] Centre for Longitudinal Studies, UCL Social Research Institute, University College London

July 2026

Survey Futures is an Economic and Social Research Council (ESRC)-funded initiative (grant 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.

Prior to citing this paper, please check whether a final version has been published in a journal. If so, please cite that version. In the meanwhile, the suggested form of citation for this working paper is:

Koerber H, Brown M & Calderwood L (2026) 'Evaluating a Respondent-Led Look-Up Tool for Occupation Coding: Evidence on Feasibility, Validity, and Reliability', *Survey Futures Working Paper* no. 25. Colchester, UK: University of Essex. Available at https://surveyfutures.net/working-papers/.

Contents

Abstract

Occupation is a key indicator in social surveys, but office coding of open-text job descriptions is costly and time-consuming. Respondent-led look-up tools could offer an efficient alternative by assigning codes during data collection, yet evidence on their performance remains limited. We evaluated the feasibility, validity, and reliability of a look-up tool using a two-wave randomised mode experiment with 1,195 adults in England who remained in the same job across waves. Participants completed surveys online, by video interview, or in person, selected an occupation code using the tool, and provided open-text descriptions for independent office coding. The tool achieved an 86.8% coding success rate, and 97.2% of respondents reported that their selected code described their job fairly or very well. However, four-digit agreement with office coding was 64.7%, compared with 95.1% agreement between office coders, and cross-wave consistency was lower for the look-up tool than for office coding (60% vs. 73.7%). Online respondents reported lower subjective accuracy than those in interviewer-assisted modes. Longer inputs were associated with lower subjective accuracy and agreement, while lexical, but not semantic, similarity predicted consistency. Overall, look-up coding is feasible but does not yet match expert office coding in validity or reliability.

Keywords

longitudinal surveys, look-up tools, occupation coding, self-coding, survey mode, text similarity

Introduction

Occupation is an important variable in social surveys, commonly used to measure social class, economic status, and labour market outcomes (Tijdens, 2022). Reliable occupational data are therefore essential for research on social inequality and mobility. However, collecting and coding occupation information remains a major methodological challenge. Given the vast number of distinct occupations, surveys typically collect this information through open-ended questions, asking respondents to describe their job title and main duties in their own words (Brugiavini et al., 2017; Simson et al., 2023). Responses are then coded to standardised occupation classification categories by expert office coders after the survey. While it is often regarded as the gold standard, post-survey office coding is a time-consuming and costly process.

As large-scale surveys increasingly move online, there is growing interest in respondent-driven and automated approaches that could improve efficiency without compromising data quality. However, online data collection introduces additional challenges for occupation measurement because there is no interviewer available to probe responses, request clarification, or ask tailored follow-up questions when job descriptions are incomplete or ambiguous.

One promising innovation in this area is the use of look-up tools whereby respondents enter keywords describing their occupation and are then presented with a tailored list of potential codes from which they select the most appropriate option (Tijdens, 2015). This method has the potential to reduce costs by providing real-time coding during data collection. Despite its promise, little empirical evidence exists on how well look-up tools perform compared to traditional office coding and how this might vary across modes of data collection.

This paper provides an experimental evaluation of a look-up tool using data from a large-scale methodological study in which participants completed surveys at two time points, two weeks apart, with random assignment to one of three modes (web, video and in-person) at each wave. On each occasion participants provided information about their occupation via a look-up tool and also provided a full open-text description which was coded by two independent office coders after the survey. The performance of the look-up tool is evaluated using four complementary quality indicators:

a. **Coding success** – the proportion of participants who successfully selected an occupational code using the look-up tool, providing an assessment of feasibility.

b. **Subjective accuracy** – respondents' own assessment of how well their selected code reflects their actual job, offering insight into the validity of self-selected codes.

c. **Agreement with office coding** – the extent to which look-up-assigned codes match those assigned by expert office coders, providing a direct comparison of quality relative to the traditional gold standard approach and a second indicator of the validity of the look-up tool.

d. **Consistency across waves** – the extent to which participants who had not changed jobs selected the same occupation code at the two time points, providing an assessment of the reliability of look-up coding.

Based on these quality indicators, four research questions guide the analysis:

RQ1: How successful are respondents at selecting an occupation code using the look-up tool and how does this vary by mode and participant input?

RQ2: How do respondents rate the accuracy of the occupation code they select using the look-up tool and how does this vary by mode and participant input?

RQ3: How closely do respondent selected look-up codes align with codes assigned by office coders and how does this vary by mode and participant input?

RQ4: Among respondents who have not changed jobs, how consistent are respondents at selecting look-up codes across two time points and how does this compare to the consistency of office coding over time? How does this consistency vary based on whether participants switched modes? To what extent is coding consistency attributable to the similarity of text inputs across time points?

By analysing the quality indicators across modes and two experimental waves, this paper contributes to methodological research on occupation coding by offering new evidence on the feasibility, validity and reliability of the look-up approach with office coding as a benchmark.

Literature review

Occupation coding in social surveys and methodological challenges

Occupation is a key variable in social surveys which is widely used as an indicator of socio-economic status, income, and lifestyle (Tijdens, 2022). Due to its strong association with social

inequality and stratification, accurate measurement of occupation is essential for high-quality survey data. Traditionally, occupation has been measured through interviewer-administered surveys, where open-ended responses are manually coded by trained office coders (Lyberg & Dean, 1992). Respondents are typically asked to provide their job title and a short description of tasks they perform (Brugiavini et al., 2017; Simson et al., 2023). After fieldwork, expert office coders code these free-text responses to standardised occupational classifications (Brugiavini et al., 2017). Widely used systems include the Standardized Occupation Classification (U.S. Bureau of Labor Statistics, n.d.), the UK Standard Occupational Classification (SOC) 2020 (Office for National Statistics, n.d.), and the International Standard Classification of Occupations (ISCO; International Labour Organization, 2010). These coding frames are structured hierarchically. For example, UK SOC codes contain four increasingly detailed classification levels. To support post-survey coding, coders often use computer-assisted tools such as CASCOT, designed to make occupational coding "simpler, quicker, and more reliable" (Brugiavini, et al., 2017; Tijdens, 2015; Office for National Statistics, 2024).

Based on the expertise involved and the systematic procedures employed, office coding has long been regarded as the "gold standard" of occupation coding (Burstyn et al., 2014). However, manual office coding is time-consuming and costly (Gweon et al., 2017). In addition, the approach is not without its limitations. Respondents may provide incomplete or ambiguous descriptions (Belloni et al., 2016; Conrad et al., 2016) and coding reliability has been questioned in several studies (Mathiowetz, 1992; Mellow & Sider, 1983; Kim & Kim, 2025). As a result, methodological research has increasingly sought ways to improve reliability and validity while trying to reduce cost and time (Kocar et al., 2025). Nevertheless, despite these efforts, the available evidence indicates that alternative approaches have yet to fully replicate the quality achieved through traditional office coding (see Kocar et al., 2025 for evidence review).

Advances towards automated and self-coding methods

Recent methodological innovations have highlighted the potential of automated and self-coding methods of occupation both for post-survey coding and during data collection. These include software-assisted coding, respondent self-coding, and automated algorithms that classify free-text (Hacking et al., 2006; Mannetje & Kromhout, 2003; Ossiander & Milham, 2006; Safikhani et al., 2023; Schierholz & Schonlau, 2021). Additionally, AI tools for occupation coding are attracting increasing attention, with

applications both during the survey (Sturgis et al., 2026) and in post-survey coding processes (Kononykhina, 2026).

Respondent coding during the interview has received growing attention as it has potential benefits for accuracy, cost efficiency, and timeliness (Schierholz et al., 2018). This method has been tested in interviewer-administered settings (Schierholz et al., 2018), self-completion and mixed-mode surveys (Peycheva et al., 2021) as well as in longitudinal studies (Peycheva et al., 2021).

Many studies adopt hybrid strategies, in which respondents attempt to code their occupation during the survey, but open text descriptions are collected for office coding when this process fails (Peycheva et al., 2021). Combining both approaches may lead to higher overall coding rates, improved data quality, and can provide a more comprehensive basis for assessing coding reliability (Burstyn et al., 2014).

Emergence of look-up coding tools

Among respondent-driven approaches, look-up coding has gained particular interest. It shifts classification responsibility from trained coders to survey participants which can reduce costs but can also introduce new sources of response error (Tijdens, 2015; Brugiavini et al., 2017). In this approach, respondents (or interviewers acting on their behalf) enter job-related information into a search interface. An algorithm then returns a list of potential job titles with associated occupation codes, which are typically ordered based on closeness of the match between text inputs and the code descriptions. Respondents are then asked to select the most appropriate option (Simson et al., 2023). In interviewer-administered modes, interviewers can assist respondents in finding the best-matching code (Peycheva et al., 2021).

Although promising, the look-up approach remains underexplored, with limited empirical evidence on its performance compared with traditional office coding and on how its effectiveness may vary across survey modes. Evaluating its performance across multiple dimensions is therefore crucial for understanding respondents' ability to self-classify occupations during data collection and for assessing its viability as an alternative to manual office coding.

Quality indicators for evaluating look-up coding

Coding rates

Coding rates, the proportion of cases in which an occupation code is successfully assigned, are one of the most widely used indicators of the feasibility and efficiency of occupation coding methods (Kocar et al., 2025; Peycheva et al., 2021; Schierholz et al., 2018; Gweon et al., 2017).

