SURVEY FUTURES
SURVEY DATA COLLECTION METHODS COLLABORATION

Working Paper 24:

Industry and occupation coding: A comparison of office-based coding and a closed-list approach

Cristian Domarchi[1], Matt Brown[2], Olga Maslovskaya[1], Lisa Calderwood[2], Curtis Jessop[3]

[1] University of Southampton; [2] University College London, [3] NatCen

July 2026

www.surveyfutures.net

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:

Domarchi C, Brown M, Maslovskaya O, Calderwood L & Jessop C (2026) 'Industry and occupation coding: A comparison of office-based coding and a closed-list approach', *Survey Futures Working Paper* no. *n*. Colchester, UK: University of Essex. Available at https://surveyfutures.net/working-papers/.

Contents

Industry and occupation coding: A comparison of office-based coding and a closed-list approach

Cristian Domarchi[1*], Matt Brown[2], Olga Maslovskaya[1], Lisa Calderwood[2], Curtis Jessop[3]

[1] Department of Social Statistics and Demography, University of Southampton, UK
[2] Centre for Longitudinal Studies, University College London, UK
[3] National Centre for Social Research (NatCen), UK

*Corresponding author. Email address: c.domarchi@soton.ac.uk. Postal address: University Road, Southampton, SO17 1BJ, United Kingdom

Abstract

Social surveys frequently collect information on industry and occupation, typically using open-ended questions about job title and duties, with responses coded by expert coders. The shift towards self-completion surveys has made this increasingly challenging, as respondents may provide insufficient detail without interviewers present to probe for clarification. Closed-list questions, in which respondents select from pre-defined categories, could reduce fieldwork and coding costs while removing ambiguities inherent to open-ended responses. This paper assesses whether closed-list questions produce classifications comparable to those derived from manually coded open-text responses. Using data from the NatCen panel, a UK probability-based panel, we find overall agreement rates between self-selected and office-coded categories are 61.7% for industries and 55.8% for occupations at the highest classification level (1-digit), with considerable variation across categories, and much lower agreement rates (45.8%) at the 2-digit level. Although these results appear to caution against the use of closed-list methods, the reduced question response times and lower costs point to clear practical advantages, particularly if highly detailed classifications are not required. However, for this approach to succeed, the categories presented to respondents would need refinement to align with how individuals conceptualise their jobs. Further research is needed to establish how to achieve this.

Keywords: Industry coding, Occupation coding, Online surveys, Self-administered surveys

1 Introduction

1.1 Background

Industry and occupation are essential measures for research across the social and economic sciences. Industry refers to the economic or business sector in which a respondent works, while occupation describes the specific tasks and activities they perform at their jobs (Park et al., 2015). Social surveys typically collect this information using open-ended questions (Office for National Statistics, 2021). For occupation, respondents are usually asked for their job title and a description of their job including main duties, tasks and responsibilities. For industry, respondents are asked about the main activity of their organisation, business, or freelance work. These responses are subsequently provided to expert office coders who classify them into standardised industry and occupation classification systems (Office for National Statistics, 2021).

Although classification systems vary across countries, all organise industries and occupations into hierarchical groups. In the UK, the most widely used industry classification is the UK Standard Industrial Classification of Economic Activities (SIC 2007), which contains 21 major categories (1-digit sections), further subdivided into 88 divisions (2-digit), 272 groups (3-digit), and 615 classes (4-digit) (Office for National Statistics, 2022). For occupation, the Standard Occupational Classification (SOC 2020) is used, comprising 10 major groups (1-digit) which are sub-divided into 26 sub-major (2-digit), 103 minor (3-digit), and 436 unit (4-digit) groups (Office for National Statistics, 2023b). Both classification systems are periodically updated to reflect evolving job roles, industries, and labour market trends (Office for National Statistics, 2023b, 2026). For example, the UK's Office for National Statistics (ONS) has recently updated the industry classification system to produce the SIC2026 classification (Office for National Statistics, 2026).

Office-coding free-text descriptions of industry or occupation, frequently assisted by specialist software, has traditionally been regarded as the "gold standard" for classification, particularly in specialist surveys examining economic activity, labour force characteristics, social inequality, or employment-related health outcomes. This approach is often expensive, time-consuming, and the requirement to provide detailed open-text responses can place a significant burden on respondents. In interviewer-administered surveys (whether in-person or by telephone), interviewers play a critical role by probing vague or ambiguous responses to obtain the specificity required for accurate office classification (Conrad et al., 2016; Kim et al., 2020). They can also ask respondents to provide additional information in case their responses are short, or the information provided lacks sufficient detail. In contrast, self-administered online surveys present a distinct challenge, as the absence of an interviewer impedes the possibility of such real-time clarification (Peycheva et al., 2021).

1.2 Methods for industry and occupation coding in self-completion surveys

Several new methods have been developed to address the challenges of manual coding in self-completion surveys, including post-survey coding, coding during the interview, and closed-list question approaches.

1.2.1 *Automated approaches for post-survey coding*

Automated approaches for post-survey coding using either word-matching techniques based on predefined rules (e.g. Jones & Elias, 2004; Jung et al., 2008; Ossiander & Milham, 2006) or advanced machine learning models (e.g. Dahl et al., 2024; Gweon et al., 2017; Russ et al., 2016; Savic et al., 2022) have been employed to classify open-text responses into standardised codes. These methods usually assign a confidence level to each suggested code, with the highest-confidence code typically being selected. For example, the CASCOT software (Jones & Elias, 2004) is widely used for industry and occupation coding in the UK and facilitates fully automated or semi-automated coding where low-confidence or ambiguous cases are referred to a human coder for manual review. More recently, large language models (LLMs) to conduct automated coding (e.g. Safikhani et al., 2023; Schierholz & Schonlau, 2021). Although some automated and semi-automated techniques report high agreement rates compared with professional manual coding (Helppie-McFall & Sonnega, 2018), there are often trade-offs. In particular, targeting high accuracy can reduce the proportion of respondents assigned a code (Gweon et al., 2017). In addition, automated coding systems may fail to capture subtle nuances in text descriptions, which can potentially lead to misclassification (Schierholz & Schonlau, 2021).

1.2.2 *Coding during the interview*

Coding during the interview is another possible alternative. In self-administered online surveys, this interactive approach typically uses look-up tables, where a list of potential occupation codes or categories is automatically generated based on the open-ended text inputs (Brugiavini et al., 2017; Kocar et al., 2025; Peycheva et al., 2021; Tijdens, 2015). Previous research has demonstrated that around 80% of participants can successfully select an occupation code using a look-up tool (Kocar et al., 2025; Peycheva et al., 2021); but agreement rates between look-up codes and office codes can be fairly low (around 78% at the 1-digit level, dropping to around 62% at the 4-digit level). The general conclusion from these studies is that self-selected look-up codes selected by the respondent should be supplemented with office coding for quality checks.

In interviewer-administered surveys, coding during the interview is usually conducted by the interviewer who enters the response input into a look-up table or a coding index which produces a list of appropriate categories that are similar to the responses provided (Belloni et al., 2016; Gweon et al., 2017; Schierholz et al., 2018). ONS uses this approach in several interviewer-administered surveys, including the Labour Force Survey, a telephone-administered survey, the largest household study in the UK which provides the official measures of employment and unemployment (Office for National Statistics, 2024).

More recently researchers have trialled use of LLMs during data collection to code occupations during the interview. LLMs can request additional details until the details provided by the respondent is deemed sufficient to allow accurate coding. ONS are currently trialling a generative AI-based prototype to obtain more information from respondents, leading to improved codability of industries and occupations obtained by open text responses. The prototype, called Survey Assist, uses LLMs to create questions designed to obtain further information from respondents, and makes background recommendations for the most likely codes (Tyrell & Arkell, 2026). Another recent study in the UK (Sturgis et al., 2025) follows a similar approach but the tool (SOCbot) dynamically adapts the number of follow-up questions, stopping when the information provided is judged sufficient which can potentially reduce respondent burden. Agreement between LLM-based coding and office coding is relatively modest at around 60% at the 4-digit level (Sturgis et al., 2025).

1.2.3 Closed-list questions

For many surveys not primarily focused on the labour market or economic analyses, highly detailed industry and occupation codes may be unnecessary, as more aggregated measures can suffice for their analytical requirements. In these cases, closed-list questions have been used as an alternative to, or in combination with, open-text questions, especially in online self-administration (Kocar et al., 2023).

Closed-list questions present respondents with a fixed list of industries (or occupations) to choose from. While the aggregated labels for industry and occupation groups on these lists can be difficult for respondents to interpret, descriptions and examples can help reduce ambiguity. While responses to closed-list questions can only collect information at the highest-level categories of industry and occupation, this question format can help reduce administration costs, response burden, and implementation difficulties, compared to collecting open-text responses for manual coding.

Methodological assessments of closed-list methods for occupation coding in self-administered surveys are scarce in the literature, with most examples of its implementation in postal surveys coming from the 1960s and 70s. The postal survey reported in Winch et al. (1968) asked a sample of US college students to match their father's occupation with one of 323 occupations from a list. They were also asked to provide an open-text response, which was independently coded. The authors calculated Duncan's Socio-Economic Indices, a widely used measure for occupational ranking (Duncan, 1961), based on both codes. The indices showed correlation of 0.85, leading the authors to conclude that little information was lost when respondents selected their father's professional category themselves, compared to office coding of open-text responses. However, including such a lengthy list of occupations would not be feasible for a typical general population survey.

Another postal survey reported by Eckhardt and Wenger (1975) asked a sample of US respondents about their own occupation and that of their partner. They used an occupational classification scheme consisting of nine levels and provided four or five specific examples for each listed category. Open-text responses were also obtained and coded separately. The results showed high levels of agreement between office and respondent coding (85% to 88%), although lower agreement rates were observed for unskilled occupations (compared to skilled occupations) and for coding a spouse's occupation (compared to coding the respondent's occupation). The authors also found that respondents whose occupations were classified as unskilled, semi-skilled, or skilled by office coders tended to place themselves in higher-level occupational categories when coding their own responses.

Taylor (1976), using data from a methodological experiment embedded in a nationally representative postal survey, asked participants to self-code into one of the ten standard US Census occupation categories. Participants also described the type of work they did in an open-text question, which was later office coded. Agreement rates between codes ranged from 55% to 61% for an exact category match, and from 70% to 83% for "approximate" matches (i.e., respondents self-coded into the same category assigned by office coders or to an adjacent one). Some systematic errors were found for self-coded occupation categories, with "operatives" (semi-skilled workers) frequently classifying themselves as "labour" (unskilled manual workers). Women and respondents with higher levels of education tended to code themselves more accurately.

The Social Inequality module of the 1987 International Social Survey Programme (ISSP Research Group, 1989) featured closed-list questions with nine categories which asked about the occupation of the respondent and their father alongside open-text questions. Using structural equation modelling with latent variables Ganzeboom (2005) reported that models using office codes produced smaller random measurement errors than models using closed-list codes from the list, but that the difference was small and supported the use of self-codes for this purpose. De Vries and Ganzeboom (2008) report similar results from analyses of four social surveys in the Netherlands.

The general argument for using closed-list questions for industry and occupation coding in self-administered surveys is that the loss of coding accuracy may be compensated by a substantial reduction in processing times and costs, especially for non-specialist surveys. Closed-list questions for occupation have been used in several self-administered surveys worldwide, including UK surveys such as the NatCen Panel (Butt et al., 2025), the Community Life Survey (Kantar Public & Department for Digital Culture Media & Sport, 2025), the DCMS Participation Survey (Kantar Public & Department for Digital Culture Media & Sport, 2019), and the Generations and Gender Survey Survey (Howe et al., 2024). However, methodological research on the quality of resulting data is limited, and the potential for mode or device effects is under-explored. Furthermore, surveys employing closed-list approaches frequently use occupation classifications that are incompatible with global classification schemes such as SOC, making direct comparisons difficult. Comparing empirical results from closed-list questions for industry and occupation coding with office codes obtained via conventional methods is therefore essential to understand the feasibility of the closed-list approach and its potential effects on accuracy and reliability.

1.3 Aims and objectives

This paper evaluates a closed-list approach for industry and occupation coding using data from a UK-based probability panel. The panel supplemented its standard open-text questions on industry and occupation, which are manually coded by office coders, with corresponding closed-list questions. In contrast to previous applications in the literature, the lists of industries and occupations presented to respondents are compatible with higher-level SIC and SOC categories. This compatibility enables a direct comparison between the two coding methods.

The main aim of the paper is to assess the level of agreement between the industries and occupations selected by respondents from closed lists and the codes assigned by specialist office coders, and to explore the factors that may help explain any differences observed. A further objective is to examine whether using different coding methods leads to differences in the resulting distributions of industries and occupations, and hence to potential analytical implications. We also investigate whether the closed-list method is effective in reducing question response time relative to the traditional open-text question approach. Finally, we consider whether respondents completing the survey using a mobile device differ systematically to those using other devices, and whether these differences are associated with agreement rates between closed-list codes and office-coded industries and occupations.

If successful, a self-coding closed-list method implemented in a device agnostic online survey could reduce implementation and processing costs, as well as question response time and respondent burden, while still providing reliable high-level data. Understanding the factors underlying mismatches between self-coded and office-coded responses is important, as it can inform improvements to the design and use of closed-list questions which could help to enhance data quality and reliability in future applications.

1.4 Research questions

To address the overreaching aim of the study, we formulate the following research questions:

- ***RQ1: Are closed-list questions associated with reduced question response time and complexity, compared to open-text questions for industry and occupation? Does this vary by device type?***

Open text questions typically require more response effort from respondents than closed-list questions, even when the options on the list are complex, as the cognitive effort of recognition (scanning and choosing from provided list) is generally perceived as lower than the effort of recall and production (generating a coherent written description from memory) (Coppens et al., 2020). Consequently, we formulate the following hypothesis:

H1.1: The closed-list questions reduced question response time, compared to the open-text questions on industry and occupation.

Participants responding to online surveys via mobile devices frequently provide open-text responses that are shorter (fewer characters) and less complex (shorter words and sentences) than those given by participants using PCs, laptops, or tablets (Lambert & Miller, 2014; Mavletova, 2013; Soszynski, 2025; Struminskaya et al., 2015). The intuitive explanation is that typing on a small touchscreen keyboard requires greater effort than typing on a physical keyboard, making respondents less inclined to elaborate their responses. Based on this evidence, we formulate the following hypotheses:

H1.2: Participants using mobile devices will provide shorter and less complex responses to the open-text questions on industry and occupation, compared with participants using other devices (PCs, laptops, tablets).