Research using look-up tools has shown promising results. For example, Peycheva et al. (2021), drawing on data from the mixed-mode Next Steps Age 25 Study in the UK reported coding rates of around 90% in web and telephone interviews, though lower rates were observed in face-to-face interviews (69%). Kocar et al. (2025), using data from the Next Steps Age 32 Survey, reported an overall look-up coding rate of 82% (across web, telephone and in-person interviews), compared to near-complete coverage for manual office coding (≈99%).

Comparing these look-up coding rates with the coding rates achieved through automated statistical coding systems suggests look-up systems perform relatively well. For instance, Helppie-McFall and Sonnega (2018) reported production rates of 60% using automated algorithms in the Health and Retirement Study, substantially below the 97% rate achieved through manual coding. Similar limitations have been reported by the UK Office for National Statistics (2021), where automated occupation coding in the 2021 Census was unable to classify around one third of entries and required substantial manual follow-up. These findings highlight a key trade-off in automation efforts: methods that remove human involvement often struggle to match the "completeness" achievable through professional office coding.

In this context, assessing the coding rate of the look-up approach as one of the quality indicators allows us to evaluate whether it represents a feasible alternative to traditional office coding.

Subjective accuracy

Subjective accuracy refers to respondents' own assessment of how well the selected occupation code reflects their job. When respondents evaluate their code choice, their judgement provides insight into whether the classification likely reflects their actual occupation and therefore serves as an important indicator of validity (Kocar et al., 2025).

Previous research suggests that subjective evaluations are strongly associated with coding accuracy. Kocar et al. (2025) found that the look-up codes selected by respondents who rated their selected code as matching their occupation "very well" were significantly more likely to be consistent with codes assigned by office coders.

Subjective accuracy offers a respondent-centred lens through which to evaluate look-up coding quality. By capturing whether participants feel the assigned code reflects their occupation, it helps assess the validity of the look-up approach and the extent to which respondents are meaningfully engaging with the classification task.

Agreement with office coding

Agreement with office coding can be used as an indicator of validity in occupation coding. The agreement rate assesses the extent to which a code assigned by the method under evaluation (e.g., the look-up tool) matches a code assigned by experienced office coders (Campanelli et al., 1997). Because office coding is typically carried out by trained experts following standardised procedures, it is commonly used as the benchmark against which alternative methods are evaluated (Helppie-McFall & Sonnega, 2018; Belloni et al., 2016; Burstyn et al., 2014; Patel et al., 2011; Russ et al., 2023; Gweon et al., 2017). Previous research suggests that agreement between codes generated by automated systems and codes assigned by office coders are typically modest. For example, Russ et al. (2023) reported a 50% match-rate between algorithm generated and office assigned 6 digit-level occupation codes.

The same benchmark has been applied to evaluate participant self-coding methods using a look-up tool (Kocar et al., 2025; Schierholz et al., 2018). Schierholz et al. (2018) compared interviewer-assisted look-up coding during telephone surveys with post-survey expert office coding and found only small differences in agreement (agreement between two office coders: 65.78%; agreement between interviewer using the look-up tool and an office coder: 61.80%). These results suggest that look-up-based occupation coding can be a viable approach. In contrast, Kocar et al. (2025) found considerably lower agreement between respondent look-up selections and independent office coders (62% at the 4-digit SOC level, compared to inter-coder agreement among office coders of approximately 90%) raising questions about the accuracy of this self-coding approach.

Consistency across time and survey waves

In longitudinal studies, consistency of occupational coding across survey waves is critical for ensuring reliable measures of occupational change. If respondents are doing the same job at two time points but their occupation is coded differently, this introduces measurement error that can distort analyses of occupational mobility (Kim & Kim, 2025). Kim and Kim (2025) examined trends in occupational coding mismatches and found that inconsistencies have increased in recent years, leading to inflated estimates of occupational mobility. Specifically, they showed that mobility appeared substantially higher when occupations were coded *independently* across waves compared to when *dependent coding* was used, that is,

when coders could refer to the respondent's previous occupation code. This suggests that part of the observed occupational mobility in longitudinal surveys may reflect coding variability rather than genuine change.

Similarly, Perales (2014) highlighted the impact of measurement error in occupational mobility studies, emphasising that potential inconsistencies in coding across time can bias conclusions about career progression and labour market dynamics. These studies demonstrate the importance of understanding coding consistency across survey waves to ensure occupation measures can be used accurately in (longitudinal) analyses.

The present study evaluates the stability of look-up-based occupational coding and office coding across two waves. By examining cross-wave coding consistency, this study contributes to understanding whether the look-up tool produces stable occupational classifications over time, and whether its consistency is comparable to that achieved through professional office coding.

Influences on the quality of occupation coding

Occupational coding quality is not only shaped by the coding approach itself but can also be influenced by contextual and respondent-related factors. Two of the most prominent influences discussed in the literature are *survey mode* and *characteristics of participant input*. These factors directly relate to the quality indicators used in the present study.

Survey Mode

As previously mentioned, occupation has traditionally been measured in interviewer-administered surveys where respondents describe their job titles and duties to trained interviewers, who transcribe this information for post-survey office coding (Lyberg & Dean, 1992; Peycheva et al., 2021). Interviewers can probe for missing or unclear details, helping to ensure sufficiently precise descriptions for accurate classification.

However, the widespread adoption of self-administered web surveys has transformed this process. Respondents are now often asked to provide job information independently, without any interviewer guidance. This shift introduces several potential challenges for data quality.

The look-up method (Tijdens, 2015), designed for both interviewer-assisted and self-administered environments, offers an important test case for examining mode effects. In interviewer-assisted modes, interviewers can help respondents navigate the look-up interface and select the most appropriate occupational category. In self-administered settings, respondents must rely on their understanding of the tool. Prior research suggests that coding

rates and accuracy tend to be higher in interviewer-assisted modes than in self-completion contexts (Kocar et al., 2025; Peycheva et al., 2021), underlining the role of interviewer assistance in maintaining coding quality. In the present study, mode effects are examined alongside the four quality indicators to assess whether and how interviewer assistance influences the quality of look-up-based coding.

Participant Input

The quality of occupation coding is reliant on the information provided by respondents, particularly the length, detail and consistency of job titles and descriptions. Previous research has suggested that overly lengthy responses can reduce coding success. For example, Conrad et al. (2016) analysed double-coded job descriptions in the U.S. Current Population Survey and found that longer responses were *less reliably* coded across independent coders, a finding echoed in subsequent research (Kocar et al., 2025).

Evidence on look-up coding is still limited, but existing studies suggest that length of participant input plays a crucial role. Kocar et al. (2025) found that intermediate job title lengths (14 to 25 characters) were associated with higher look-up coding rates (feasibility) whereas longer keyword descriptions (more than 25 characters) negatively affected both look-up coding rates and agreement between look-up and office-assigned codes.

Understanding what constitutes high-quality input is essential to determine how best to prompt and guide respondents when using the look-up tool. To address this, the present study examines how the length of participants' job titles and job descriptions influences coding success, subjective accuracy, and agreement with office coders for look-up coding.

Participant input may also affect coding consistency over time. When respondents describe the same job using different wording across waves, this can lead to different codes being assigned. For look-up coding, different keywords will generate different drop-down lists, potentially leading respondents to select different codes. Office coding may also be affected, as varying descriptions can lead coders to different classification decisions, particularly where coders use assistive coding software that generates code suggestions based on the text provided and as such may produce different recommendations for the same job. The present study therefore examines whether changes in participants text input across survey waves are associated with lower coding consistency.

Methods

Data

This study draws on data from a large-scale experimental study conducted by the UCL Centre for Longitudinal Studies (CLS) to assess the impact of survey mode on several key measures used in recent waves of two major UK longitudinal cohort studies: *Next Steps* (Age 32 survey) (Wu et al., 2024) and the *Millennium Cohort Study* (MCS, Age 23 survey) (Connelly & Platt, 2014; Joshi & Fitzsimons, 2016). The target population was adults aged 20–40 years living in England, selected to ensure broad comparability with participants in Next Steps and MCS. The questionnaire included measures of participants' occupation, income, financial literacy, mental health, other socio-demographic characteristics and a cognitive assessment.

Participants were recruited using in-home or in-street screening on a free-find basis in clustered pre-defined sample points which were designed to reflect the regional distribution of the population across England. Soft quotas were applied to ensure broad representativeness by gender, age, and region, with additional monitoring of education, employment status, and social grade. In total, 1,800 participants were recruited across 200 clustered sample areas.

Participants were asked to complete two surveys, with an approximately two-week interval between wave 1 and wave 2. Respondents received a £20 voucher for each completed wave. Survey mode (in-person, video or web) was randomly assigned at each wave creating nine experimental groups covering all possible combinations, including repeating the same mode twice. At the time of recruitment, respondents were not made aware of their assigned survey modes. The full mode allocation structure is shown in shown in Table 1.

Additionally, to correct for sampling imbalances, survey weights were produced to align the sample with population distributions for gender, age, working status, region, education, and social grade. The calibration targets used to construct the survey weights are provided in Supplementary Table A1.