- ***RQ2: Does the closed-list approach generate similar distributions of industry and occupations as the office-coded open-text responses?***

Previous research (De Vries & Ganzeboom, 2008; Ganzeboom, 2005) suggests that switching between self-selected and office-coded occupation codes has a relatively minor impact on substantive analyses. However, there is evidence indicating that some differences exist (Eckhardt & Wenger, 1975; ISSP Research Group, 1989; Taylor, 1976). Consequently, we formulate the following hypothesis:

H2.1: The distributions of industries and occupations will differ by coding method.

For industries, the Office for National Statistics (ONS) acknowledges that the SIC 2007 system of classification predates many of today's major economic sectors and does not adequately account for organisations operating across multiple traditional industry boundaries (Mitton, 2025). As a result, industries that span sectors such as professional, administrative, and support services face an increased risk of misclassification, since respondents may be unable to associate their organisation's activities with a single category when selecting from a closed list. An ONS quality analysis confirmed this potential for mismatch (Office for National Statistics, 2023a). Among industries with a sufficient number of observations, the highest mismatch rates between SIC codes derived from business survey responses and those held by the UK's tax office (HM Revenue and Customs) were found in Sections S ("Other service activities") and N ("Administrative and support service activities"). This suggests that these categories may be inherently ambiguous for respondents. We therefore propose the following hypothesis.

H2.2: Participants will be more likely to self-select into professional, administrative, and support service industries than to be classified into these industries through office coding.

Given the mixed empirical evidence regarding self-selection of <u>occupations</u> and its relation to the skill levels required for each, two possible predictions are plausible. On one hand, participants may self-select into occupations associated with lower skill levels than those to which they were assigned by office coding, as in Taylor (1976). On the other hand, participants may self-select into higher skill levels, as reported in Eckhardt and Wenger (1975). Accordingly, we test the directional nature of this effect without a priori commitment to a single outcome. We formulate our hypothesis as follows:

H2.3: Participants will self-select into occupations associated with different skill levels than those to which they are assigned by office coding.

- ***RQ3: What are the overall agreement rates between the closed-list method and office coding for industry and occupation? Do these rates vary by industry or occupation? Do they vary by device type?***

Based on earlier evidence (Eckhardt & Wenger, 1975; ISSP Research Group, 1989; Taylor, 1976), we anticipate some level of disagreement between codes selected from the closed-list question and those assigned by office coders, for both industry and occupation. Although early self-administered paper-based surveys reported relatively high agreement rates (between 70% and 88%), these studies date from the 1970s, when the labour market was arguably less complex and more readily classifiable than it is today. We therefore expect lower agreement rates under online administration, for both industries and occupations. We propose the following hypothesis:

H3.1: Agreement rates between list-based and office-coded codes will be below 70% for both industry and occupation.

For <u>occupations</u> specifically, and drawing on Taylor (1976), we expect that respondents who fail to select the category from the list that corresponds to their office code will often choose categories that appear similar. More precisely, we anticipate that a high proportion of respondents with mismatched occupation codes at the 2-digit level will nonetheless have matching codes at the 1-digit level. Our hypothesis is:

H3.2: Among participants with mismatched occupations at the 2-digit SOC level, a substantial proportion will select categories within the same 1-digit SOC level.

- ***RQ4: Do sociodemographic characteristics (sex, age, educational qualifications) and survey paradata related to text length and complexity, influence the agreement rates between closed-list and office coded industries and occupations? Does this vary by device type?***

Following the limited evidence available (Eckhardt & Wenger, 1975; Taylor, 1976), we expect that the agreement rates between self-coding and office codes will be higher for women and participants with higher educational attainment than for men and those with lower levels of education. We therefore formulate the following hypothesis:

H4.1: Agreement between list-based and office-based codes will be higher for women and participants with higher educational qualifications, compared to men and participants with lower or no educational qualifications.

We also expect that longer open-text responses will be associated with higher disagreement rates between closed-list codes and office codes based on open-text responses. Conrad et al. (2016) analysed double-coded occupations from open-text responses and concluded that disagreement rates between coders increase with the length of the occupation description. They hypothesise that providing more words may offer more opportunities for coders to disagree, unless the descriptions closely align with the official definition of the job category. In addition, longer job descriptions are likely to be an indicator of the complexity of the occupation itself, which may in turn be associated with lower agreement rates between coders. Massing et al. (2019) and Helppie-McFall and Sonnega (2018) support this finding, as additional information (whether in the form of longer descriptions or through supplementary questions or probes) was associated with lower levels of coder agreement in these studies. Following the same logic, the complexity of the open-text response (typically measured by indices of sentence readability and word length, such as the SMOG Index, discussed in Section 2.3) is also likely associated with lower agreement rates. We thus formulate the following hypotheses:

H4.2: Longer and more complex open-text responses will be associated with lower agreement between list-based and office codes, compared to shorter and simpler open-text responses.

Following on from the previous hypotheses *H4.1* and *H1.2*, we further propose that agreement rates between self-selected and office-assigned codes will be higher for respondents using a smartphone than for those using other devices, for both industry and occupation. This follows from the observation (discussed above) that mobile respondents tend to provide shorter responses to open-text questions, and that shorter responses are associated with lower coding agreement.

This result may initially appear counterintuitive, given evidence of lower data quality for mobile respondents compared with those responding via other devices, in terms of indicators like item non-response, careless responding, and shorter open-text responses (Peytchev & Hill, 2009; Struminskaya et al., 2015). However, recent evidence suggests that device effects on data quality are weaker than previously thought (Antoun et al., 2017; Clement et al., 2020; Décieux & Sischka, 2024; Sommer et al., 2016). Furthermore, we are unable to analyse industry and occupation data quality – not least because there is no "true value" against which to compare the codes or list selections. Instead, our focus is on the relative ease or difficulty of assigning respondents to a code based on the information they provide. As mentioned above, longer responses are not necessarily more helpful for accurate coding; rather, they can be difficult to interpret within the framework of an industrial or occupational classification and can make it difficult to decide on a unique code (Massing et al., 2019).

Considering this evidence, we would expect device differences, both in terms of the specificity of office-based codes (an indicator of coding specificity), and in terms of agreement rates between closed-list and office-based codes. We therefore formulate the following hypotheses:

H4.3: Agreement rates for participants using mobile devices will be higher than for participants using other devices.

H4.4: The proportion of inaccurate SIC and SOC codes obtained from open-text responses, as measured by the proportion of "not elsewhere classified" codes, or "less specific codes", will be lower for mobile respondents compared to respondents using other devices.

2 Data

2.1 Survey description

This paper uses data from a survey conducted as part of the NatCen Panel, a probability-based panel of adults in the United Kingdom. This survey employed two distinct approaches for collecting data on industry and occupation: open-ended questions and closed-list questions. The open-ended questions followed standard practice. A single open-ended question was used for industry: "*What does the firm or organisation you work for mainly make or do at the place where you work? Please give a full description. For example, do you work in 'manufacturing', 'processing' or 'distributing'? What are the main goods you produce and materials you use? Is it wholesale or retail?"*. For occupation, two open-ended questions were used, one for job title ("*What is the name of your job or your job title?*") and one for job description ("*In your own words, what do you mainly do in your job? Please give a full description*"). For the industry and job description questions, the questionnaire also included a soft check prompting participants to provide more information when the response entered had fewer than 15 characters. The soft check prompt read "*it is important that you enter enough information for us to clearly understand [where you work] OR [your job]. Please ensure you have fully described what you do*". Professional coders manually coded the responses to these questions, assisted by the semi-automated coding software CASCOT (Jones & Elias, 2004). All three open-text responses available (industry, job title, and job description) were used by coders to code industry and occupation. The office coding process included quality controls. For example, supervisors blindly coded a random 10% of cases and then compared the results, intervening if a particular coder was experiencing a high number of discrepancies. Similarly, any open-text responses that were impossible for a coder to code were escalated to a supervisor for coding.

Additionally, participants were presented with closed lists of industries and occupations to select from. For industry, they were asked, "*Which of the following best describes the sector you work in?*" with a list of 21 options. These options correspond to the 1-digit sections of the SIC 2007 classification, as listed in Table 1. Respondents were offered the choice to read additional information about each option when they clicked on a help link. For example, for Section A (Agriculture, forestry and fishing), the additional text read: "*Typically involves working on farms, in forests or on fishing boats, and is often based outdoors on land. Examples include crop farming, livestock farming, logging, silviculture, and marine and freshwater fishing.*" The full text for the information associated with each option is provided in Appendix A.

Table 1. List of options for the closed-list industry question (1-digit SIC 2007 sections)

Section	Industry
A	Agriculture, forestry and fishing
B	Mining and quarrying
C	Manufacturing
D	Electricity, gas, steam & air conditioning
E	Water supply, sewerage and waste management
F	Construction
G	Trade (retail sales and wholesales) and vehicle repair
H	Transportation and storage
I	Accommodation and food service
J	Information and communication
K	Finance and insurance
L	Real estate
M	Professional, scientific and technical
N	Administrative and support services
O	Public administration and defence
P	Education
Q	Human health and social work
R	Arts, entertainment and recreation
S	Other service activities
T	Employed by a household as domestic staff
U	International organisations and bodies

For occupation, participants were asked, "*Which of the following best describes the main paid job you had in the seven days preceding Sunday?*" with a list of 26 options. These correspond to the 2-digit *sub-major groups* of SOC 2020, listed in Table 2. Similar to the industry closed-list question, respondents were offered the choice to read additional information about each option in the list. For example, for category 11 ("Corporate managers and directors"), the additional text read "*such as Chief Executive, Production manager, Financial manager, marketing manager, officers in the armed forces, senior police officers, health, public health and social care managers".* The full text for the information associated with each option is provided in Appendix A.

Within the SOC system of occupation classification, *sub-major groups* (2-digit codes) can be grouped into *major groups* (1-digit level), which are a set of broad occupational categories designed to be useful in bringing together groups of occupations which are similar in qualifications, training, skills and experience commonly associated with the competent performance of work tasks (Office for National Statistics, 2023b). Sub-major groups can also be associated with a *skill level*, which in turn is usually approximated by "*the time taken to gain necessary formal qualifications or the required amount of work-based training*". Skill levels required for each sub-major group are listed in the last column of Table 2. Jobs belonging to major groups 1 and 2 normally require a degree or equivalent period of relevant work experience. Occupations in groups 3 and 5 require a period of post-compulsory education but not normally to degree level, while occupations in groups 4, 6, 7, and 8 require completion of compulsory education and a longer period of work-related training or work experience, while elementary occupations (group 9) only require compulsory education.

The questions were asked in the following order: 1) Industry (closed list) 2) Industry (open text) 3) Job title (open text) 4) Occupation (closed list) 5) Job description (open text).

Table 2. List of options for the closed-list occupation question (1-digit SOC 2020 codes)

Major group		Sub-major group		Skill
Code	**Definition**	**Code**	**Definition**	**level**
1	Managers, directors and senior officials	11	Corporate managers and directors	4
		12	Other managers and proprietors	3
2	Professional occupations	21	Science, research, engineering and technology professionals	4
		22	Health professionals	4
		23	Teaching and other educational professionals	4
		24	Business, media and public service professionals	4
3	Associate professional occupations	31	Science, engineering and technology associate professionals	3
		32	Health and social care associate professionals	3
		33	Protective service occupations	3
		34	Culture, media and sports occupations	3
		35	Business and public service associate professionals	3
4	Administrative and secretarial occupations	41	Administrative occupations	2
		42	Secretarial and related occupations	2
5	Skilled trades occupations	51	Skilled agricultural and related trades	3
		52	Skilled metal, electrical and electronic trades	3
		53	Skilled construction and building trades	3
		54	Textiles, printing and other skilled trades	3
6	Caring, leisure and other service occupations	61	Caring personal service occupations	2
		62	Leisure, travel and related personal service occupations	2
		63	Community and civil enforcement occupations	2
7	Sales and customer service occupations	71	Sales occupations	2
		72	Customer service occupations	2
8	Process, plant and machine operatives	81	Process, plant and machine operatives	2
		82	Transport and mobile machine drivers and operatives	2
9	Elementary occupations	91	Elementary trades and related occupations	1
		92	Elementary administration and service occupations	1

**The options provided in the closed-list question were the sub-major groups. Major groups are provided for context. The last column in the table provides the skill level required for each sub-major group, according to the SOC classification. Occupations at level 1 require completion of compulsory education. Level 2 requires the same level of education as in Level 1, but with a longer period of work-related training or work experience. Level 3 require a period of post-compulsory education but not normally to degree level, while Level 4 occupations require a degree.*

2.2 Dataset description

Our analysis is based on data collected during the March 2022 wave of the NatCen panel survey. The NatCen panel is a probability-based survey comprising participants from Great Britain and Northern Ireland. In Great Britain, panel members are recruited from the British Social Attitudes survey (BSA), a high-quality random probability survey of individuals aged 18 or over. Between 2015 and 2019, the BSA was conducted face-to-face, with participants within each household selected at random. Since 2020, however, it has used a "push-to-web" approach, with up to two adults in a household invited to take part and encouraged to complete the survey online, with an opt-in telephone option available. In Northern Ireland, panel members were recruited from the Consumer Detriment Survey (CDS), which interviewed individuals aged 18 or over in 2021. Like the post-2020 BSA, the CDS is a random probability survey that uses a "push-to-web" design, with up to two adults per household allowed to participate and encouraged to complete the survey online, with an opt-in telephone mode available.

Respondents interviewed in the BSA and CDS are invited to join the NatCen panel at the end of their interview. For the March 2022 wave of the NatCen panel, all panel members aged 18 to 66 recruited from BSA (from 2015 onwards) and CDS, excluding those who had left the panel or become inactive, and a random sub-sample from BSA 2021, were issued ($n =$

$5{,}884$). Participants who had not been in any paid employment since January 2020 were also screened out.

Fieldwork for the March 2022 wave of the panel followed a sequential mixed mode design. Panel members were initially invited to participate online and received multiple reminders by post, email, and/or text message. Thos who had not responded after two weeks, and for whom telephone numbers were available, were contacted and encouraged to complete the survey online or offered the option of a telephone interview. Conditional on completion, participants received a £10 e-voucher as a token of appreciation.

As we were interested in industry and occupation codes from online respondents, we excluded the small fraction of interviews ($n = 187$) that were conducted by telephone. Similarly, we excluded all respondents who reported not being in employment at the time of the interview. Our initial sample contains a total of 4,776 observations. 99.7% of participants in the initial sample selected one of the options in the closed-list industry question, while 98.9% selected an occupation. Similarly, 99.8% of respondents provided open-text responses to the industry questions, while over 99.0% provided an open-text response to the occupation questions (job description and job title). Coding rates were similarly high, albeit slightly higher for industry (99.3%) than for occupation (98.6%).

We constructed two separate analytical samples - one for SIC and one for SOC - to account for differences in coding rates. Each sample includes only observations with both a closed-list selection and a corresponding office-coded value for the relevant domain. This approach allows cases with valid industry data but missing occupation data (and vice versa) to be retained in the analysis. In addition and given the relatively low levels of item non-response in both surveys, our analytical samples excluded observations with missing values in key sociodemographic variables, including sex, age group, UK region, urban/rural classification, highest educational qualification, ethnicity, economic activity, and household size.

Table 3 presents the sample sizes of our two analytical samples. The initial sample includes all online respondents who were working on the interview date, and the analytical sample excludes observations as described above. The table also provides the percentage of observations excluded due to each exclusion criteria.

Table 3. Sizes of initial and analytical samples and percentage of missing observations

Sample	Sample size		% of missing observations (with respect to the initial sample)		
	Initial	**Analytical**	**Due to missing industry or occupation data**	**Due to missing sociodemographic data**	**Total missing**
SIC	4,776	4,669	1.1	1.1	2.2
SOC		4,589	2.7	1.2	3.9

Our analytical samples retain 97.8% of the initial SIC sample and 96.1% of the initial SOC sample. A descriptive analysis of sociodemographic variables for both samples is provided in Appendix B.

We also make use of paradata. Specifically, we have information on the device used to complete the survey (mobile or other), as well as navigation data related to each of the industry and occupation questions. This includes the elapsed question response time, the number of error messages (if any), and the number of timeouts due to prolonged inactivity (if applicable). In addition, we have access to the full open-text responses for all industry and occupation questions.

2.3 Methods

To compare *question response time and complexity* (**RQ1**) between the two approaches, we analysed three types of indicators: question response times, number of errors and timeouts, text length and complexity.

We started by comparing question response times for the closed-list and open-text questions, for both industry and occupation. For occupation, the time each respondent spent on the open-text question was calculated as the sum of the time taken to write the open-text response to the job title and the job description questions. We also compare response times for all questions across devices.

We then compared the total number of error messages received by the respondent, and the number of timeouts (associated with prolonged inactivity in the response). We computed paired-samples t-tests to determine whether the question response time and complexity indicators differed between the two question types. We also compared the total number of error messages and timeouts across devices.

Finally, we processed the full text entered for each of the three open-text questions (industry description, job title and job description) using the *quanteda* package in R (Benoit et al., 2018) to generate length and complexity indicators for each response. To measure response length, we recorded the number of characters, syllables and words in each text response. Open-text response complexity was assessed the SMOG Index (McLaughlin, 1969), one of the most widely used text complexity indicators of text readability and clarity. The SMOG Index is calculated by counting the number of polysyllabic words (those with three or more syllables) in a sentence. A higher SMOG score indicates more complex vocabulary and denser text, suggesting it is suitable for readers with more advanced education. Specifically, a SMOG score of 8 implies that the text can be understood by someone with approximately an eighth-grade reading level in the US, while a score of 12 or higher indicates material appropriate for readers with higher educational qualifications.

We compared text length and complexity across all three open-text questions (industry description, job title, job description) and calculated paired-means t-tests to compare the results across devices.

To assess potential differences in the *distributions of industries and occupation* (**RQ2**), we first computed the frequencies of selected and coded categories using a χ^2 test. Additionally, we calculated the mean dissimilarity index D (Duncan & Duncan, 1955) for both industry and occupation. This index was calculated as:

$$D = \frac{1}{2}\sum_{i=1}^{n}|s_i - c_i|, \tag{1}$$

where s_i is the proportion of respondents who selected category i, c_i is the proportion of respondents who were coded into category i, and n is the total number of (industry or occupation) categories. This index measures the unevenness between the two distributions and represents the percentage of one group that would need to move to another category to make the two distributions equal. We identified "problematic" categories of industry and occupation as those with the highest contribution to this dissimilarity index.

We analysed *agreement rates* (**RQ3)** between selected and coded industries by defining a binary variable coded as 1 when the selected category (SIC section) matched the coded category, and 0 otherwise. For occupations, we defined two similar variables: one for the major group (1-digit SOC code) and one for sub-major groups (2-digit SOC codes). We then computed the mean values of agreement rates across all categories of industry and occupation and identified the categories with the highest and lowest agreement levels. Furthermore, we built two-way cross-classification tables. By examining the off-diagonal cells of these tables, we identified the most frequent error patterns between selected and coded industries and occupations for participants where a discrepancy was observed.

Finally, to identify the *factors associated with agreement rates* **(RQ4)**, we estimated two logistic regression models (one for industry and one for occupation, both at the 1-digit level), in which the dependent variable was agreement as defined above. The models controlled for the sociodemographic variables available in the dataset (sex, age, household size, ethnicity, educational qualifications, age category, UK region, and urban or rural location) as well as the coded categories for industry and occupation. More importantly, the models also accounted for paradata, including the device used to complete the survey, the time taken to respond to the list question, and the length and complexity of the open-text responses. For the open-text questions, the indicators of text length, complexity, and elapsed time were highly correlated. In light of this, we tested various model configurations to determine the indicator that provided a better model fit. The final version of each model included a single indicator for open-text responses (one of character length, SMOG index or elapsed time), along with the time elapsed when responding to the list question and the device used for responding to the survey, as covariates.

For coding specificity, we focused on descriptions of occupations that may be too abstract to be assigned to a more specific category. Generally, SIC or SOC codes identified as "not elsewhere classified" (n.e.c.) contain a mix of industries or occupations which are not in sufficient numbers to merit their own group. For SOC, these are usually identified as 4-digit codes ending in '9'. Although there is no similar rule for SIC, industries "not elsewhere classified" can be identified as such in the written description of each code. Similarly, some SOC codes at the lower level of disaggregation may be equivalent to the broader level, indicating that the coding could not be achieved at a more detailed level. These are usually identified by SOC codes ending in '0'. For example, unit group (4-digit) SOC code 3560 is no more specific than its corresponding minor group (3-digit) SOC code 356, as they both refer to "public services associate professionals". We identified these as "less specific codes". It must be noted that we cannot construct a similar indicator for SIC, since 3-digit classifications (industry "divisions") are the most disaggregated codes available for some activities (e.g. 85100 for pre-primary education, and 85200 for primary education) and no more detailed codes are provided in the framework. Thus, using a higher-level SIC division does not necessarily imply inaccurate coding. We compute the proportions of "not elsewhere classified" and "less specific" codes and compare them across devices.

3 Results

3.1 Question response time and complexity (RQ1)

Table 4 summarises the three types of indicators of question response time and complexity (completion times, number of errors and timeouts, and text length and complexity) across question types and devices for the industry and occupation questions in the survey.

Table 4. Summary of indicators of question response time and complexity across question types and devices

Category	Indicator	Open-text questions				Closed-list questions				p-value (question)
		Mobile	Others	p-value (device)	Full sample	Mobile	Others	p-value (device)	Full sample	
Industry										
Response time	Elapsed time (seconds)	51.2	60.3	0.053	**54.5**	41.8	46.1	0.464	**43.4**	<0.001-
Errors and Timeouts	Error messages (n)	0.27	0.19	<0.001	**0.24**	0.003	0.004	0.724	**0.003**	<0.001
	Timeouts (n)	0.04	0.06	0.007	**0.045**	0.03	0.05	0.007	**0.035**	0.003
Text length and complexity	Number of characters	43.5	50.2	<0.001	**45.9**	-	-	-	-	-
	Number of syllables	12.4	14.4	<0.001	**12.9**	-	-	-	-	-
	Number of words	6.4	7.1	<0.001	**6.6**	-	-	-	-	-
	SMOG Index	9.4	10.3	<0.001	**9.7**	-	-	-	-	-
										Occupation
Response time	Elapsed time (seconds)	71.2	82.9	0.007	**75.5**	40.9	39.7	0.727	**40.5**	<0.001
Errors and Timeouts	Error messages (n)	0.08	0.06	0.004	**0.075**	0.006	0.01	0.156	**0.007**	0.021
	Timeouts (n)	0.09	0.15	<0.001	**0.108**	0.04	0.07	0.003	**0.06**	0.037
Text length and complexity (Job title)	Number of characters	19.4	20.5	0.002	**19.8**	-	-	-	-	-
	Number of syllables	6.1	6.4	0.035	**6.2**	-	-	-	-	-
	Number of words	2.5	2.6	0.097	**2.5**	-	-	-	-	-
	SMOG Index	8.6	9.0	<0.001	**8,8**	-	-	-	-	-
Text length and complexity (Job description)	Number of characters	68.8	78.1	<0.001	**72.2**	-	-	-	-	-
	Number of syllables	18.8	21.7	0.021	**19.9**	-	-	-	-	-
	Number of words	10.2	11.5	0.037	**10.7**	-	-	-	-	-
	SMOG Index	10.5	11.7	<0.001	**11.0**	-	-	-	-	-

Note: *The table presents the mean values of each indicator, broken down by question format (open-text versus closed-list) and by device type (mobile versus other devices including PC, laptop, and tablet). It also presents the mean values of each indicator for the full sample. The "p-value (device)" column provides the p-values from unpaired t-tests assessing whether the mean values differ between respondents using mobile devices and those using other devices. The "p-value (question)" column provides the p-values from unpaired t-tests assessing whether the mean values for the full sample differ between open-text questions and closed-list questions. The latter comparison is only available for completion time and errors and timeouts indicators.*

As expected, the greater complexity of responding to an open-text question, compared with a closed-list format, resulted in participants spending, on average, 11 seconds more time on the open-text industry question than on the closed-list version. The difference is significant. The number of error messages and timeouts was also higher for the open-text industry question compared with the closed-list question. All the differences are statistically significant at the 95% confidence level. We observe similar results for occupation. The two open-text questions take, on average, 35 seconds longer to answer than the closed-list occupation question. The number of error messages and timeouts is also significantly higher for the open-text occupation questions compared with the closed-list versions. All these differences are statistically significant at the 95% confidence level. These results confirm that closed-list questions on industry and occupation are successful in reducing question response time compared with traditional open-text questions, supporting *H1.1*.

In the SIC sample, 60.3% of respondents completed the survey using a mobile phone, with the remainder using other devices such as tablets, desktop computers, or laptops. Mobile respondents spent 9.1 fewer seconds answering the open-text industry question compared with those using other devices, although this difference is non-significant. Mobile users also experienced fewer timeouts but received significantly more error messages than respondents on other devices. Both differences were significant at 95% level. Device-related differences were less pronounced for the closed-list industry question, with elapsed times not differing across devices. For the open-text industry question, mobile respondents provided notably shorter answers (6.7 fewer characters on average) and significantly less complex responses (as measured by the SMOG index) than those using other devices. These findings collectively support *H1.2*.

Similar patterns emerge in the SOC sample, where 63.6% of respondents used a mobile phone. Mobile respondents spent less time answering the open-text occupation questions (11.7 seconds less on average), experienced fewer timeouts, and received more error messages than those using other devices. The number of timeouts was significantly lower for mobile respondents in the closed-list format. Mobile respondents supplied significantly shorter job titles and job descriptions in terms of characters and syllables. Although the difference in job title word length between mobile and other devices is non-significant, job descriptions from mobile users contain significantly fewer characters, syllables, and words. Moreover, open-text responses for both job titles and descriptions are significantly less complex for mobile respondents. These findings are also consistent with *H1.2*.

3.2 Distributions of industries and occupation (RQ2)

Table 5 summarises the agreement and specificity rates between respondent-selected and office-coded categories, both for industry and occupation, disaggregated by device type, and additionally presented for the full sample. Agreement rates for industry are significantly higher for mobile respondents compared to those using other devices. There is no significant difference in agreement rates for occupation between mobile respondents and those using other devices.

Table 5. Coding specificity and agreement rates for industry and occupation

Indicator	Dimension	Mobile	Others	p-value	Full sample
Coding agreement (%)	Industry	63.5	58.4	<0.001	**61.7**
	Occupation (Major group)	56.7	54.2	0.105	**55.8**
	Occupation (Sub-major group)	46.7	44.3	0.105	**45.8**
Coding specificity – Less specific codes (%)	Occupation	4.0	4.4	0.596	**4.2**
Coding specificity – Not elsewhere classified (%)	Industry	5.3	5.4	0.540	**5.5**
	Occupation	12.4	13.8	0.170	**12.9**

****Note:** The table provides the p-values from unpaired t-tests assessing whether the mean values (percentages) differ between respondents using mobile devices and those using other devices.*

Table 6 presents the distribution of selected and coded industries, along with the agreement rates, and the sample size for each industry. As expected, the distribution differs significantly according to the coding method ($\chi^2 = 376.1, p < 0.001$). Consistent with *H2.1*, the method of collecting industries has a direct impact on the distribution of industries in both samples.