Fieldwork was managed by Ipsos UK and conducted by Kantar Global Services Ltd (KGS) between 30 November 2023 and 21 February 2024. The mean interval between Wave 1 and Wave 2 responses was 15.7 days (SD = 3.3 days). A total of 1,692 respondents participated in Wave 1, while 1,510 completed Wave 2, as recruitment ceased shortly after the target sample size of 1,500 was reached. The analytical sample for this study was comprised of the 1195 individuals who were working in the same job at both time points.

Table 1

Mode groups

Group	Survey 1	Survey 2
1	Online	Online
2	Online	Teams
3	Online	In-person
4	Teams	Online
5	Teams	Teams
6	Teams	In-home
7	In-person	Online
8	In-person	Teams
9	In-person	In-home

Variables

Occupation coding

In both survey waves, occupations were coded using two methods: (1) Coding via a look-up tool during the survey, and (2) office coding by two independent coders after the survey. This approach allows for a direct comparison of the look-up method with traditional office coding.

Look-up coding

In each survey, participants were first asked to code their occupation using the look-up tool during the survey. In interviewer-assisted modes (video and in-person) the data entry was conducted by the interviewers, and they assisted participants with selecting their occupation code. The process worked as follows:

a. **Data entry:** Respondents entered their *job title* (maximum 50 characters) and *keywords describing what they do in their job* (maximum 200 characters). These questions referred to the respondent's *current or main job*.

b. **Search algorithm:** The look-up tool used a trigram search function to process text entries. Input strings were broken into overlapping three-character sequences (trigrams) to accommodate partial matches and spelling variations. These trigrams were compared against the UK job title index, generating a ranked list of potential

occupational codes. Occupations with the fewest missing words compared to the respondent's input appeared first. Up to 35 occupation options were displayed.

c. **Refinement:** Respondents could edit their job title or keywords to refresh the list of suggested occupations if a suitable match was not initially found.

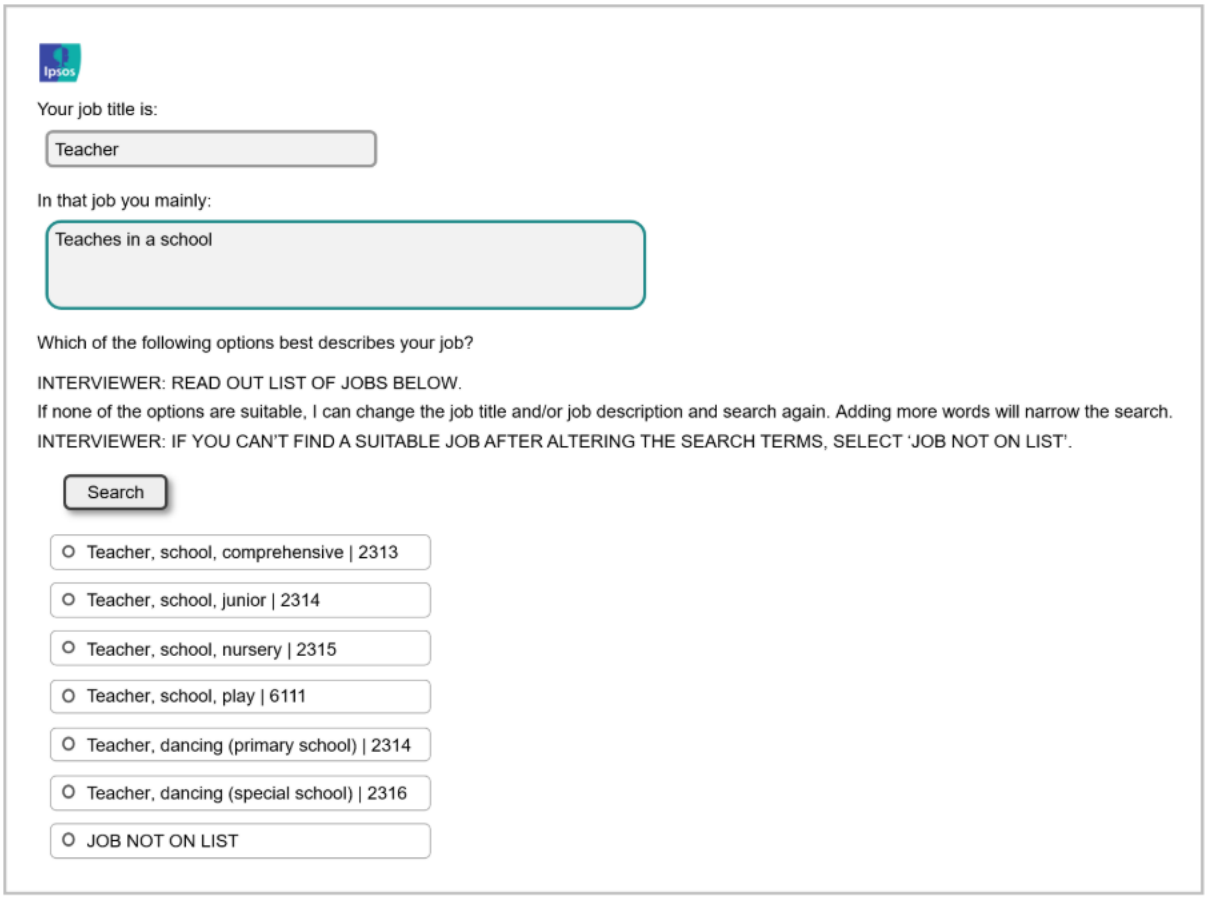
d. **Selection:** Respondents (or interviewers, after discussion with respondents) selected the occupation that best described their job. If no appropriate match was available, participants could select "Job not in the list."

e. **Subjective Accuracy:** To assess the subjective accuracy of the selected job code, respondents were asked: *"How well do you think the option you selected actually describes the job that you do?"*. The response options were "very well", "fairly well", "not well", and "not at all well".

See figure 1 for an example of the Look-up tool interface.

Figure 1

Example of Look-up coding interface during the survey



Office coding

After completing the look-up self-coding process, all respondents were asked to provide an open-text description of their job for later office coding. Specifically, respondents were asked:

"This approach to collecting information about your job is new and we are testing it out. To help us check whether it is working, could you also describe in your own words what you mainly do in your job? Please describe in detail (for example, the type of work, the department you are in, and what level you work at)."

Office coding was conducted post interview using the job title that was entered into the look-up tool and the open-text description. Occupation was coded blindly by two experienced office coders. The office coders remained the same for both survey waves to ensure consistency. Coding was conducted with assistance of the CASCOT software. CASCOT uses a rule-based and index-matching system for free-text descriptions. The software compares input text to a structured coding index containing occupational titles and associated codes. It applies predefined linguistic rules (e.g., handling synonyms, abbreviations, spelling variants, and word order) to standardise and interpret the text. CASCOT then calculates similarity scores between the processed input and indexed entries, ranks potential matches, and provides a list of occupations with a confidence score (0–100) indicating the strength of the match. The system relies on structured matching and rule logic rather than statistical machine-learning training (IER, 2018).

In this study, CASCOT was used in a semi-automated mode: the software's suggestions and confidence scores were visible, but final coding decisions were made manually based on coder judgement. This setup mirrors professional office-coding standards. An illustration of the CASCOT-assisted coding interface is shown in Figure 2.

Figure 2

Example of CASCOT coding interface

Note. The example is taken from the CASCOT website (https://cascotweb.warwick.ac.uk/#/classification/soc2020). Although the screenshot displays 6-digit codes, this study used 4-digit SOC codes. The input included both participants job titles and job descriptions.

The parallel use of look-up and office coding in both survey waves provides a robust framework for evaluating the look-up tool across all key quality indicators: coding rates, subjective accuracy, agreement with office coding, and coding consistency over time.

Table 2 outlines the technical differences between the look up tool and CASCOT system.

Table 2

Mechanisms of CASCOT and the Look-up tool

Step	**CASCOT**	**Look-up tool**
Underlying approach	Rule-based scoring using occupation knowledge and index matching	Character-level fuzzy matching using trigrams
Input processing	Words/ phrases (e.g., “senior”, “nurse”, “manager”)	Breaks text into overlapping 3-character sequences (e.g., “nur”, “urs”, “rse”)
Data resources used	Structure file (occupation hierarchy), index file (job titles → codes), rule files (heuristics and weights)	Job-title index with linked occupation codes
Matching logic	Compares parsed words to index entries; applies rules about word importance (core occupation vs modifiers); weights matches	Compares trigrams in input string to trigrams in index entries; counts character similarity

Quality indicators of occupation coding

a. ***Coding success***

Coding success is a measure of the feasibility of the method. For look-up coding, success is defined as the participant (or interviewer) successfully selecting an occupation code using the look-up tool during the survey. For office coding, it refers to the successful assignment of an occupation code by an office coder post-survey.

b. ***Subjective accuracy***

After selecting a job code using the look-up tool, participants were asked how well the chosen code described their current job. This subjective accuracy measure, available only for the look-up tool, provides insight into the validity of look-up coding.

c. ***Agreement with office coding***

Agreement with office coding provides a second indicator of the validity of look-up coding. Agreement rates were calculated between look-up coding and the first office coder for Wave 1 for all four SOC digit levels. As a benchmark, this was compared to the agreement between the two independent office coders for Wave 1, again at all four SOC digit levels.

d. ***Coding consistency across waves***

The reliability of the two coding approaches was assessed through coding consistency, measured by whether the same occupation code was selected in both survey waves. For look-up coding, consistency was defined as participants selecting the same SOC code at each SOC digit level. For office coding, consistency was assessed by comparing the codes assigned by the same office coder (office coder 1) across both waves. As mentioned previously, participants were asked in Survey 2 whether they had changed jobs, and those who reported a job change were excluded from the analysis.