Table 6. Distribution of coded and selected industries

Code	Description	Selected (%)	Coded (%)	Agreement rate (%)	N (coded)
A	Agriculture, forestry and fishing	0.7	0.3	80.0	15
B	Mining and quarrying	0.2	0.1	75.0	4
C	Manufacturing	5.7	9.3	51.3	433
D	Electricity, gas, steam & air conditioning	1.2	0.6	80.0	25
E	Water supply, sewerage and waste management	0.7	0.8	72.2	36
F	Construction	3.2	3.6	58.7	167
G	Trade (retail sales and wholesales) and vehicle repair	7.0	9.8	57.1	461
H	Transportation and storage	3.6	3.3	62.8	156
I	Accommodation and food service activities	2.8	2.8	68.5	127
J	Information and communication	4.6	5.3	48.8	248
K	Finance and insurance	5.9	4.4	86.5	207
L	Real estate	0.8	1.3	41.0	61
M	Professional, scientific and technical	10.9	8.2	44.0	382
N	Administrative and support services	7.4	3.0	16.2	142
O	Public administration and defence	6.6	9.2	53.8	433
P	Education	14.0	14.6	85.0	687
Q	Human health and social work	14.9	18.4	71.1	855
R	Arts, entertainment and recreation	2.6	2.6	57.1	119
S	Other service activities	6.8	2.0	38.8	93
T	Employed by a household as domestic staff	0.4	0.3	33.3	12
U	International organisations and bodies	0.4	0.1	16.7	6

The mean dissimilarity index across all industries is 15.0%. This indicates that 15.0% of selected responses would need to move to a different category to reproduce the distribution of coded industries. Categories with the highest contribution to this mean dissimilarity index include S ("Other services", 4.8%), C ("Manufacturing", 3.6%), Q ("Human health and social work"), G ("Trade (retail sales and wholesales) and vehicle repair", 2.8%), and M ("Professional, scientific and technical"). Respondents appear to self-code their organisations more frequently into categories M, S, and N ("Administrative and support services") than they are office coded into those categories. Categories M, N, and S likely represent activities common across many organisations (e.g., professional, administrative, or service roles). This suggests that some respondents may have been thinking about their own role within their organisation, rather than the primary activity of their employer. Specifically, 60% of mismatches in the industry were office-coded into job categories including administrative occupations; science, research, engineering and technology professionals (and associate professionals); business, media and public service professionals (and associate professionals); corporate managers and directors; and elementary administration and service occupations. Conversely, respondents tended to under-select four categories: C, G, Q, and O ("Public administration and defence"). All these findings are consistent with *H2.2.*

Table 7 presents the distribution of selected and coded occupations at the major group (1-digit) level, along with their agreement rates and the corresponding sample sizes for each category. As expected, the distributions differ significantly by coding method ($\chi^2 = 250.4$, $p < 0.001$). Similar to the patterns observed for industries, the method in which occupations are collected has a direct impact on the resulting occupational distribution in the sample.

Table 7. Distribution of coded and selected occupations (1-digit level)

Major group	Description	Selected (%)	Coded (%)	Agreement rate (%)	N (coded)
1	Managers, directors, and senior officials	13.9	10.2	56.5	467
2	Professional occupations	37.0	31.7	74.6	1,454
3	Associate professional and technical occupations	10.5	16.2	23.2	745
4	Administrative and secretarial occupations	12.4	13.1	61.2	603
5	Skilled trade occupations	5.4	5.9	57.9	271
6	Caring, leisure and other service occupations	5.7	8.5	38.7	390
7	Sales and customer service occupations	9.5	5.8	77.7	265
8	Process, plant and machine operatives	4.4	4.2	63.9	194
9	Elementary occupations	1.2	4.4	15.5	200

At the one-digit level, the mean dissimilarity index across all occupations is 12.9%. The categories with the highest contribution to the dissimilarity index include major group 1 ('Managers, directors, and senior officials') and major group 2 ('Professional occupations'). Respondents tend to self-code into these groups more frequently than they are office-coded into the same categories. Participants also place themselves in major group 7 ('Sales and customer service occupations') more often than office coders do. Conversely, participants are more frequently office-coded into major groups 3 ('Associate professional and technical occupations'), 6 ('Caring, leisure and other service occupations'), and 9 ('Elementary occupations'). In *H2.3*, we hypothesised that there would be a difference between the skills required for self-selected occupations and coded occupations. The data confirms that participants tend to self-select into occupations with *higher* skill levels than those to which they were assigned by office coders.

Distributions at the 2-digit level are presented in Appendix C. At this level, the distributions of coded and selected occupations are also significantly different ($\chi^2 = 585.1$, $p < 0.001$), and the mean dissimilarity index is 19.2%.

3.3 Agreement rates (RQ3)

As presented in Table 5, agreement rates between respondent-selected and office codes are relatively low at 61.7% for industry, 55.8% for 1-digit level occupation and 45.8% for 2-digit level occupation. All these values are lower than the 70% threshold suggested by the earlier literature, supporting *H3.1.*

As listed in Table 6,agreement rates are highly variable across <u>industries</u>. Agreement rates for industry categories P (“Education”), K (“Finance and insurance”), D (“Electricity, gas, steam and air conditioning”), and A (“Agriculture, forestry and fishing”) scoring exceeded 80% while agreement rates for industry categories, J (“Information and communication”), L (“Real estate”), and M (“Professional, scientific and technical”) were a little under 50%. We observe the lowest agreement rates in category N (“Administrative and support services”). Category U (“International organisations and bodies”) also had a low agreement rate (16.7%), but only 6 respondents were coded into this category. This category includes organisations that carry out rental and leasing activities, employment activities, tour operations, security and investigation activities, services to building and landscaping, and office/business support. These activities are frequently confused by respondents in the list selection with categories M (“Professional, scientific, and technical”) O (“Public administration and defence”), and Q (“Human health and social work”).

A clearer look at these differences arises when focusing on some consistent patterns of disagreement. Cross-classification tables (provided in Appendix D) present the correspondence between self-coded industries, selected from the closed-list, and office-coded industry categories. In these tables, the sum of the elements of the diagonal corresponds to the overall agreement rate. Focusing on the off-diagonal elements, Table 8 lists the most frequently selected categories among the 1,790 respondents with mismatched industry codes, along with the proportion of respondents who selected each category. In addition, this table lists the categories to which participants with mismatched industry codes were most frequently assigned.

Table 8. Most frequently selected and coded industries among respondents with mismatched industry codes

Category	Code	Industry description	% of responses (Among respondents with mismatched industry codes, N = 1,790)
Most frequently selected	M	Professional, scientific and technical	18.9
	N	Administrative and support services	18.1
	S	Other service activities	15.5
Most frequently coded	Q	Human health and social work	14.3
	C	Manufacturing	12.4
	M	Professional, scientific and technical	12.4
	O	Public administration and defence	11.8
	G	Trade (retail sales and wholesales) and vehicle repair	11.3

****Note:** Only industry categories representing over 10% of responses among respondents with mismatched industry codes on either survey are represented in this table.*

52.5% of mismatched respondents selected one of three categories: M (“Professional, scientific and technical”), N (“Administrative and support services”), or S (“Other service activities”). Notably, category M appears among both the most over- and under-selected codes. This could indicate a degree of question misinterpretation, as respondents often assign their organisation to this category incorrectly but also fail to assign it when they should.

Similarly, there is substantial variability in agreement rates across occupation categories, as listed in Table 7. At the 1-digit level, the highest agreement rates are found in major groups 7 (“Sales and customer service occupations”) and 2 (“Professional occupations”), both of which exceed 70%. High levels of agreement are also observed for major groups 8 (“Process, plant and machine operatives”) and 4 (“Administrative and secretarial occupations”), each with agreement rates above 60%. In contrast, the lowest agreement rates are observed in major groups 3 (“Associate professional and technical occupations”) and 9 (“Elementary occupations”). The cross-classification table indicates that approximately 3.0% of participants were office-coded into major group 3 while self-coding into major group 2. Conversely, an additional 5.3% self-coded into major group 3 while being office-coded into major group 2. This suggests that many participants had difficulty distinguishing between these two occupational groups. A similar pattern emerges between major groups 1 and 2, with 5.3% of participants self-coding into one while being office-coded into the other.

Table 9 presents the cross-classification of occupations at the major group (1-digit) level. The corresponding cross-classification table at the sub-major group (2-digit) level is provided in Appendix D. As with industries, these figures are lower than the reference values reported in the literature for comparisons between self-coded and office-coded occupation categories. These results support *H3.1* for both industries and occupations and suggest that the closed-list method does not provide high levels of accuracy for obtaining codes at these levels of aggregation.

Table 9. Cross-classification table of selected and coded occupations (1-digit level) (percentage of responses)

Selected	Major group	Coded								
		1	2	3	4	5	6	7	8	9
1	Managers, directors, and senior officials	5.8	3.3	2.2	1.1	0.5	0.3	0.3	0.2	0.2
2	Professional occupations	2.0	23.6	5.3	1.4	0.7	3.1	0.2	0.3	0.3
3	Associate professional and technical occupations	0.7	3.0	3.8	1.0	0.3	1.1	0.2	0.1	0.2
4	Administrative and secretarial occupations	0.4	1.2	2.1	8.0	0.1	0.1	0.3	0.1	0.2
5	Skilled trade occupations	0.2	0.1	0.5	0.1	3.4	0.2	0.0	0.6	0.2
6	Caring, leisure and other service occupations	0.2	0.2	0.9	0.2	0.2	3.3	0.1	0.0	0.6
7	Sales and customer service occupations	0.8	0.3	1.3	1.0	0.2	0.2	4.5	0.1	1.2
8	Process, plant and machine operatives	0.1	0.0	0.2	0.2	0.4	0.1	0.0	2.7	0.7
9	Elementary occupations	0.0	0.0	0.0	0.2	0.1	0.1	0.0	0.1	0.7

****Note:*** *The sum of the cell values in this table is 100%.*

We examine some common patterns of disagreement at the 1-digit level in Table 10, which lists the most frequently selected occupation categories among the 2,486 respondents with mismatched occupation codes, along with the proportion of respondents who selected each category. The table also reports the categories to which these participants were most frequently office-coded.

Consistent with the findings presented earlier, 50.3% of respondents with mismatched codes selected either category 1 ("Managers, directors and senior officials") or category 2 ("Professional occupations"). Meanwhile, 33.7% of respondents with mismatches were office-coded into these two categories, further indicating that many participants had difficulty distinguishing between them. "Associate professional and technical occupations" also appear frequently among both the selected and office-coded categories for respondents with mismatched 1-digit occupation codes.

Table 10. Most frequently selected and coded industries among respondents with mismatched industry codes

Category	Code	Industry description	% of responses (Among respondents with mismatched major group occupation codes, N = 2,486)
Most frequently selected	2	Professional occupations	31.1
	1	Managers, directors, and senior officials	19.2
	3	Associate professional and technical occupations	14.3
	7	Sales and customer service occupations	11.7
Most frequently coded	3	Associate professional and technical occupations	24.9
	2	Professional occupations	21.3
	1	Managers, directors, and senior officials	12.4
	4	Administrative and secretarial occupations	11.3

****Note:*** *Only 1-digit occupation categories representing over 10% of responses among respondents with mismatched industry codes on either survey are represented in this table.*

It must further be noted that, among respondents whose sub-major group (two-digit) occupation codes were mismatched, only 456 (18.3 per cent) were coded into the same major group (one-digit) occupation category they had selected from the list, and thus most mismatches at the two-digit level occur across different major sub-groups. This indicates that the main sources of ambiguity for respondents arise from the major group definitions, rather than from the more specific sub-major groups, contradicting *H3.2*.

3.4 Factors associated with agreement rates (RQ4)

The full results of the logistic regression models of agreement between selected and coded industries are presented in Appendix E. The marginal plots in Figure 1 illustrate that, even after controlling for sociodemographic, location, and paradata variables, certain industries are more likely to result in mismatches between respondent selected and office coded industry categories. Specifically, respondents whose organisation was office-coded into categories C, J, L, M, N, and "other" industries (S, T, and U) are significantly less likely to have selected a code that matches the office code compared to the base category (industries A and B). These industries correspond to those in which the agreement rates were lower in Table 6. In addition, participants living in rural areas had a higher probability of agreement between selected and coded industries compared to those in urban areas, and participants with educational qualifications below A-levels had a lower probability of agreement compared to those with higher qualifications. While the latter finding partially supports *H4.1*, we did not find any significant differences between men and women.

As anticipated, more complex text responses (as measured by the SMOG index) were generally associated with a lower likelihood of agreement between selected and coded industries, which supports *H4.2*. Similarly, the time elapsed when selecting a category from the list was also negatively associated with agreement. Both these findings indicate that agreement between selected and coded industries is more likely when respondents find it easier to describe the activity conducted in their workplace. For more complex and/or diverse industries, participants will require more effort to select a category from the closed list, and/or to provide an open-text description of it. This will likely result in a lower probability of agreement. On a similar note, mobile respondents have a significantly higher probability of agreement between selected and coded industries compared to those responding via other devices, which supports *H4.3*. This is consistent with the fact that mobile respondents tend to provide shorter responses to the open-text questions. These, in turn, are generally associated with higher agreement rates.

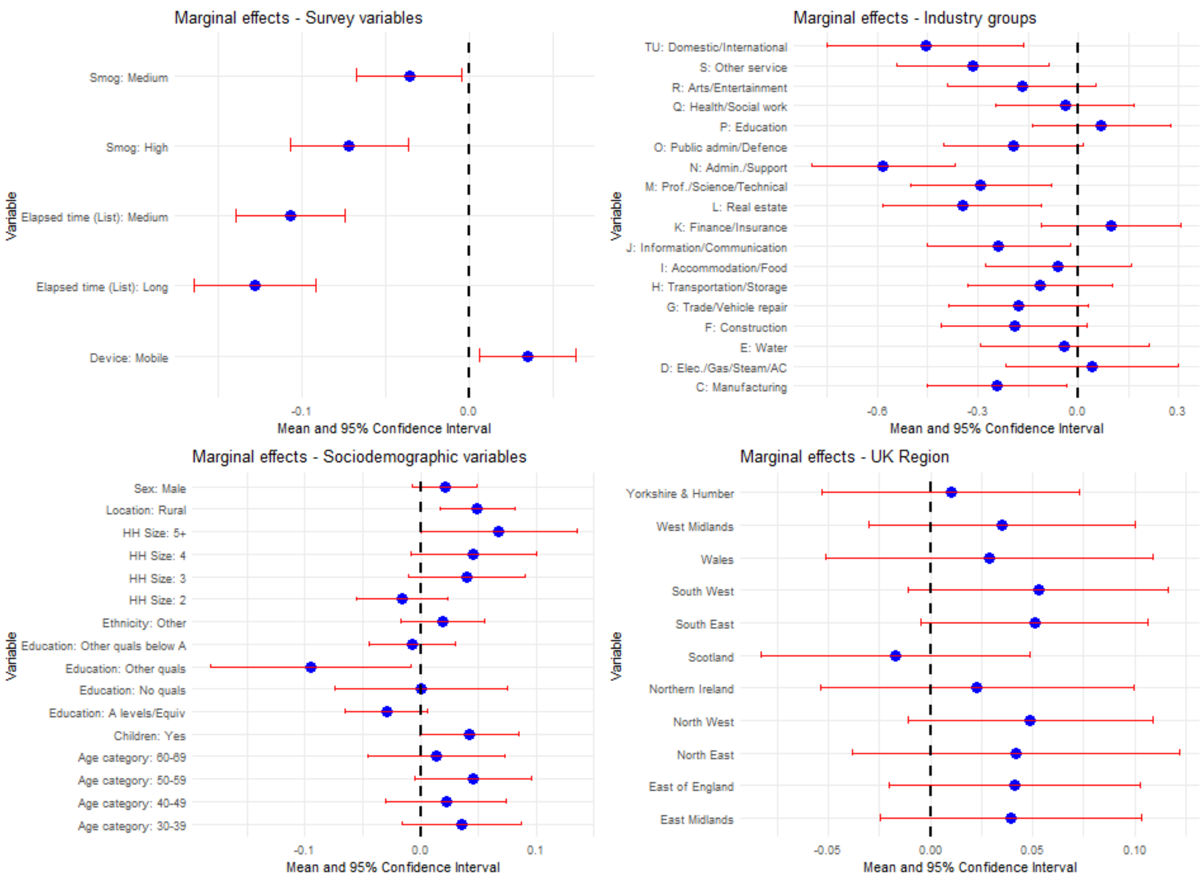

Figure 1. Logistic regression models of agreement between industries

The marginal plots for the logistic regression models of agreement between selected and coded <u>occupations</u> are illustrated in Figure 2, with full results provided in Appendix E. Being office-coded into major groups 3, 6, and 9 reduced the likelihood of a match with the self-selected code (compared to the reference category 1), while being office coded into major groups 2 and 7 increased the likelihood of a match. These results are consistent with the agreement rates reported Table 7. None of the sociodemographic covariates has a significant effect on agreement rates, which does not support *H4.1*, although some regional differences do appear.