Participant Input

To understand how participant input influences the quality of the occupation coding methods the length of the job title and job keywords were combined into a measure of text input length. The two inputs were combined as this is also how the look-up tool processes the text. This input length measure was used to examine whether the amount of information provided affects the feasibility and accuracy of look-up coding. Length was measured by character counts. To address non-linearity, the text length variable was recoded into categorical groups with five groups of approximately equal size.

Text similarity analysis

To evaluate the impact of text similarity on coding consistency, lexical and semantic similarity scores were calculated by comparing participant text inputs at waves 1 and 2. Semantic similarity scores were computed using the spaCy natural language processing library in Python with the pretrained English model en_core_web_lg. Semantic similarity was calculated using cosine similarity between the aggregated vector representations (Montani et al, 2023). Lexical similarity was measured using the Jaccard similarity coefficient, calculated on lemmatised text inputs and ranging from 0 (no shared terms) to 1 (identical terms).

For logistic regression analyses, lexical and semantic similarity were modelled as continuous variables to retain information across their full 0–1 scales. Correlations and variance inflation factors were examined before model estimation. As there was no evidence of problematic multicollinearity between the two indicators (r = .60; VIF = 1.64), both were included simultaneously in the regression models. The linearity of each predictor in the logit was assessed using the Box–Tidwell procedure. There was no evidence of non-linearity in the look-up model, although potential non-linearity was identified for both lexical and semantic similarity in the office-coding model. Quadratic sensitivity models and predicted-probability plots were therefore examined. The quadratic lexical-similarity term was statistically significant, whereas the quadratic semantic-similarity term was not. As the alternative specifications did not change the substantive conclusions, the continuous linear specification was retained for parsimony and comparability across coding methods. Full linearity checks and sensitivity analyses are reported in Supplementary Tables C1–C2 and Supplementary Figure C1.

Mode and mode switching

For coding success, subjective accuracy and agreement with office coding, analyses examined whether outcomes varied by mode. For coding consistency, analyses focused on whether consistency varied for participants who switched modes compared to those who stayed in the same mode. Specifically, three mode groups were compared: 1) those who stayed in the same mode, 2) those who switched from one interviewer-assisted mode to another interviewer-assisted mode (video to in-person or vice versa), 3) those who switched from online to an interviewer-assisted mode (video or in-person) or vice versa.

Statistical analysis

Descriptive statistical analyses were conducted for all four quality indicators to assess the performance of the look-up tool. For coding success, agreement with office coding,

and coding consistency, descriptive statistics were also produced for office coding to provide a benchmark against which look-up coding could be compared. To examine the impact of participant input on look-up performance, logistic and multinomial regression models were estimated. Python Spacy was used to generate semantic similarity indicators for text similarity of participant input at waves 1 and 2. To test for differences across survey modes, Chi-squared tests were conducted for each quality indicator. For all analyses, weights were used.

Results

The results are structured according to the four quality indicators used to evaluate the performance of the look-up tool: coding success, subjective accuracy, agreement with office coding, and consistency across survey waves. For each indicator except for subjective accuracy, results are presented for the look-up tool with office coding serving as the benchmark. Results are further disaggregated by survey mode to examine potential mode effects and by characteristics of participant input to assess their influence on coding quality. Additional descriptive results for the participant-input analyses are provided in Supplementary Tables B1–B4.

Coding success

To address RQ1 we compare look-up coding rates with office coding rates. The look-up coding rate reflects the proportion of participants (or interviewers) that successfully selected a job code with the look up tool. The office coding rate shows the proportion of jobs that were successfully office coded using the open text that participants provided. Furthermore, we look at whether coding rates differed by mode and participant input.

As shown in Table 3, the overall coding rate for the look-up tool was 86.75%, compared to 100% for office coder 1. Look-up coding rates were highest in the video mode (88.09%) and lowest in the in-person mode (86%), although variation across modes was not significant (x^2 (1,1192) = .0989, p = .7532). These findings indicate that the majority of participants were able to successfully select a job code using the look-up tool, demonstrating a high level of feasibility despite being lower than office coding rates.

Table 3

Survey 1 coding Success rates by mode for the Look-up tool and office coder 1

Mode	Look-up coding rate	Office coding rate
Web (n = 401)	86.16%	100%
Video (n = 410)	88.09%	100%
In-person (n = 384)	86%	100%
Total (n = 1,195)	86.75%	100%

To examine whether participant input affected coding success, a logistic regression model was estimated with combined input length (job title and job keywords length summed) as the predictor (Table 4). Compared with the reference category (0–35 characters), none of the input length categories were significantly associated with successful coding. These findings suggest that respondents' ability to select an occupational code using the look-up tool was not influenced by the amount of text they entered.

Table 4

Logistic regression for participant input on coding success

Predictor	OR	95%CI	p-value
Text length			
0 – 35 characters	*Reference category*		
36 – 49 characters	.92	[.45; 1.89]	.825
50 – 63 characters	1.09	[.47; 2.5]	.841
64 – 86 characters	1.11	[.53; 2.4]	.774
87 – 230 characters	.78	[.39; 1.57]	.49
constant	6.78	[3.95; 11.63]	p<.001

Note. Reference category: 0 - 35 characters due to largest number of observations.

Subjective accuracy

To address RQ2 we examined the second quality indicator – subjective accuracy – which captures participants' perception of how well the selected look-up occupation code represents their actual job. As above, we also explore whether this varies by mode and participant input.

As seen in Table 5, most participants indicated their selected job code reflects their actual job either very well (54.77%) or fairly well (42.47%) suggesting generally positive perceptions of code accuracy. Comparing across modes, a X^2 test indicates significant differences (F (3.89, 4123.77) = 4.1818, p = .0025). Online respondents were significantly less likely to rate their selected job code as reflecting their job "very well" (45.55%) compared to in-person respondents (54.48%) and video respondents (64.05%), suggesting that in the absence of interviewer assistance online participants were less confident about the accuracy of the selected code.

Table 5

Subjective Accuracy by mode for the Look-up tool for survey 1

Mode	Subjective accuracy		
	Very well	**Fairly well**	**Not very/ at all well**
Online	45.55%	51.07%	3.38%
In-person	54.48%	42.54%	2.98%
Video	64.05%	34%	1.95%
Total	54.77%	42.47%	2.75%

Note. "Not very well" and "not at all well" were combined into one category due to low number of responses.

To examine whether input length influenced perceived accuracy, a multinomial logistic regression model was estimated with *"very well"* as the reference outcome category (Table 6) and 0–35 characters as the reference input category.

For the comparison between *"fairly well"* and *"very well,"* participants providing longer inputs - either 64–86 characters (OR = 1.89, p = .013) or 87–230 characters (OR = 1.94, p = .010) - had significantly higher odds of reporting that the occupation code described their job *fairly well* rather than *very well* compared to those providing the shortest responses. The shorter

intermediate categories (36–49 and 50–63 characters) were not significantly different from the reference group.

For the lowest accuracy category (*"not very/not at all well"* vs. *"very well"*), longer input lengths were associated with substantially higher odds of lower perceived accuracy. Responses containing 50–63 characters (OR = 11.37, p = .028), 64–86 characters (OR = 19.88, p = .007), and 87–230 characters (OR = 34.68, p = .001) showed significantly increased odds of being rated as poorly fitting compared to the shortest inputs. The 36–49-character category was not statistically significant (p = .070).

Overall, these findings indicate that longer text inputs were associated with lower levels of confidence in the accuracy of the selected look-up code.

Table 6

Multinomial logistic regression of subjective accuracy on combined input length for the look-up tool

Predictor		**OR**	**95%CI**	**p-value**
Very well	*Base outcome*			
Fairly well	0 – 35	*Reference category*		
	36–49	1.25	[0.75; 2.10]	.382
	50–63	1.14	[0.69; 1.89]	.598
	64–86	1.89	[1.15; 3.11]	.013
	87–230	1.94	[1.17; 3.22]	.010
	cons	.55	[0.39; 0.79]	.001
Not very/ not at all well	0 – 35	*Reference category*		
	36–49	8.20	[0.83; 81.70]	.073
	50–63	11.37	[1.31; 98.79]	.028
	64–86	19.88	[2.27; 173.36]	.007
	87–230	34.68	[4.32; 278.21]	.001
	cons	.004	[0.001; 0.027]	<.001

Note. Reference category: 0 – 35 characters due to largest number of observations.

Agreement with office coding

To address RQ3 we examine the third quality indicator – agreement between look-up coding and office coding. As office coding is regarded as the gold standard, high agreement rates should indicate high quality.

Table 7 shows agreement rates between the look-up code selected at wave 1 and the first office coder – at all four SOC digit levels. At the one-digit level, the agreement between look-up and office coding is 77.67% but it drops down to 64.73% at the more precise 4-digit level. This is much lower than the agreement rates between the two office coders which range from 98.39% at 1-digit level to 95.13% at the 4-digit level.

Agreement rates were also examined across modes. There were no significant differences across modes for either the agreement between look-up and office coding or for the agreement between two office coders.

Table 7

Agreement between look-up coding and office coder 1 and agreement between the two office coders by mode for survey 1

Mode	Agreement look-up tool & 1st office coder				Agreement 1st office coder & 2nd office coder			
	1-digit	2-digit	3-digit	4-digit	1-digit	2-digit	3-digit	4-digit
Online	80.10%	74.82%	72.03%	65.72%	98.33%	96.76%	96.54%	94.97%
In-person	74.99%	70.79%	68.83%	63.40%	98.42%	97.15%	96.89%	95.09%
Video	77.92%	74.19%	69.74%	65.05%	98.43%	97.33%	97.06%	95.32%
Total	77.67%	73.27%	70.19%	64.73%	98.39%	97.08%	96.83%	95.13%

To examine whether participant input influenced four-digit agreement between look-up and office coding, a logistic regression model was estimated with combined input length as a categorical predictor (Table 8). Responses containing 0–35 characters served as the reference category.

Compared with the shortest responses, most input length categories were not significantly associated with agreement. However, where participants provided the longest inputs (87–230 characters) the odds of an exact four-digit match between look-up and office

coding was significantly lower (less than half as likely) (OR = 0.45, p = .002) than those who provided the shortest inputs.

Overall, these findings suggest that agreement between look-up and manual coding decreases only for very long responses, while moderate increases in input length do not significantly affect coding agreement.

Table 8

Logistic regression of four-digit agreement between look-up and office coding on combined input length

Predictor	OR	95%CI	p-value
Input length			
0 – 35	*Reference category*		
36 – 49	1.05	[.61; 1.80]	.860
50 - 63	.69	[.41; 1.16]	.161
64 – 86	.67	[.40; 1.12]	.129
87 – 230	.45	[.27; .75]	.002
constant	2.52	[1.75; 3.64]	<.001

Note. Reference category: first category: 0 – 35 characters due to largest number of observations.

Consistency across waves

To address RQ4, coding consistency across survey waves was examined for both the look-up tool and office coding. As shown in Table 9, coding consistency declined with increasing digit precision for both methods, from 76.12% to 59.96% for the Look-up tool and from 83.05% to 73.71% for office coding. This pattern reflects the greater difficulty of reproducing exact 4-digit codes across waves - as the number of digits increases, the greater the number of possible categories, and so both variations in how jobs are described and the inherent subjectivity in the classification decision becomes more likely to result in different codes being assigned. Across all digit levels, office coding remained more consistent than the look-up tool, indicating that manual office coding produces more stable and reliable occupational classifications over time.

Table 9

Coding consistency by digit level for Look-up coding and office coding

Digit Level	Look-up Coding consistency	Office Coding consistency
1-digit	76.12%	83.05%
2-digit	70.68%	79.95%
3-digit	67.13%	78.23%
4-digit	59.96%	73.71%

Table 10 further examines whether mode-switching between waves influenced coding consistency. For the look-up tool, there were no statistically significant differences in consistency across mode-switching patterns at any SOC level (e.g., 4-digit: $\chi^2 = 2.61$, $p = .074$). This suggests that look-up coding performance was relatively stable regardless of whether participants completed both waves in the same mode or switched modes. In contrast, switching between online and interviewer-assisted modes significantly reduced the consistency of office coding, particularly at the 2- and 3-digit levels ($\chi^2 = 5.71$ and 5.85, both $p < .01$). These findings indicate that manual coding was more sensitive to mode changes, possibly due to differences in how job information was expressed across survey formats.

Text Similarity analysis

To better understand the drivers of coding consistency observed across waves, we examined the extent to which consistency in occupational codes reflected consistency in the text inputs provided by respondents. Specifically, we assessed lexical similarity (overlap in wording) and semantic similarity (similarity in meaning) between Wave 1 and Wave 2 occupational descriptions and investigated whether these measures were associated with coding consistency.

As seen in Table 11, text similarity between Wave 1 and Wave 2 showed a clear pattern: semantic similarity was consistently much higher than lexical similarity for both the Look-up tool and office coding.

Table 10

Coding consistency by mode-switching for office and look-up coding

Digit	Same mode	Switch video & in-person	Switch online & interviewer-assisted	Rao-Scott corrected Chi2	p-value
Look-up					
1-digit	76.70%	79.90%	73.67%	.9522	.3844
2-digit	73.74%	74.35%	66.32%	1.9986	.1363
3-digit	71.90%	69.71%	62%	2.6791	.0692
4-digit	66.30%	58.06%	55.93%	2.6095	.0740
Office					
1-digit	88.36%	82.86%	79.15%	4.0315	.0180
2-digit	85.85%	82.12%	74.43%	5.7121	.0034
3-digit	84.44%	80.26%	72.55%	5.8528	.0029
4-digit	77.77%	75.81%	69.62%	2.5832	.0759

Table 11

Lexical and semantic word similarity across wave 1 and wave 2 for Look-up and office coding text inputs

Similarity analysis	Mean similarity LU text inputs	Mean similarity Office coding text inputs
Job title lexical similarity	.64	.64
Job title semantic similarity	.84	.84
Job description lexical similarity	.26	.25
Job description semantic similarity	.77	.77
Combined lexical	.35	.35
Combined semantic	.85	.85

Note. The Look-up tool used job title and keywords describing participants occupation. Office coding used the Look-up job title and a separate open text description.

This difference was particularly pronounced for job keywords (for the look-up tool) and job descriptions (for office coding). While exact word overlap across waves was relatively low, semantic similarity remained high, indicating that respondents often rephrased their job descriptions but retained the same underlying meaning. The same pattern was observed for the combined input measure and was nearly identical for office coding. This suggests that respondents' occupational descriptions are conceptually stable over time, even when wording varies. Importantly, neither the Look-up tool nor CASCOT (used for office coding) incorporates semantic similarity measures, relying instead primarily on lexical matching.

A logistic regression model was estimated to examine whether lexical and semantic similarity between text input of Wave 1 and Wave 2 were associated with four-digit coding consistency for the Look-up tool. Results showed that lexical similarity was significantly associated with coding consistency (OR = 16.79, $p < .001$), indicating that higher lexical overlap between responses substantially increased the odds of receiving the same occupational code across waves, holding semantic similarity constant. In contrast, holding lexical similarity constant, semantic similarity was not significantly associated with coding consistency (OR = 1.10, p = .92, 95% CI [0.15, 8.36]). These findings suggest that coding consistency in the Look-up tool is primarily driven by similarity in wording rather than similarity in meaning.

Table 12

Logistic regression of four-digit look-up coding consistency on lexical and semantic text similarity across waves

Predictor	OR	p-value	95% CI
Lexical word similarity	16.79	<.001	[5.65; 49.88]
Semantic word similarity	1.10	.92	[.15; 8.36]
cons	.52	.41	[.11; 2.43]

A second logistic regression model examined whether lexical and semantic similarity were associated with four-digit coding consistency for office coding. Consistent with the Look-up results, holding semantic similarity constant, lexical similarity was significantly associated with coding consistency (OR = 4.48, p = .008, 95% CI [1.47, 13.65]), indicating that greater overlap in wording across waves increased the odds of consistent occupational coding. Semantic similarity, however, was not significantly associated with coding consistency (OR = 0.89, p =

.93, 95% CI [0.12, 6.62]), holding lexical similarity constant. Overall, these results indicate that for office coding as well, the association between coding consistency and text similarity seems to be primarily driven by lexical overlap rather than semantic similarity between responses.

Table 13

Logistic regression of four-digit office-coding consistency on lexical and semantic similarity across waves

Predictor	OR	p-value	95% CI
Lexical word similarity	4.48	.008	[1.47; 13.65]
Semantic word similarity	.89	.93	[.12; 6.62]
cons	1.86	.44	[.40; 8.85]

Discussion and conclusion

This study evaluated the performance of a look-up tool for occupation coding using four quality indicators: coding success, subjective accuracy, agreement with office coding, and coding consistency across two survey waves. By integrating experimental mode variation at two time points as well as participant input length and input similarity across survey waves, the study provides a comprehensive assessment of both feasibility and data quality.

The coding success rate of the look-up tool was broadly consistent with previous research at around 87% (e.g., Peycheva et al., 2021; Kocar et al., 2025; Schierholz, 2018), indicating that respondents are generally able to select an occupation code. As in earlier studies (Helppi & Sonnega, 2018; Kocar et al., 2025), office coding achieved near-complete coding success, confirming its robustness as a benchmark. In contrast to some earlier findings (Peycheva et al., 2021; Schierholz, 2018), coding success did not vary by survey mode. This likely reflects the randomised mode allocation in this study design, which reduces confounding between respondent characteristics and survey mode. Furthermore, input length was not found to impact upon coding rates. Compared with short responses, longer inputs neither increased nor decreased the odds of successful coding. While these findings indicate that the look-up tool is operationally feasible, they also reinforce a practical limitation noted in prior research (Peycheva et al., 2021; Kocar et al., 2025): self-coding alone does not guarantee complete coverage. To avoid loss of data, look-up coding should therefore be supplemented with office coding for cases where no code is assigned.