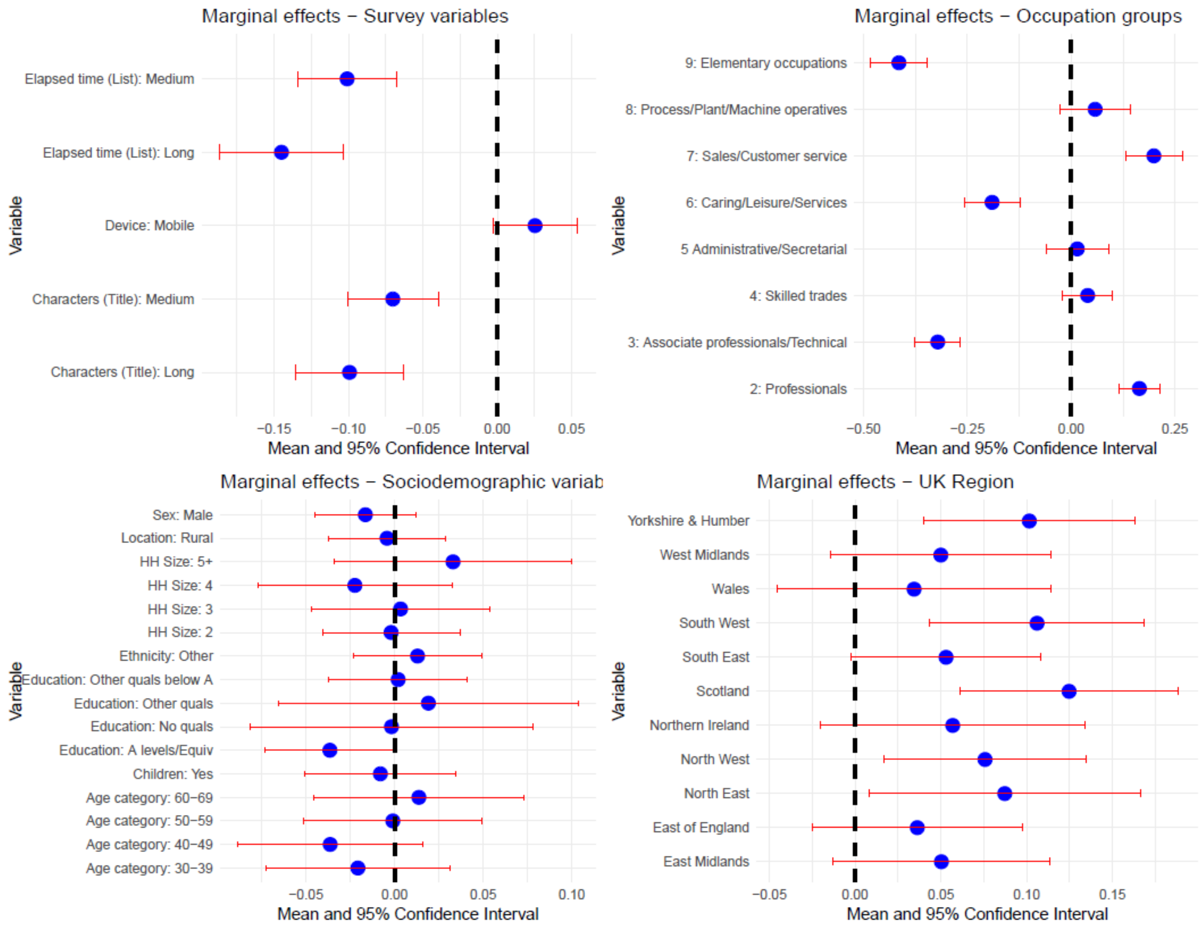

Figure 2. Logistic regression models of agreement between industries

Consistently with the industry results, longer open-text responses were generally associated with a lower likelihood of agreement between self-selected and office coded occupations, which supports *H4.2*. There was also a significant negative association between the time elapsed when responding to the list question for occupations and the likelihood of agreement between self-selected and office coded occupations. We did not find a significant effect of the device used on the probability of agreement in occupations.

Finally, as presented in Table 5, we did not find any significant device differences in "not elsewhere classified" SIC or SOC codes, or in less specific SOC codes across devices. This finding does not support *H4.4*.

4 Discussion

4.1 Conclusions

Using data from a probability-based panel in the UK, this paper compared industry and occupation coding obtained via a closed-list question with results from open-text questions that were office-coded by expert coders. The overall agreement rates between self-selected and office-coded categories were 61.7% for SIC sections (industries at the 1-digit level), 55.8% for SOC major groups (1-digit occupations), and 45.8% for SOC sub-major groups (2-digit occupations). These rates were lower than those reported in earlier self-administered surveys using closed-list occupation questions (which ranged from 70% to 88%), and in other studies comparing different methods for industry and occupation coding in online surveys (Kocar et al., 2025; Peycheva et al., 2021; Schierholz et al., 2018). While including example job titles and offering additional information via help links may have assisted some respondents, it is unclear whether these features meaningfully reduced confusion. More broadly, the low agreement rates point to the need to develop response options which are meaningful to respondents and better reflect how they conceptualise their own jobs, while still maintaining alignment with SOC categories.

The distributions of industries and occupations differed by coding method, but the dissimilarity indices were relatively low (15.0% for industries, 12.9% for occupations at the 1-digit level, and 19.2% at the 2-digit level). This suggests that closed-list questions may provide an acceptable approximation of population distributions for surveys not focused on detailed economic or occupational data. However, the results achieved through self-selected categories are unlikely to suffice for surveys where high levels of accuracy at an individual level are required. They could also be problematic if bi-variate analyses are required.

Agreement rates varied considerably by industry category. Industries that span multiple sectors (such as professional and technical services, and administration) showed particularly low agreement, with many respondents self-coding into these categories while being assigned to other categories by coders. Beyond the complexity of industry descriptions, the data suggest that some respondents may have confused the industry question with the question about their own job (occupation). The ONS has recently (Office for National Statistics, 2026) updated SIC categories to reflect recent economic changes which will undoubtedly benefit industry analysis in surveys, but we argue that further research will still be needed into how respondents approach the industry question and interpret the categories offered. Maintaining consistency with SIC categories is essential for analysis, but it is equally important to ensure that respondents understand the categories presented to them.

Similar challenges emerged with the list of occupation categories. Although we anticipated confusion between sub-major groups (2-digit) within the same major group (1-digit), mismatches occurred more frequently across major groups. This suggests that respondents may have misunderstood the options, and that presenting the major group label alongside the full category definition did not mitigate this confusion. Notable overlaps were observed between professionals and associate professionals, and between professionals and managers or directors.

We found only weak evidence that sociodemographic attributes influenced agreement rates for industry or occupation. However, paradata proved valuable in revealing how the complexity of a given industry or occupation (or more generally, the difficulty of describing it in simple terms) contributed to lower agreement between self-selected and coded categories. Specifically, longer response times for the closed-list question were associated with a lower probability of agreement. This extended processing time likely reflects greater difficulty in selecting an appropriate category, either because multiple options seemed to fit the respondent's situation, or because none appeared suitable. Consistently, indicators of length and complexity in the open-text responses were also linked to lower agreement probabilities, suggesting that industries encompassing diverse activities (or occupations involving a wide range of tasks) are less likely to align with a pre-selected option. In the literature, such mismatches have often been attributed to respondents failing to provide relevant information for coding. However, it may also be that the coding frames themselves inadequately capture complex organisations spanning multiple activities, or multifaceted roles comprising varied tasks.

Device type was also associated with notable results. Mobile respondents tended to provide shorter and less complex descriptions of their industry and occupation, and spent less time answering these questions, compared to those using other devices. Since shorter open-text responses are commonly associated with higher inter-coder reliability (or greater agreement between different coding methods), mobile phone use was positively associated with higher agreement between self-selected and coded categories. However, this does not necessarily imply that mobile respondents provide higher data quality than respondents using other devices, or that their responses are closer to a “true value”; instead, mobile respondents appear to provide less complex information which may be easier for coders to interpret and code.

Respondents took, on average, 11 seconds less to answer the closed-list industry question than to provide an open-text response; the time saving was more pronounced for occupation questions, at 35 seconds. These findings indicate that closed-list questions effectively reduce question response time. Combined with lower data processing costs (by eliminating the need for office coding), this further supports the use of closed-list approaches in non-specialist surveys.

4.2 Limitations and suggestions for further research

This paper has provided insight into the reasons for mismatches between self-selected and office-coded industries and occupations. The findings suggest that at least some of the confusion arises from issues with the classification systems themselves, which are not designed to be respondent facing and appear misaligned with how respondents understand their workplaces and jobs.

Although the low agreement rates observed here might initially be interpreted as a strong argument against deploying closed-list methods in surveys, further research is needed before discarding the approach, especially considering its effectiveness in reducing question response time and processing costs. Experimental studies testing alternative configurations of category lists which are designed for respondents, for instance, could help determine whether agreement rates can be improved. Recent qualitative work conducted at the ONS in the UK (Watson & Jones, 2026) adopts a respondent-centred design perspective to examine industry and occupation survey questions. This work investigates how participants understand and describe their jobs, and how these interpretations translate into survey respondents. Developing a clearer understanding of these response processes may also support the creation of category lists that are more intuitive and easier for respondents to interpret.

The extent and nature of device effects represent another important avenue for further research. Shorter responses (such as those provided by mobile respondents) are positively associated with higher agreement between office-assigned and self-selected codes. However, this speaks only to the ease with which such responses can be coded, and not necessarily to their accuracy. Intuitively, more complex occupations involving multiple tasks (or those that span multiple industries) should require more detailed open-text responses for accurate office coding. The relationship between job complexity, open-text responses, and coding accuracy thus warrants further investigation. At the same time, it is important to note that the absence of a "gold standard" against which to evaluate new methods undermines any assessment of industry and occupation coding methods. In this case, discrepancies between self-selected and office-coded categories may reflect errors in either.

Finally, large language models and artificial intelligence more generally will undoubtedly play an important role in achieving more accurate coding with lower respondent burden and reduced fieldwork and processing costs. However, such models still depend on robust classification schemes that require ongoing updating to reflect changes in the labour market. They also rely on collecting information from respondents, who need to understand what is being asked of them and to provide the essential details required for these models to produce accurate results. Qualitative research is therefore essential to understand how respondents think about their workplaces and jobs, and how these insights can inform more accurate classification of industries and occupations from survey data.

Acknowledgements

The work was supported by the UK Research and Innovation (UKRI)/Economic and Social Research Council (ESRC) research grant ES/X014150/1 for "Survey data collection methods collaboration: securing the future of social surveys", known as Survey Futures. 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. We would also like to extend our sincere gratitude to the National Centre for Social Research (NatCen) for facilitating access to the data in our analysis and for helping us understand the survey design and data collection processes.

References

Antoun, C., Couper, M. P., & Conrad, F. G. (2017). Effects of mobile versus PC Web on survey response quality. *Public Opinion Quarterly*, *81*(S1), 280–306. https://doi.org/10.1093/poq/nfw088

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

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30). https://doi.org/10.21105/joss.00774

Brugiavini, A., Belloni, M., Martens, M., & Buia, R. E. (2017). The "Job Coder". In *SHARE Wave 6: Panel innovations and collecting Dried Blood Spots* (pp. 51–70). MEA, Max Planck Institute for Social Law and Social Policy. https://iris.unive.it/bitstream/10278/3694317/2/SHARE_WAVE6_V01.pdf

Butt, S., Moss, S., Bourguignon, N., Jessop, C., Keyes, A., & Hussey, D. (2025). *Skills and Employment Survey 2024: Technical report*. Prepared for: Economic and Social Research Council, National Centre for Social Research (NatCen). https://wiserd.ac.uk/publication/ses-2024-natcan-technical-report/

Clement, S. L., Severin-Nielsen, M. K., & Shamshiri-Petersen, D. (2020). Device effects on survey response quality. A comparison of smartphone, tablet and PC responses on a cross sectional probability sample. *Survey Methods: Insights from the Field*. https://surveyinsights.org/?p=13585

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

Coppens, L., de Jonge, M., van Gog, T., & Kester, L. (2020). The effect of practice test modality on perceived mental effort and delayed final test performance. *Journal of Cognitive Psychology*, *32*(8), 764–770. https://doi.org/10.1080/20445911.2020.1822366

Dahl, C. M., Johansen, T., & Vedel, C. (2024, 2024/04/02). *Breaking the HISCO Barrier: Automatic Occupational Standardization with OccCANINE*. arXiv. http://arxiv.org/abs/2402.13604

De Vries, J., & Ganzeboom, H. B. G. (2008). Hoe meet ik beroep? Open en gesloten vragen naar beroep toegepast in een statusverwervingsmodel. *Mens en Maatschappij*, *83*(1), 71–96.

Décieux, J. P., & Sischka, P. E. (2024). Comparing data quality and response behavior between smartphone, tablet, and computer devices in responsive design online surveys. *SAGE Open*, *14*(2). https://doi.org/10.1177/21582440241252116

Duncan, O. D. (1961). A socioeconomic index for all occupations. In J. Scott (Ed.), *Class: Critical Concepts, Volume I* (pp. 388–426). Routledge.

Duncan, O. D., & Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, *20*(2), 210–2017. https://doi.org/10.2307/2088328

Eckhardt, K. W., & Wenger, D. E. (1975). Respondent coding of occupation. *The Public Opinion Quarterly*, *39*(2), 246–254. https://doi.org/https://www.jstor.org/stable/2748152

Ganzeboom, H. B. G. (2005). On the cost of being crude: comparison of detailed and coarse occupational coding in the ISSP 1987 data. In J. H. P. Hoffmeyer-Zlotnik & J. Harkness (Eds.), *Methodological aspects in cross-national research* (pp. 241–257). Mannheim, Germany. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-49161-7

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

Helppie-McFall, B., & Sonnega, A. (2018). Feasibility and reliability of automated coding of occupation in the Health and Retirement Study. In. Rochester, NY: Social Science Research Network.

Howe, S., Maatook, N., & Dyer, A. (2024). *Generations and Gender Study UK: Technical report*. National Centre for Social Research, UKRI Economic and Social Research Council, University of Southampton, Generations and Gender Programme. https://doc.ukdataservice.ac.uk/doc/9247/mrdoc/pdf/9247_ukggs_technical_report_mar24.pdf

ISSP Research Group. (1989). *International Social Survey Programme: Social Inequality I - ISSP 1987* Version 1.0.0) GESIS Data Archive, Cologne. https://doi.org/10.4232/1.1680

Jones, R., & Elias, P. (2004). *CASCOT: Computer-assisted Structured Coding Tool*. University of Warwick, UK.

Jung, Y., Yoo, J., Myaeng, S.-H., & Han, D.-C. (2008, 2008). A web-based automated system for industry and occupation coding.

Kantar Public, & Department for Digital Culture Media & Sport. (2019). *DCMS Participation Survey 2021/22: Annual Technical Note October 2021–March 2022*. https://assets.publishing.service.gov.uk/media/645b6f242226ee000c0ae555/DCMS_Participation_Survey_2021-22_-_Annual_Technical_Note__28.07.22___2_.pdf

Kantar Public, & Department for Digital Culture Media & Sport. (2025). *Community Life Survey Technical Report 2024-25*. https://www.gov.uk/government/statistics/community-life-survey-202425-annual-publication/community-life-survey-202425-technical-report

Kim, C., Kim, J., & Ban, M. (2020). Do you know what you do for a living? Occupational coding mismatches between coders in the Korean General Social Survey. *Research in Social Stratification and Mobility*, *70*, 100467. https://doi.org/10.1016/j.rssm.2019.100467

Kocar, S., Brown, M., & Calderwood, L. (2023). *Occupation coding in self-completion surveys: Evidence review* (Survey Futures Report, No. 3). Colchester, UK: University of Essex. https://surveyfutures.net/wp-content/uploads/2025/02/report-3-occupation-coding-self-self-completion-surveys-evidence-review.pdf

Kocar, S., Peycheva, D., Brown, M., Sakshaug, J. W., Bhaumik, C., & Calderwood, L. (2025, 14–18 July). *An evaluation of the look-up approach to occupation coding: Evidence from the Next Steps study in the UK* 11th Conference of the European Survey Research Association (ESRA), Utrectht, The Netherlands: Utrectht University.

Lambert, A. D., & Miller, A. L. (2014). Living with Smartphones: Does Completion Device Affect Survey Responses? *Research in Higher Education*, *56*(2), 166–177. https://doi.org/10.1007/s11162-014-9354-7

Massing, N., Wasmer, M., Wolf, C., & Zuell, C. (2019). How standardized is occupational coding? A comparison of results from different coding agencies in Germany. *Journal of Official Statistics*, *35*(1), 167–187. https://doi.org/10.2478/jos-2019-0008

Mavletova, A. (2013). Data quality in PC and mobile Web surveys. *Social Science Computer Review*, *31*(6), 725–743. https://doi.org/10.1177/0894439313485201

McLaughlin, G. H. (1969). SMOG grading – a new readability formula *Journal of Reading*, *12*(8), 639–646. https://www.jstor.org/stable/40011226

Mitton, D. (2025). Mapping the economy: Using industrial classifications to deliver the UK Industrial Strategy. *Government Transformation*. https://www.government-transformation.com/data/mapping-the-economy-using-industrial-classifications-to-deliver-the-uk-industrial-strategy

Office for National Statistics. (2021). *Occupation, industry and travel to work question development for Census 2021*. Retrieved 18 August 2025 from https://www.ons.gov.uk/census/censustransformationprogramme/questiondevelopment/occupationindustryandtraveltoworkquestiondevelopmentforcensus2021

Office for National Statistics. (2022). *UK SIC 2007 [Online]*. Retrieved 18 August 2025 from https://www.ons.gov.uk/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007

Office for National Statistics. (2023a). *Investigating the accuracy of Standard Industrial Classification using BRES*. Retrieved 4 May 2023 from https://www.ons.gov.uk/businessindustryandtrade/business/activitysizeandlocation/methodologies/investigatingtheaccuracyofstandardindustrialclassificationusingbres

Office for National Statistics. (2023b). *UK SOC 2020 [Online]*. Retrieved 18 August 2025 from https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2020

Office for National Statistics. (2024). *Labour Force Survey User Guide Volume 1 - LFS Background and Methodology 2024 Update*. Retrieved 20 June 2024 from https://www.ons.gov.uk/file?uri=/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/methodologies/labourforcesurveyuserguidance/volume1backgroundandmethodology2024.pdf

Office for National Statistics. (2026). *UK SIC 2026 guidance*. Retrieved 19 June 2026 from https://www.ons.gov.uk/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2026guidance

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

Park, R., Grant, D., Jans, M., Frausto, M., & Rauch, J. (2015, 2015/08/13/). Comparing manual and automated industry and occupation coding: Accuracy and cost from the perspective of the California Health Interview Survey. Proceedings of the Joint Statistical Meetings 2015 Survey Research Methods Section,

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

Peytchev, A., & Hill, C. A. (2009). Experiments in mobile Web survey design: Similarities to other modes and unique considerations. *Social Science Computer Review*, *28*(3), 319–335. https://doi.org/10.1177/0894439309353037

Russ, D. E., Ho, K.-Y., Colt, J. S., Armenti, K. R., Baris, D., Chow, W.-H., Davis, F., Johnson, A., Purdue, M. P., Karagas, M. R., Schwartz, K., Schwenn, M., Silverman, D. T., Johnson, C. A., & Friesen, M. C. (2016). Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. *Occupational and Environmental Medicine*, *73*(6), 417–424. https://doi.org/10.1136/oemed-2015-103152

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), 6. https://doi.org/10.1007/s44163-023-00050-y

Savic, N., Bovio, N., Gilbert, F., Paz, J., & Guseva Canu, I. (2022). Procode: A machine-learning tool to support (re-)coding of free-texts of occupations and industries. *Annals of Work Exposures and Health*, *66*(1), 113–118. https://doi.org/10.1093/annweh/wxab037

Schierholz, M., Gensicke, M., Tschersich, N., & Kreuter, F. (2018). Occupation coding during the interview. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *181*(2), 379–407. https://doi.org/10.1111/rssa.12297

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

Sommer, J., Diedenhofen, B., & Musch, J. (2016). Not to be considered harmful: Mobile-device users do not spoil data quality in Web surveys. *Social Science Computer Review*, *35*(3), 378–387. https://doi.org/10.1177/0894439316633452

Soszynski, M. (2025). Smartphone or laptop? Device type choice and data quality in surveys of commercial and residential utility customers. *Survey Practice*, *18*(1), 1–12. https://doi.org/10.29115/sp-2024-0031

Struminskaya, B., Weyandt, K. W., & Bosnjak, M. (2015). The effects of questionnaire completion using mobile devices on data quality. Evidence from a probability-based general population panel. *Methods, data, analyses : a journal for quantitative methods and survey methodology (MDA)*, *9*(2), 261–292. https://doi.org/10.12758/mda.2015.014

Sturgis, P., Robinson, T., Fung, L., & Roberts, C. (2025). SOCbot: Using Large Language Models to measure and classify occupations in surveys. In: OSF.

Taylor, D. G. (1976). The accuracy of respondent-coded occupation. *Public Opinion Quarterly*, *40*(2). https://doi.org/10.1086/268293

Tijdens, K. (2015). Self-identification of occupation in web surveys: requirements for search trees and look-up tables. *Survey Methods: Insights from the Field (SMIF)*. https://doi.org/10.13094/SMIF-2015-00008

Tyrell, K., & Arkell, J. (2026, 4 June 2026). *Using Generative AI to help respondents complete industry and occupation questions* Occupation and Industry Coding in Online Surveys: Evidence, Practice, and Challenges (Survey Futures Workshop), London, UK.

Watson, D., & Jones, B. (2026, 4 June 2026). *Using mental models research to review how we collect SIC and SOC information* Occupation and Industry Coding in Online Surveys: Evidence, Practice, and Challenges (Survey Futures Workshop), London, UK.

Winch, R. F., Mueller, S. A., & Godiksen, L. (1968). Reliability of respondent-coded occupational prestige. *American Sociological Review*, *34*(2), 245–251. https://www.jstor.org/stable/2092181

Supplementary material

Appendix A: Scripts for the closed-list questions on industry and occupation

The following text contains the script used for the closed-list questions on industry and occupation coding in the March 2022 NatCen Panel questionnaire:

CurEmpInd2

"The following questions refer to the main paid job you had in the seven days ending Sunday {date of preceding Sunday, e.g.: "6th February.

{IF PdWrk = 1}

Which of the following best describes the sector you work in?

G_ReadOut_II1"

1. Agriculture, forestry and fishing [HELPLINK: "Typically involves working on farms, forests or fishing boats and often based outdoors on the land. Examples include crop farming, livestock farming, logging, silviculture, marine and freshwater fishing."]

2. Mining and quarrying [HELPLINK: "Involves removing rock and other minerals from the ground. Examples include underground mining, surface mining, petrol and natural gas extraction and support services for mining and quarrying."]

3. Manufacturing [HELPLINK: "Creating goods, by hand or machine, for sale. This includes all goods production and processing such as meat, fish, baked goods and other food products, beverages, tobacco, textiles, clothes, wood, metal, furniture, printed materials, recorded media, chemical and pharmaceutical products, computer and electrical equipment, vehicles and transport equipment."]

4. Electricity, gas, steam & air conditioning [HELPLINK: "Generating and distributing electricity, gas, steam and air conditioning. This includes production and transmission of electricity, gas and gaseous fuels and steam and air conditioning.

5. Water supply, sewerage and waste management [HELPLINK: "Includes sewerage, waste management and water cleaning and processing activities."]

6. Construction [HELPLINK: "The process of building something such as a house, bridge, tunnel. This includes building construction, civil engineering, demolition, electrical and plumbing installation, plastering, joinery, roofing, scaffolding and other specialised construction activities."]

7. Trade (retail sales and wholesales) and vehicle repair [HELPLINK: "This includes both wholesale (on a fee or contract basis) and retail trade (in specialised and non-specialised stores) and refers to all goods such as foods, fuels, household goods, clothing, hardware and so on. It also includes retail sales and wholesale, as well as repair of, motor vehicles and motorcycles."]

8. Transportation and storage [HELPLINK: "Moving passengers or freight by land, sea or air. Also includes supporting warehousing and storage activities as well as postal and courier services."]

9. Accommodation and food service [HELPLINK: "This includes hotels, holiday villages and other short-stay accommodation like youth hostels etc. Food and beverage services include things such as restaurants, mobile food services and event catering."]

10. Information and communication [HELPLINK: "Creating and sharing information. This includes activities such as publishing books, journals, software etc.; Recording, producing and publishing movies, music, TV; broadcasting; telecommunications; computing and data processing and hosting"]

11. Finance and insurance [HELPLINK: "Includes activities and organisations such as central banking, holding companies, trusts, funds, insurance and reinsurance, as well as pensions and work associated with financial markets and brokerage."]

12. Real estate [HELPLINK: "Involves property which is either land or buildings. This includes buying and selling, letting and operating of private or leased real estate."]

13. Professional, scientific and technical [HELPLINK: "Involves performing specialised professional, scientific or technical activities for others, need a high level of expertise and training. This includes activities such as legal, accounting, head offices,

management consultancy, architecture, engineering, scientific research and development, advertising, market research and veterinary services.”]

14. Administrative and support services [HELPLINK: This includes activities such as renting and leasing goods, equipment and vehicles, employment agencies and HR, travel agencies, tour operators, security systems and investigation, services to buildings and landscapes, office administration, office support and other business support services.”]

15. Public administration and defence [HELPLINK: “This involves activities by government bodies, carried out by public administration, local or national. It includes compulsory social security activities, civil and criminal courts of law, tax, national defence, public order and safety, foreign affairs, immigration and the administration of government programmes.”]

16. Education [HELPLINK: “Includes all levels of education from pre-primary to degree and postgraduate level, as well as sports, cultural, driving education etc.”]

17. Human health and social work [HELPLINK: “Includes hospital activities, GP practices, dental practices, residential care and social work without accommodation.”]

18. Arts, entertainment and recreation [HELPLINK: “Establishments and facilities that provide cultural, entertainment and recreational services. This includes involvement in live performances, events or exhibits for public viewing; displaying objects and sites of interest; providing services for taking part in leisure activities. Examples include theaters, libraries, museums, sports clubs, theme parks etc.”]

19. Other service activities [HELPLINK: “Includes membership organisations (professional bodies, trade unions etc), computer and household goods repair, and other services such as dry cleaning, hairdressing, beauty treatments etc.”]

20. Employed by a household as domestic staff [HELPLINK: “This category is to allow you to state the activity of your employer, even though your employer is an individual. Includes all goods and services which may be produced by the household for their own use.”]

21. International organisations and bodies [HELPLINK: “This includes activities of international organisations such as the UN, The World Bank, the European Free Trade Association etc. As well as diplomatic activities and consular missions.”]