Most respondents reported that their selected look-up occupation code described their job either very well or fairly well. Consistent with Kocar et al. (2025), subjective accuracy ratings varied by survey mode, with online respondents expressing lower confidence in their assigned codes compared to respondents in interviewer-assisted modes. This pattern suggests that interviewer presence may increase respondents' confidence in the coding outcome. This is likely due to interviewers being able to ask probing questions about participants jobs during the interview (Conrad et al., 2016) and assist respondents in choosing an occupation. When examining the role of input characteristics, differences in subjective accuracy emerged primarily for longer response categories. Compared with short inputs, respondents providing longer descriptions were more likely to rate the assigned occupation code as less accurate. Moderately long responses did not significantly differ from short inputs, suggesting that decreases in perceived accuracy occur mainly once responses become relatively long rather than increasing gradually with input length. Several explanations may account for this finding. Longer text inputs may indicate more complex occupations that require more elaborate descriptions and are inherently more challenging to code. Alternatively, the look-up tool may function more effectively with concise input, while extended inputs may introduce greater ambiguity and additional sources of error. Importantly, subjective accuracy reflects respondents' perceptions rather than objective validity. As noted by Kocar et al. (2025), perceived accuracy is positively associated with agreement with office coding, but high confidence does not necessarily imply correct classification. Subjective evaluations, while informative, should be interpreted cautiously.

Agreement between look-up and office coding followed patterns observed in previous studies (Kocar et al., 2025; Schierholz et al., 2018; Russ et al., 2023): agreement was moderate and declined at more detailed digit levels, whereas agreement between trained office coders remained high (4-digit LU agreement: 65% vs 95% Office coder agreement). Longer inputs reduced the odds of four-digit agreement compared to short descriptions. Again, this could suggest that more elaborate descriptions may signal more complex or ambiguous occupations that are harder to classify consistently or that, rather than improving precision, additional detail may introduce ambiguity if different aspects of the job point toward different occupational categories. It is important to also emphasise that the high inter-coder agreement observed for office coding (95% at the 4-digit level) likely reflects the use of standardised training, identical software, and harmonised procedures (Kocar et al., 2025; Kim & Kim, 2025). As noted in previous work (Schierholz et al., 2018; Conrad et al., 2016; Kim et al., 2020), agreement rates decline substantially when coding systems differ. Thus, the gap

between look-up and office coding may partly reflect the exceptional consistency of the office coding process rather than being explained fully by poor performance of the look-up tool. Nevertheless, the findings underscore that self-coding tools cannot yet fully replicate expert office coding procedures.

Coding consistency over time declined as digit-level precision increased for both the look-up tool and office coding. For the look-up tool, consistency dropped to approximately 60% at the 4-digit level. This is notable given that respondents effectively coded their own occupations. One might expect greater stability under self-coding, yet a substantial proportion of participants selected different occupational codes across waves. This suggests that even when individuals report on the same job, small differences in interpretation, search strategies, or understanding of occupational classifications can lead to different coding outcomes. Office coding exhibited greater overall stability than the look-up tool; however, its consistency across two survey waves remained considerably lower than its inter-coder agreement within the same wave. At the 4-digit level, inter-coder agreement within a single wave reached 95%, whereas consistency across waves was only 74%. This discrepancy underscores an important methodological distinction: high agreement between professional coders at a single time point does not necessarily translate into stability over time. Reliability depends not only on coding procedures and training but also on the consistency of respondents' descriptions, survey conditions and coder judgements across waves.

For office coding, mode switching had a significant effect on coding consistency. Coding consistency was significantly lower when respondents switched between online self-completion and interviewer-assisted modes. Interviewer-administered surveys typically involve probing and clarification, which can elicit job descriptions which provide the detail required for accurate coding. In contrast, self-completion modes provide less guidance, potentially resulting in more variable, incomplete or tailored responses (Conrad et al., 2016). In large-scale longitudinal surveys—where mixed-mode designs are increasingly common—such effects may have important implications. Changes in survey mode across waves could increase the likelihood of spuriously classifying participants as occupationally mobile when in fact they remain in the same job. These findings highlight the importance of accounting for survey design features when interpreting longitudinal occupational change.

The text similarity analyses help explain why occupational coding is challenging, regardless of whether coding is carried out by respondents or trained office coders. In both cases, coding consistency was more closely related to lexical similarity than semantic similarity. This is perhaps unsurprising given that both respondent look-up tools and many office coding

systems, including CASCOT, rely heavily on the specific words used in occupational descriptions. As a result, they may struggle to recognise that differently worded descriptions are referring to the same occupation.

These findings also raise questions about treating office coding as a stable benchmark against which alternative approaches should be assessed. Although agreement between coders was very high, coding consistency across waves was considerably lower. Even among respondents who had remained in the same job, changes in how that job was described could lead to different occupational codes being assigned.

For respondent-led coding, this issue may be exacerbated by the fact that occupational classification systems are not always intuitive to survey participants. Recent mental-models research suggests that many respondents find the terminology and underlying structure of SOC and SIC classifications difficult to navigate, creating a disconnect between how they think about their work and how classification systems organise occupations (Watson & Jones, 2026). This may help to explain why some respondents selected different occupation codes across waves despite providing very similar job descriptions.

The findings should be interpreted in light of several limitations. The analyses draw on two survey waves involving participants aged 20 to 40, conducted within a specific experimental survey setting and using a single occupational classification framework. The extent to which the results generalise to other populations, survey contexts, or coding systems therefore remains unclear. In addition, while the text similarity measures provide useful evidence about one source of coding inconsistency, they cannot tell us how respondents understood occupational categories, how they searched for occupations within the look-up tool, or how office coders arrived at their final coding decisions when reviewing software-generated suggestions.

Overall, the results suggest that respondent-led look-up tools are not yet able to replace expert office coding when highly accurate occupational measures are required. At the same time, their performance was encouraging. The vast majority of occupations were coded successfully, and respondents generally considered the resulting classifications to be accurate. The next stage of development is therefore less about extending coding coverage and more about improving coding accuracy and consistency over time.

Future work should focus on both the technical design of coding tools and the experience of the people using them. This could involve providing clearer guidance, helping respondents produce concise but informative occupational descriptions, and exploring coding frameworks that

better reflect how people understand and describe their jobs. AI-based approaches may be particularly promising in this respect. Large language models have the potential to interpret meaning rather than relying solely on keyword matching, ask tailored follow-up questions, and adapt dynamically to respondent input. If these capabilities can be applied successfully to occupational coding, they may reduce the impact of variation in wording while retaining the efficiency benefits of respondent-led approaches. As online and mixed-mode surveys become increasingly common, developing more robust approaches to occupational coding will be critical to maintaining the quality of occupational data.

References

Belloni, M., Brugiavini, A., Meschi, E., & Tijdens, K. (2016). Measuring and Detecting Errors in Occupational Coding: an Analysis of SHARE Data. *Journal of Official Statistics*, *32*(4), 917–945. https://doi.org/10.1515/jos-2016-0049

Brugiavini, A., Belloni, M., Martens, M., & Buia, R. E. (2017). The "Job Coder." In F. Malter & A. Börsch-Supan (Eds.), *SHARE Wave 6: Panel innovations and collecting dried blood spots* (pp. 51–70). Munich Center for the Economics of Aging, Max Planck Institute for Social Law and Social Policy. https://share-eric.eu/fileadmin/user_upload/Methodology_Volumes/201804_SHARE-WAVE-6_MFRB.pdf

Burstyn, I., Slutsky, A., Lee, D. G., Singer, A. B., An, Y., & Michael, Y. L. (2014). Beyond crosswalks: Reliability of exposure assessment following automated coding of Free-Text job descriptions for occupational epidemiology. *The Annals of Occupational Hygiene*, *58*(4), 482–492. https://doi.org/10.1093/annhyg/meu006

Campanelli, P., Thomson, K., Moon, N., & Staples, T. (1997). The quality of occupational coding in the United Kingdom. *Wiley Series in Probability and Statistics*, 437–453. https://doi.org/10.1002/9781118490013.ch19

Connelly, R., & Platt, L. (2014). Cohort profile: UK Millennium Cohort Study (MCS). *International Journal of Epidemiology*, *43*(6), 1719–1725. https://doi.org/10.1093/ije/dyu001

Conrad, F. G., Couper, M. P., & Sakshaug, J. W. (2016). Classifying Open-Ended Reports: Factors affecting the reliability of occupation codes. *Journal of Official Statistics*, *32*(1), 75–92. https://doi.org/10.1515/jos-2016-0003

Emery, T., Cabaco, S., Fadel, L., Lugtig, P., Toepoel, V., Schumann, A., Lück, D., & Bujard, M. (2023). Breakoffs in an hour-long, online survey. *Survey Practice*, *16*(1), 1–12. https://doi.org/10.29115/sp-2023-0008

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics*, *33*(1), 101–122. https://doi.org/10.1515/jos-2017-0006

Hacking, W., J. Michiels, and S. Jansen, S. 2006. "Computer Assisted Coding by Interviewers." In Proceedings of the 10th International Blaise Users Conference, IBUC 2006, 9–12 May, Arnhem, The Netherlands. Available at: http://blaiseusers.org/2006/ Papers/291.pdf (accessed May 2026)

Helppie-McFall, B., & Sonnega, A. (2018). *Feasibility and reliability of automated coding of occupation in the Health and Retirement Study* (Michigan Retirement Research Center Research Paper No. 2018-392). University of Michigan. https://doi.org/10.2139/ssrn.3338502

Institute for Employment Research. (2018). *CASCOT: Computer assisted structured coding tool*. University of Warwick. Retrieved February 18, 2026, from https://cascotweb.warwick.ac.uk/

Joshi, H., & Fitzsimons, E. (2016). The Millennium Cohort Study: the making of a multi-purpose resource for social science and policy. *Longitudinal and Life Course Studies*, *7*(4). https://doi.org/10.14301/llcs.v7i4.410

Kim, A. T., & Kim, C. (2025). The Rise in Occupational Coding Mismatches and Occupational Mobility, 1991–2020. *Sociological Methods & Research*, *55*(2), 659–699. https://doi.org/10.1177/00491241241303517

Kirby, G., Carson, J., Dunlop, F., Dibben, C., Dearle, A., Williamson, L., Garrett, E., & Reid, A. (2015). Automatic methods for coding historical occupation descriptions to standard classifications. In G. Bloothooft, P. Christen, K. Mandemakers, & M. Schraagen (Eds.), *Population Reconstruction* (pp. 43–60). Springer. ISBN 978-3-319-19883-5.

Kocar, S., Peycheva, D., Brown, M., Sakshaug, J. W., Bhaumik, C., & Calderwood, L. (2025). *Evaluating the feasibility and quality of participant self-coding of occupation in a mixed-mode survey*. [Manuscript submitted for publication].

Kocar, S., Brown, M., & Calderwood, L. (2025). 'Occupation coding in self-completion surveys: Evidence review'. *Survey Futures Report no. 3.* Colchester, UK: University of Essex. Available at https://surveyfutures.net/reports/.

Kononykhina, O. (2026, June). *Can LLMs Advance Occupational Coding? Evidence and Methodological insights.* Survey Futures Occupation Coding Workshop, Universtiy College London, London, United Kingdom. https://surveyfutures.net/wp-content/uploads/2026/06/Can-LLMs-Advance-Occupational-Coding.pdf

Lyberg, L., & Dean, P. (1992). *Automated coding of survey responses: An international review* (R&D Report 1992:2). Statistics Sweden. https://www.scb.se/contentassets/7c4edb581f8745e3a081e1ba9b332eb4/rnd-report-1992-02-green.pdf

Mannetje, A. & Kromhout, H. (2003). The use of occupation and industry classifications in general population studies. *International Journal of Epidemiology*, *32*(3), 419–428. https://doi.org/10.1093/ije/dyg080

Mathiowetz, N. A. (1992). Errors in reports of occupation. *Public Opinion Quarterly*, *56*(3), 352. https://doi.org/10.1086/269327

Mellow, W., & Sider, H. (1983). Accuracy of response in labor market surveys: Evidence and implications. *Journal of Labor Economics*, *1*(4), 331-344.

Montani, I., Honnibal, M., Honnibal, M., Boyd, A., Van Landeghem, S., & Peters, H. (2023). explosion/spaCy: v3.7.2: Fixes for APIs and requirements. *Zenodo (CERN European Organization for Nuclear Research)*. https://doi.org/10.5281/zenodo.1212303

Office for National Statistics. (2024). "CASCOT: Computer Assisted Structured Coding Tool", [online] Available at https://warwick.ac.uk/fac/soc/ier/data_group/cascot/

Office for National Statistics. (2023). *Automated Text Coding: Census 2021*. Retrieved August 2025 from https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/automatedtextcodingcensus2021

Ossiander, E. M., & Milham, S. (2006). A computer system for coding occupation. *American Journal of Industrial Medicine*, *49*(10), 854–857. https://doi.org/10.1002/ajim.20355

Patel, M. D., Rose, K. M., Owens, C. R., Bang, H., & Kaufman, J. S. (2011). Performance of automated and manual coding systems for occupational data: A case study of historical

records. *American Journal of Industrial Medicine*, *55*(3), 228–231. https://doi.org/10.1002/ajim.22005

Perales, F. (2014). How wrong were we? Dependent interviewing, self-reports and measurement error in occupational mobility in panel surveys. *Longitudinal and Life Course Studies*, *5*(3), 299–316. https://doi.org/10.14301/llcs.v5i3.295

Peycheva, D. N., Sakshaug, J. W., & Calderwood, L. (2021). Occupation Coding during the interview in a Web-First sequential Mixed-Mode survey. *Journal of Official Statistics*, *37*(4), 981–1007. https://doi.org/10.2478/jos-2021-0042

Russ, D. E., Josse, P., Remen, T., Hofmann, J. N., Purdue, M. P., Siemiatycki, J., Silverman, D. T., Zhang, Y., Lavoué, J., & Friesen, M. C. (2023). Evaluation of the updated SOCcer v2 algorithm for coding free-text job descriptions in three epidemiologic studies. *Annals of Work Exposures and Health*, *67*(6), 772–783. https://doi.org/10.1093/annweh/wxad020

Safikhani, P., Avetisyan, H., Föste-Eggers, D., & Broneske, D. (2023). Automated occupation coding with hierarchical features: a data-centric approach to classification with pre-trained language models. *Discover Artificial Intelligence*, *3*(1). https://doi.org/10.1007/s44163-023-00050-y

Schierholz, M., & Schonlau, M. (2021). Machine Learning for Occupation Coding—A comparison Study. *Journal of Survey Statistics and Methodology*, *9*(5), 1013–1034. https://doi.org/10.1093/jssam/smaa023

Schierholz, M. (2018). Eine Hilfsklassifikation mit Taetigkeitsbeschreibung fuer Zwecke der Berufskodierung. A StA Wirtschafts- und Sozialstatistisches Archiv, 12(3-4), 285-298.

Simson, J., Kononykhina, O., & Schierholz, M. (2023). OccupationMeasurement: A comprehensive toolbox forInteractive occupation coding in surveys. *The Journal of Open Source Software*, *8*(88), 5505. https://doi.org/10.21105/joss.05505

Sturgis, P., Robinson, T. S., Fung, L., & Roberts, C. (2026). SOCBot: Using large language models to dynamically measure and classify occupations in surveys. *Sociological Methods & Research*. https://doi.org/10.1177/00491241261461516

Tijdens, K. (2015). Self-identification of occupation in web surveys: requirements for search trees and look-up tables. *UvA-DARE (University of Amsterdam)*, *2015*, 11. https://doi.org/10.13094/smif-2015-00008

Tijdens, K. (2022), “The Importance of Occupation Coding Quality: Lessons for EU-SILC from SHARE and Other International Surveys,” in Improving the Measurement of Poverty and

Social Exclusion in Europe: Reducing Non-Sampling Errors, eds. P. Lynn, and L. Lyberg, Luxembourg: Publications Office of the European Union, pp. 413-426.

Walzenbach, S., Burton, J., Couper, M. P., Crossley, T. F., & JÄckle, A. (2023). Experiments on multiple requests for consent to data linkage in surveys. *Journal of Survey Statistics and Methodology*, *11*(3), 518–540. https://doi.org/10.1093/jssam/smab053

Watson, D., & Jones, B. (2026, 4 June). Using mental models research to review how we collect Standard Inducstrial (SIC) and Standard Occupational Classification (SOC) information [Workshop presentation]. Survey Futures Occupation Coding Workshop, London, United Kingdom. https://surveyfutures.net/wp-content/uploads/2026/06/Using-mental-models-research-to-review-how-we-collect-Standard-Industrial-Classification-.pdf

Wu, A. F., Henderson, M., Brown, M., Adali, T., Silverwood, R. J., Peycheva, D., & Calderwood, L. (2024). Cohort Profile: Next Steps—the longitudinal study of people in England born in 1989–90. *International Journal of Epidemiology*, *53*(6). https://doi.org/10.1093/ije/dyae152

Appendix A. Sample weights

Experimental mode groups were weighted to targets for gender within age and working status, region, social grade and education. Targets were derived from the Office for National Statistics Annual Population Survey, October 2022 to September 2023, for adults aged 20–40 years in England, and from PAMCo 1 2021 for social grade.

Table A1

Calibration targets applied within each experimental mode group

Characteristic	Category	Total	Male	Female
Total	All	100.00%	50.51%	49.49%
Social grade	AB	25.44%	—	—
	C1	32.17%	—	—
	C2	20.77%	—	—
	DE	21.62%	—	—
Age	20–24	21.68%	11.05%	10.63%
	25–29	24.15%	12.29%	11.86%
	30–34	25.21%	12.67%	12.54%
	35–40	28.96%	14.50%	14.46%
Region	North West	12.70%	—	—
	North East	4.55%	—	—
	Yorkshire and the Humber	9.63%	—	—
	West Midlands	10.43%	—	—
	East Midlands	8.17%	—	—
	South East	14.88%	—	—
	East of England	10.32%	—	—
	South West	8.93%	—	—
	Greater London	20.38%	—	—
Working status	Working (codes 1, 2, 3, 6)	81.13%	42.77%	38.35%
	Not working (codes 4, 5, 7–10)	18.87%	7.74%	11.14%
Education	Graduate (codes 7–9)	45.64%	—	—
	Non-graduate (codes 1–6)	54.36%	—	—

Note. A dash indicates that no sex-specific target was applied for that characteristic.

Appendix B. Additional descriptive results

The following tables provide additional survey-weighted descriptive results for the participant-input analyses reported in the manuscript.