{IF PdWrk = 1}

CurIndust

“And what does the firm or organisation you work for mainly make or do at the place where you work?

{IF Mode = WEB: “Please give a full description. For example, do you work in “manufacturing”, “processing” or “distributing”? What are the main goods you “produce and materials you use? “Is it wholesale or retail?”}

INTERVIEWER: PROBE FOR DETAILS: E.G. MANUFACTURING, PROCESSING OR DISTRIBUTING; MAIN GOOD PRODUCED; MATERIALS USED; WHOLESALE OR RETAIL ETC.”

OPEN

SOFTCHECK: IF NUMBER OF CHARACTERS < 15: ‘It is important that you enter enough information for us to clearly understand where you work. Please ensure you have fully described what the firm or organisation you work for mainly make or do.’

{IF PdWrk = 1}

CurJbTtl

“What is the name of your job or your job title?”

OPEN

{IF PdWrk = 1}

CurEmpOcc3

"Which of the following best describes the main paid job you had in the seven days ending Sunday {date of preceding Sunday, e.g.: "6th February"}?

G_ReadOut_II1"

1. Corporate manager or Director [HELPLINK: "Such as Chief Executive, Production manager, Financial manager, marketing manager, officers in the armed forces, senior police officers, health, public health and social care managers."]

2. Other manager or owner [HELPLINK: "Such as managers and proprietors in agriculture, fishing, forestry, hotels, public houses, leisure and sports centres, travel agencies, health care practices, transport, warehousing, housing, garages, beauty salons, creative industries etc."]

3. Professional – Science, research, engineering, technology [HELPLINK: "Such as natural and social scientists, engineers, IT professionals, Web and graphic design professionals, Research & Development professionals.

4. Professional – Health [HELPLINK: "Hospital doctors, GPs, physiotherapists, speech and language therapists, midwives, community nurses, veterinarians, pharmacists, optometrists, paramedics, dentists."]

5. Professional - Education [HELPLINK: "Such as primary school teachers, SEN teachers, Headteachers, School inspectors, Teachers of English as a foreign language."]

6. Professional – Business, media, public services [HELPLINK: "Such as barristers, solicitors, lawyers, accountants, management consultants, business analysts, statisticians, actuaries, business and financial project management professionals, architects, surveyors, social workers, librarians, regulatory professional, editors, PR professionals."]

7. Associate professional - Science, engineering, technology "[HELPLINK: "Such as technicians, database administrators and web content technicians."]

8. Associate professional – Health and social care [HELPLINK: "Such as dispensing opticians, pharmaceutical technicians, dental and medical technicians, youth workers, counsellors, higher level teaching assistants, early education and childcare practitioners, veterinary nurses."]

9. Protective services [HELPLINK: "Such as non-commissioned officers, police officers (sergeant and below), fire service officers (watch manager and below), prison service officers (below principal officer)."]

10. Culture, media and sport [HELPLINK: "Such as artists, authors, translators, actors, presenters, dancers, musicians, photographers, arts officers and producers, directors, designers, athletes."]

11. Associate professional – Business and public services [HELPLINK: "Such as pilots, air traffic controllers, ship officers, legal associates, brokers, insurance underwriters, financial accounts manager, data analysts, procurement officers, estate agents, sales accounts managers, events managers, HR officers, careers advisors."]

12. Administration [HELPLINK: Such as government administration roles, book-keepers, payroll managers, bank and post office clerks, records clerks and assistants, office managers, customer service managers, sales administrators, data entry."]

13. Secretarial [HELPLINK: "Such as medical secretaries, legal secretaries, school secretaries, company secretaries and administrations, personal assistants, receptionists, typists."]

14. Skilled worker – Agriculture, gardening [HELPLINK: "Such as farmers, gardeners, groundsmen."]

15. Skilled worker – metal, electrical, electronics [HELPLINK: "Such as sheet metal workers, smiths, tool makers, vehicle technicians, mechanics, electricians."]

16. Skilled worker – construction and building [HELPLINK: "Such as bricklayers, roofers, plumbers, carpenters, glaziers, plasterers."]

17. Skilled worker – textiles, printing and other [HELPLINK: "Such as upholsterers, tailors and dressmakers, printers, butchers, bakers, florists."]

18. Caring personal services [HELPLINK: "Such as early education and childcare assistants, teaching assistants, childminders, pest control officers, nursing auxiliaries, dental nurses, care workers."]

19. Leisure, travel and personal services [HELPLINK: "Such as travel agents, air travel assistants, hairdressers, beauticians, housekeepers, caretakers."]

20. Community and civil enforcement role [HELPLINK: "Such as police community support officers, parking and civil enforcement."]

21. Sales [HELPLINK: "Such as sales assistants, cashiers, telephone salesperson, pharmacy and optical dispensing assistants, visual merchandiser, market and street trader, shopkeeper."]

22. Customer services [HELPLINK: "Such as call and contact centre occupations, telephonists, market research interviewers."]

23. Factory, plant or machine operative [HELPLINK: "Such as process operatives (food, drink, textiles, metal making etc), machine operatives, assemblers, inspectors and testers, sewing machinists, scaffolders, stagers and riggers."]

24. Drivers and transport operatives [HELPLINK: "Such as large goods vehicle drivers, bus and coach drivers, taxi drivers, driving instructors, crane drivers, fork-lift truck drivers, train and tram drivers, marine, air and rail transport operatives."]

25. Elementary trade [HELPLINK: "Such as farm workers, ground workers, packers, canners, fillers."

26. Elementary administration or services [HELPLINK: "Such as postal workers, mail sorters, messengers, window cleaners, launderers, refuse and salvage, security guards, shelf fillers, warehouse operatives, delivery operatives, hospital porters, kitchen and catering assistants, bar staff."]

{IF PdWrk = 1}CurOccup

"And in your own words, what do you mainly do in your job?"

{IF Mode = WEB: "Please give a full description."}

INTERVIEWER: PROBE FOR DETAILS"

OPEN

SOFTCHECK: IF NUMBER OF CHARACTERS < 15: 'It is important that you enter enough information for us to clearly understand your job. Please ensure you have fully described what you do.'

Appendix B: Sample composition

Table B1 provides a descriptive analysis of the two analytical samples. The distributions of key sociodemographic variables are not statistically different across both samples for any confidence level.

Table B1. Summary of sample composition per sample

Variable	Category	SIC (N = 4,669)	SOC (N = 4,589)
Sex	Female	57.8	58.1
	Male	42.2	41.9
Age group	18-29	9.6	9.6
	30-39	23.8	23.8
	40-49	27.8	27.6
	50-59	27.6	27.7
	60-69	11.2	11.3
Highest education level	Degree or equivalent, and above	58.1	58.2
	A levels or vocational level 3 or equivalent and above, but below degree	18.3	18.5
	Other qualifications below A levels or vocational level 3 or equivalent	17.8	17.6
	Other qualification	2.4	2.5
	No qualifications	3.3	3.2
Ethnicity	White British	82.8	82.8
	Other	17.2	17.2
Monthly income (£)	£1000 or less	12.3	12.4
	£1001 to £1500	16.4	16.3
	£1501 to £2500	31.4	31.3
	More than £2500	37.8	38.0
	Prefer not to answer	2.1	2.1
Household size	1	16.9	17.0
	2	34.0	34.0
	3	20.2	20.0
	4	21.7	21.8
	5+	7.2	7.2
UK Region	Greater London	9.4	9.4
	North East	4.0	4.1
	North West	11.1	11.2
	Yorkshire and The Humber	9.2	9.0
	East Midlands	8.9	9.0
	West Midlands	7.7	7.8
	South East	15.1	14.9
	South West	8.5	8.7
	Wales	4.0	4.0
	Scotland	7.8	7.7
	Northern Ireland	4.3	4.3

Appendix C: Coded and selected occupations at the 2-digit level

Table C1. Distribution of coded and selected occupations (2-digit level)

Code	Description	Selected (%)	Coded (%)	Agreement rate (%)	N (coded)
11	Corporate managers and directors	5.1	7.1	27.3	326
12	Other managers and proprietors	8.7	3.1	49.6	141
21	Science, research, engineering and technology professionals	9.1	7.7	59.6	354
22	Health professionals	8.8	6.4	90.1	292
23	Teaching and other educational professionals	11.3	7.3	89.9	335
24	Business, media and public service professionals	7.8	10.3	31.7	473
31	Science, engineering and technology associate professionals	1.7	2.7	15.3	124
32	Health and social care associate professionals	2.8	2.7	24.8	125
33	Protective service occupations	1.1	1.1	47.9	48
34	Culture, media and sports occupations	1.4	1.8	28.4	81
35	Business and public service associate professionals	3.5	8.0	8.2	367
41	Administrative occupations	11.4	11.0	56.6	505
42	Secretarial and related occupations	1.0	2.1	35.7	98
51	Skilled agricultural and related trades	0.7	0.5	76.0	25
52	Skilled metal, electrical and electronic trades	1.8	2.4	45.5	110
53	Skilled construction and building trades	2.0	1.6	73.0	74
54	Textiles, printing and other skilled trades	0.9	1.4	22.6	62
61	Caring personal service occupations	3.6	7.0	37.0	322
62	Leisure, travel and related personal service occupations	1.6	1.4	37.5	64
63	Community and civil enforcement occupations	0.6	0.1	25.0	4
71	Sales occupations	4.2	3.9	52.5	181
72	Customer service occupations	5.3	1.8	61.9	84
81	Process, plant and machine operatives	2.2	2.0	46.2	93
82	Transport and mobile machine drivers and operatives	2.2	2.2	75.2	101
91	Elementary trades and related occupations	0.4	0.3	15.4	13
92	Elementary administration and service occupations	0.9	4.1	12.3	187

Appendix D: Cross-classification tables of selected and coded industries and occupations

Table D1. Cross-classification table of selected and coded industries (percentage of responses)

Selected	Industry description	Office coded																				
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A	Agriculture, forestry and fishing	0.3	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
B	Mining and quarrying	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C	Manufacturing	0.0	0.0	4.8	0.0	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
D	Electricity, gas, steam & air conditioning	0.0	0.0	0.1	0.4	0.0	0.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
E	Water supply, sewerage and waste management	0.0	0.0	0.1	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F	Construction	0.0	0.0	0.3	0.0	0.0	2.1	0.2	0.0	0.0	0.0	0.0	0.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
G	Trade (retail sales and wholesales) and vehicle repair	0.0	0.0	0.4	0.0	0.0	0.0	5.6	0.2	0.1	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.1	0.1	0.0	0.0
H	Transportation and storage	0.0	0.0	0.5	0.0	0.0	0.1	0.4	2.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0
I	Accommodation and food service activities	0.0	0.0	0.1	0.0	0.0	0.0	0.3	0.0	1.9	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.1	0.0	0.1	0.0	0.0
J	Information and communication	0.0	0.0	0.2	0.0	0.0	0.0	0.2	0.0	0.0	2.6	0.0	0.0	0.5	0.1	0.3	0.0	0.3	0.2	0.1	0.0	0.0
K	Finance and insurance	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.1	0.0	0.1	3.8	0.1	1.1	0.1	0.2	0.1	0.0	0.0	0.0	0.0	0.0
L	Real estate	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
M	Professional, scientific and technical	0.0	0.0	1.4	0.0	0.1	0.2	0.5	0.1	0.0	1.3	0.2	0.2	3.6	0.3	1.0	0.7	0.8	0.1	0.3	0.0	0.0
N	Administrative and support services	0.0	0.0	0.3	0.0	0.0	0.2	0.6	0.2	0.1	0.4	0.1	0.1	0.8	0.5	1.2	0.7	1.8	0.1	0.2	0.0	0.0
O	Public administration and defence	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.1	0.1	0.4	0.2	5.0	0.1	0.2	0.1	0.0	0.0	0.0
P	Education	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.1	0.1	12.5	0.6	0.1	0.1	0.0	0.0
Q	Human health and social work	0.0	0.0	0.2	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.6	0.2	13.0	0.1	0.2	0.0	0.0
R	Arts, entertainment and recreation	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.1	0.1	0.0	0.1	0.1	1.5	0.1	0.0	0.0
S	Other service activities	0.0	0.0	0.4	0.0	0.1	0.2	0.7	0.3	0.3	0.2	0.0	0.1	0.7	0.8	0.6	0.2	1.0	0.2	0.8	0.0	0.0
T	Employed by a household as domestic staff	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.0	0.0	0.1	0.0
U	International organisations and bodies	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0

***Note:** The sum of the cell values in this table is 100%.

Table D2. Cross-classification table of selected and coded occupations (2-digit level) (percentage of responses)

Selected	Occupation description	Coded																									
		11	**12**	**21**	**22**	**23**	**24**	**31**	**32**	**33**	**34**	**35**	**41**	**42**	**51**	**52**	**53**	**54**	**61**	**62**	**63**	**71**	**72**	**81**	**82**	**91**	**92**
11	Corporate managers and directors	1.9	0.4	0.4	0.0	0.1	1.0	0.1	0.0	0.0	0.1	0.6	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
12	Other managers and proprietors	1.9	1.5	0.6	0.0	0.1	1.1	0.1	0.1	0.0	0.2	1.0	0.7	0.0	0.0	0.1	0.0	0.3	0.1	0.1	0.0	0.3	0.0	0.1	0.1	0.0	0.2
21	Science, research, engineering and technology professionals	0.3	0.1	4.6	0.1	0.2	1.5	0.7	0.0	0.0	0.1	0.5	0.1	0.0	0.0	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0
22	Health professionals	0.3	0.1	0.3	5.7	0.0	0.4	0.0	0.4	0.0	0.1	0.2	0.1	0.1	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
23	Teaching and other educational professionals	0.2	0.0	0.1	0.1	6.6	0.3	0.1	0.6	0.0	0.0	0.4	0.2	0.0	0.0	0.0	0.0	0.1	2.1	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.3
24	Business, media and public service professionals	0.8	0.2	0.4	0.0	0.0	3.3	0.2	0.1	0.1	0.3	1.4	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
31	Science, engineering and technology associate professionals	0.0	0.0	0.4	0.0	0.0	0.2	0.4	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
32	Health and social care associate professionals	0.0	0.1	0.0	0.3	0.0	0.4	0.0	0.7	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1
33	Protective service occupations	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.5	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
34	Culture, media and sports occupations	0.0	0.0	0.2	0.0	0.0	0.2	0.0	0.0	0.0	0.5	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
35	Business and public service associate professionals	0.4	0.0	0.2	0.0	0.0	1.0	0.1	0.1	0.1	0.0	0.7	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
41	Administrative occupations	0.2	0.1	0.3	0.1	0.2	0.6	0.4	0.2	0.1	0.0	1.3	6.2	0.9	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.2
42	Secretarial and related occupations	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
51	Skilled agricultural and related trades	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
52	Skilled metal, electrical and electronic trades	0.0	0.1	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0
53	Skilled construction and building trades	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	1.2	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.1	0.0
54	Textiles, printing and other skilled trades	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0
61	Caring personal service occupations	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	2.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.2
62	Leisure, travel and related personal service occupations	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.3
63	Community and civil enforcement occupations	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1
71	Sales occupations	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	2.1	0.0	0.0	0.0	0.0	0.5
72	Customer service occupations	0.2	0.1	0.1	0.0	0.0	0.2	0.1	0.1	0.0	0.0	0.2	0.8	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.0	1.3	1.1	0.0	0.0	0.0	0.7
81	Process, plant and machine operatives	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.9	0.1	0.1	0.5
82	Transport and mobile machine drivers and operatives	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	1.7	0.0	0.2
91	Elementary trades and related occupations	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
92	Elementary administration and service occupations	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5

*Note:** The sum of the cell values in this table is 100%.