Table B1

Coding success and four-digit agreement with office coding by combined text length

Combined text length (characters)	Coding success	Agreement with office coding
0–35	87.0%	72.0%
36–49	86.0%	73.0%
50–63	88.0%	64.0%
64–86	88.0%	63.0%
87–230	84.0%	53.0%
Total	87.0%	64.7%

Note. Agreement is defined as an exact four-digit match between the look-up code and the code assigned by office coder 1.

Table B2

Subjective accuracy by combined text length

Combined text length (characters)	Very well	Fairly well	Not very/not at all well
0–35	64.0%	36.0%	0.2%
36–49	58.0%	40.0%	2.0%
50–63	60.0%	38.0%	3.0%
64–86	47.0%	49.0%	3.0%
87–230	45.0%	49.0%	6.0%
Total	55.0%	42.0%	3.0%

Note. Percentages may not sum to 100 because of rounding. “Not very well” and “not at all well” were combined because of small cell sizes.

Table B3

Mean combined text length by subjective-accuracy rating

Subjective-accuracy rating	Mean combined text length (characters)
Very well	59.67
Fairly well	68.16
Not very/not at all well	81.54

Table B4

Mean combined text length by four-digit agreement with office coding

Agreement status	Mean combined text length (characters)
No agreement	69.67
Agreement	60.71

Appendix C. Text similarity and sensitivity analyses

Lexical and semantic similarity were entered as continuous predictors because both measures range from 0 to 1 and categorisation would discard information and require arbitrary cut-points. The two measures could be included together without problematic multicollinearity (r = .60; VIF = 1.64). Linearity in the logit was assessed using Box–Tidwell transformations, followed by quadratic sensitivity analyses where indicated.

Table C1

Box–Tidwell tests of linearity in the logit

Coding method	Transformed term	p-value	Conclusion
Look-up tool	Lexical similarity × ln(lexical similarity)	.644	No evidence of non-linearity
Look-up tool	Semantic similarity × ln(semantic similarity)	.734	No evidence of non-linearity
Office coding	Lexical similarity × ln(lexical similarity)	.045	Potential non-linearity
Office coding	Semantic similarity × ln(semantic similarity)	.042	Potential non-linearity

Note. Results are from survey-weighted logistic regressions. N = 970 for the look-up model and N = 1,191 for the office-coding model.

Table C2

Unweighted quadratic sensitivity model for four-digit office-coding consistency

Predictor	b	SE	z	p-value	95% CI
Lexical similarity	3.84	1.09	3.52	< .001	[1.71, 5.98]
Lexical similarity^2	−2.35	1.11	−2.12	.034	[−4.53, −0.18]
Semantic similarity	5.13	5.23	0.98	.327	[−5.13, 15.39]
Semantic similarity^2	−3.61	3.54	−1.02	.308	[−10.56, 3.34]
Constant	−1.56	1.92	−0.81	.417	[−5.32, 2.20]

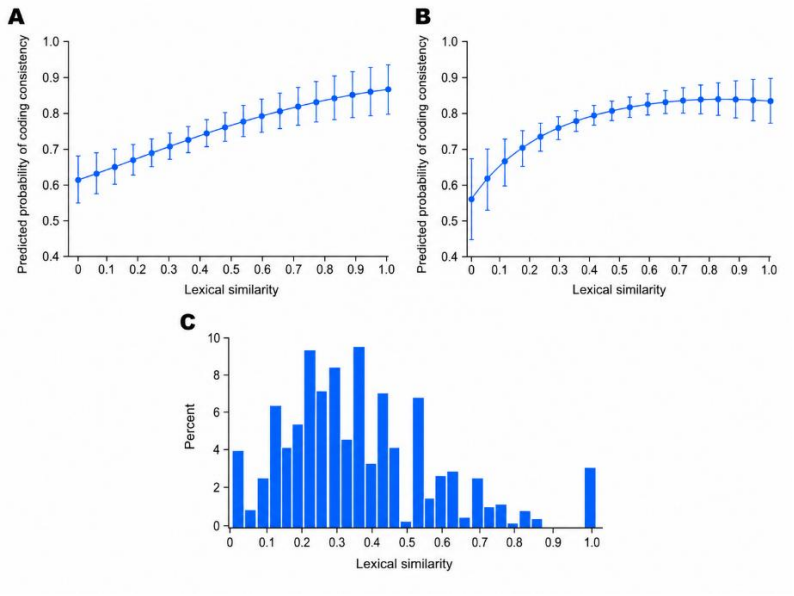
Note. N = 1,191. Adding the two squared terms improved fit relative to the linear model, LR $\chi^2(2) = 6.43$, $p = .040$. The lexical-similarity squared term was significant; the semantic-similarity squared term was not.

The quadratic specification indicated that the association between lexical similarity and office-coding consistency increased across most of the observed range and flattened at high levels of

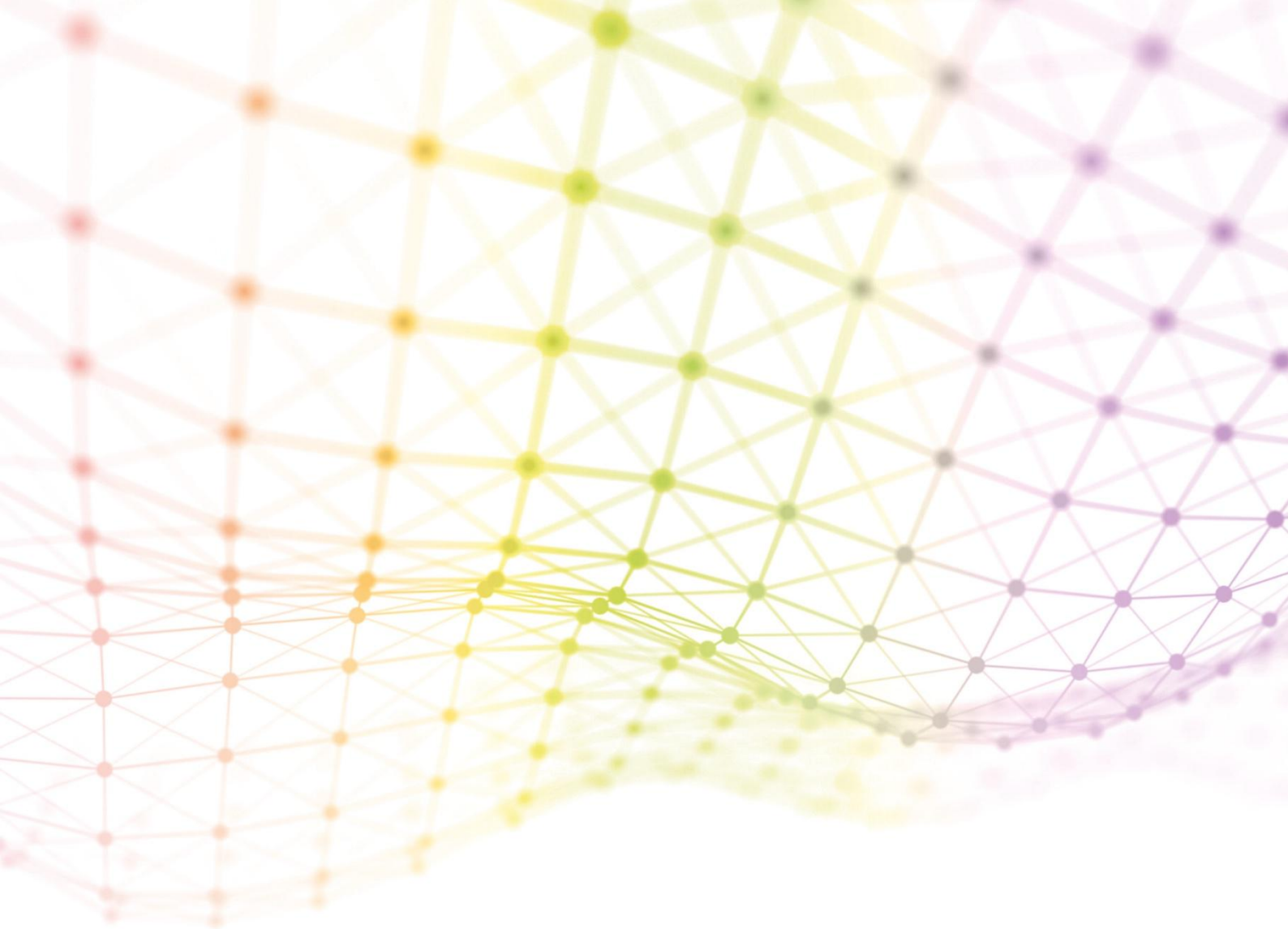
similarity. Predicted probabilities from the linear and quadratic specifications produced the same substantive conclusion.

Figure C1

Linear and quadratic predicated probabilities of the association between lexical similarity and four-digit office-coding consistency



The primary analyses therefore retained lexical and semantic similarity as continuous linear predictors. This specification preserved the full scale, allowed direct comparison between the look-up and office-coding models, and was more parsimonious than categorised or quadratic alternatives. The sensitivity analyses did not change the substantive finding that lexical similarity was associated with coding consistency whereas semantic similarity was not. Because the measures range from 0 to 1, the reported odds ratios represent the change in the odds of consistency associated with moving across the full observed similarity scale.



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