Appendix E: Logistic regression models

Both tables include the estimated coefficients (Coef), their significance level (Sig), the standard errors (SE), and odds ratio (OR). Significance levels are coded as follows: ***p < 0.001, **p<0.01, *p< 0.05.

Table E1. Logistic regression model of agreement between selected and coded industries

Variable	Category	Coef.	Sig.	SE	OR
Intercept	-	1.005		0.626	2.731
Coded industry	A/B: Agriculture, forestry and fishing / Mining and quarrying	*Ref.*			
	C: Manufacturing	-1.118		0.584	0.327
	D: Electricity, gas, steam & air conditioning	0.248		0.768	1.282
	E: Water supply, sewerage and waste management	-0.217		0.689	0.805
	F: Construction	-0.896		0.598	0.408
	G: Trade (retail sales and wholesales) and vehicle repair	-0.837		0.584	0.433
	H: Transportation and storage	-0.564		0.601	0.569
	I: Accommodation and food service activities	-0.309		0.608	0.734
	J: Information and communication	-1.092		0.591	0.336
	K: Finance and insurance	0.648		0.613	1.911
	L: Real estate	-1.544	*	0.634	0.214
	M: Professional, scientific and technical	-1.314	*	0.585	0.269
	N: Administrative and support services	-2.777	***	0.620	0.062
	O: Public administration and defence	-0.912		0.584	0.402
	P: Education	0.429		0.587	1.535
	Q: Human health and social work	-0.214		0.581	0.807
	R: Arts, entertainment and recreation	-0.795		0.606	0.451
	S: Other service activities	-1.414	*	0.615	0.243
	T/U: Employed by a household as domestic staff	-2.047	**	0.785	0.129
UK Region	Greater London	*Ref.*			
	North East	0.203		0.200	1.225
	North West	0.238		0.148	1.269
	Yorkshire & Humb	0.047		0.154	1.048
	East Midlands	0.191		0.158	1.210
	West Midlands	0.171		0.161	1.186
	East of England	0.200		0.153	1.222
	South East	0.248		0.138	1.282
	South West	0.257		0.159	1.294
	Wales	0.139		0.199	1.149
	Scotland	-0.082		0.160	0.921
	Northern Ireland	0.109		0.189	1.115
Sex	Female	*Ref.*			
	Male	0.103		0.071	1.109
Age category	18-29	*Ref.*			
	30-39	0.175		0.127	1.191
	40-49	0.107		0.128	1.113
	50-59	0.223		0.124	1.250
	60-69	0.068		0.146	1.070
Highest qualifications	Degree or equivalent and above	*Ref.*			
	A-levels, Vocational level 3, equivalent but below degree	-0.144		0.089	0.866
	Other qualifications below A levels, Vocational level 3 or equivalent	-0.035		0.093	0.966
	Other qualifications	-0.454	*	0.208	0.635
	No qualifications	0.002		0.188	1.002
Ethnicity	White British	*Ref.*			
	Other	0.096		0.091	1.101
Household size	1	*Ref.*			
	2	-0.078		0.097	0.925
	3	0.196		0.127	1.217

Variable	Category	Coef.	Sig.	SE	OR
	4	0.226		0.138	1.254
	5+	0.338	.	0.173	1.402
Device	Other	*Ref.*			
	Mobile	0.171	*	0.072	1.187
SMOG Index (Open-text question)	Low complexity (< 9)	*Ref.*			
	Medium complexity (>=9 & <= 12)	-0.176	*	0.079	0.839
	High complexity (> 12)	-0.347	***	0.086	0.707
Elapsed time (Closed-list question)	Short (< 17 seconds)	*Ref.*			
	Medium (>= 17 & <= 37 characters)	-0.532	***	0.085	0.587
	Long (< 37 characters)	-0.632	***	0.092	0.532
Residual deviance	5530.7				
Null deviance	6214.3				
AIC	5632.7				

Table E2. Logistic regression model of agreement between selected and coded occupations

Variable	Category	Coef.	Sig.	SE	OR
Intercept	-	0.803		0.258	2.233
Coded occupation	1: Managers, directors and senior officials	Ref.			
	2: Professional occupations	0.760	***	0.115	2.138
	3: Associate professional occupations	-1.433		0.131	0.238
	4: Administrative and secretarial occupations	0.170		0.131	1.185
	5: Skilled trades occupations	0.061		0.164	1.063
	6: Caring, leisure and other service occupations	-0.802		0.147	0.449
	7: Sales and customer service occupations	0.953		0.181	2.595
	8: Process, plant and machine operatives	0.249		0.188	1.283
	9: Elementary occupations	-2.032	***	0.226	0.131
UK Region	Greater London	Ref.			
	North East	0.433		0.203	1.541
	North West	0.374		0.149	1.454
	Yorkshire & Humb	0.506		0.157	1.659
	East Midlands	0.247		0.159	1.281
	West Midlands	0.246		0.162	1.279
	East of England	0.178		0.153	1.194
	South East	0.261		0.139	1.298
	South West	0.529		0.160	1.697
	Wales	0.168		0.200	1.184
	Scotland	0.626		0.165	1.871
	Northern Ireland	0.280		0.194	1.323
Sex	Female	Ref.			
	Male	-0.083		0.073	0.921
Age category	18-29	Ref.			
	30-39	-0.104		0.134	0.901
	40-49	-0.182		0.134	0.834
	50-59	-0.005		0.130	0.995
	60-69	0.070		0.153	1.072
Highest qualifications	Degree or equivalent and above	Ref.			
	A-levels, Vocational level 3, equivalent but below degree	-0.182		0.093	0.834
	Other qualifications below A levels, Vocational level 3 or equivalent	0.010		0.100	1.010
	Other qualifications	0.096		0.219	1.101
	No qualifications	-0.009		0.204	0.991
Ethnicity	White British	Ref.			
	Other	0.065		0.093	1.067
Household size	1	Ref.			
	2	-0.010		0.099	0.990

Variable	Category	Coef.	Sig.	SE	OR
	3	0.017		0.129	1.017
	4	-0.112		0.139	0.894
	5+	0.168		0.174	1.182
Device	Other	Ref.			
	Mobile	0.127		0.073	1.136
Character length	Short (< 15)	Ref.			
(Job title)	Medium (>=15 & <= 25)	-0.356	***	0.080	0.700
	Long (> 25)	-0.499	***	0.093	0.607
Elapsed time	Short (< 20 seconds)	Ref.			
(2 job questions)	Medium (>= 20 & <= 100 seconds)	-0.512	***	0.088	0.599
	Long (> 100 seconds)	-0.725	***	0.106	0.484
Residual deviance	5364.7				
Null deviance	6300.6				
AIC	5446.7				

Appendix

Table A1. Questions included in the item non-response and substantive analysis

Mode	Module	Order	Variable name	Question	Variable type	Group selected for analysis
Telephone	Sociodemographic	1	Religion	*What is your religion?*	Single choice	No religion
		2	ResTme	*How long have you lived at this address?*	Single choice	Five years or more
		3	HSat	*And, how satisfied are you with this accommodation?*	Single choice	Fairly satisfied / Very satisfied
	Internet	4	IntHhHave	*Does your household have access to the internet at home?*	Binary	Yes
		5	IntPersUse	*Do you personally use the internet at home, work, or elsewhere?*	Binary	Yes
	Welsh	6	WelUndSpk	*Can you understand spoken Welsh?*	Binary	Yes
		7	WelSpk	*Can you speak Welsh?*	Binary	Yes
		8	WelRead	*Can you read Welsh?*	Binary	Yes
		9	WelWrite	*Can you write Welsh?*	Binary	Yes
	Occupation/Education	10	HighEd	*Are you currently studying for a university degree or equivalent level qualification?*	Binary	Yes
	Transport	11	CarUse	*Is there a car or van normally available for use by you or anyone in your household?*	Binary	Yes
		12	Bus12M	*In the last 12 months, have you used bus services in Wales?*	Binary	Yes
		13	Train12M	*In the last 12 months, have you used train services in Wales?*	Binary	Yes
		14	BusOverSat	*Overall, how satisfied or dissatisfied are you with bus services? We're interested in your views, even if you don't use buses.*	Single choice	Fairly satisfied / Very satisfied
		15	TrainOverSat	*Overall, how satisfied or dissatisfied are you with train services? We're interested in your views, even if you don't use trains.*	Single choice	Fairly satisfied / Very satisfied
	Arts/Culture/ Heritage	16	ArchAtt	*In the last 12 months, have you used an archive or records office in Wales or elsewhere, either in person or online?*	Binary	Yes
	Climate change	17	CliChanView	*You may have heard the idea that the world's climate is changing due to increases in temperature over the past 100 years. What is your personal opinion on this? Is it…*	Single choice	Definitely changing / Probably changing
		18	CliChanCause2	*Do you think that climate change is caused by natural processes, human activity, or both?*	Single choice	Entirely by human activity / Mainly by human activity
		19	CliChanCon	*How concerned, if at all, are you about climate change?*	Single choice	Fairly concerned / Very concerned
		20	CliChanWhen	*When, if at all, do you think climate change will start to have an impact in Wales?*	Single choice	Already having an impact
		21	CliChanRespPublic	*How much responsibility do you think each of the following has to tackle climate change in Wales?: The general public*	Single choice	A lot
		22	CliChanRespBus	*How much responsibility do you think each of the following has to tackle climate change in Wales?: Business*	Single choice	A lot
		23	CliChanRespGov	*How much responsibility do you think each of the following has to tackle climate change in Wales?: The government*	Single choice	A lot
		24	CliChanGov	*Thinking now about how the government is responding to climate change, do you think the government is doing…*	Single choice	A lot
	Social care	25	Carer	*Do you look after or give any help or support to family members, friends, neighbours or others because of long-term physical or mental ill-health or disability, or problems related to old age?*	Binary	Yes
		26	SCUseYN	*In the last 12 months, have you received help for yourself from care and support services in Wales?*	Binary	Yes

Mode	Module	Order	Variable name	Question	Variable type	Group selected for analysis
		27	SCUse2	*In the last 12 months have you received any help from care and support services in Wales to care for, or arrange care for, someone else?*	Binary	Yes
	Health/GP/Hospital	28	GenHealth	*How is your health in general; is it...*	Single choice	Good / Very good
		29	GpSeenDr	*Thinking about the last 12 months, have you had an appointment with a GP or family doctor about your own health?*	Binary	Yes
		30	GpSeenPract	*Have you had an appointment with any other health professional based at your GP surgery in the last 12 months*	Binary	Yes
		31	HspHadApp	*In the last 12 months, have you had an NHS hospital appointment?*	Binary	Yes
		32	HspAE	*have you attended an Accident and Emergency or a Minor Injury Unit for your own health?*	Binary	Yes
Online	Recycling	33	UaRecycQual	*How satisfied or dissatisfied are you with the recycling collection service provided by [council]?*	Single choice	Fairly satisfied / Very satisfied
		34	UaRecycComm	*[council] keeps me informed about its recycling collection service*	Single choice	Strongly agree / Tend to agree
		35	RecycHHClo	*In the last 12 months, have you carried out or arranged a repair or alteration of clothing that would otherwise have been unused or thrown away?*	Binary	Yes
		36	RecycHHFrn	*In the last 12 months, have you carried out or arranged a repair of any household items, such as furniture, a fridge or a kettle, that would otherwise have been unused or thrown away?*	Binary	Yes
	Physical punishment of children	37	SmackAgree	*To what extent do you agree or disagree that it is sometimes necessary to smack a child?*	Single choice	Tend to agree / Strongly agree
		38	SmackHarmSelf	*Which of these statements comes closest to your opinion on whether it's appropriate for a parent to smack a child in the following circumstances? To stop them doing something which is dangerous or harmful to themselves*	Single choice	Yes, whenever necessary / Yes, as a last resort
		39	SmackHarmAnoth	*Which of these statements comes closest to your opinion on whether it's appropriate for a parent to smack a child in the following circumstances? To stop them doing something which is dangerous or harmful to another child*	Single choice	Yes, whenever necessary / Yes, as a last resort
		40	SmackOOC	*Which of these statements comes closest to your opinion on whether it's appropriate for a parent to smack a child in the following circumstances? To stop behaviour that's out of control – for example, a tantrum or meltdown.*	Single choice	Yes, whenever necessary / Yes, as a last resort
		41	SmackPun	*Which of these statements comes closest to your opinion on whether it's appropriate for a parent to smack a child in the following circumstances? As a punishment for misbehaving.*	Single choice	Yes, whenever necessary / Yes, as a last resort

Note: In this table, "Order" indicates the question number appearing in the item non-response plots.

Table A2. Logistic regression model of participation in online modules – Full results

Variable	Category	Beta	SE	OR
Knock-to-nudge status	Non-nudged (Reference)	-	-	-
	Nudged	-0.598***	0.063	0.550
Sex	Female (Reference)	-	-	-
	Male	-0.235***	0.055	0.791
Age	16-24 (Reference)	-	-	-
	25-34	0.472	0.152	1.116
	35-49	0.133	0.151	1.255
	50-64	0.316	0.157	1.170
	65+	0.402	0.163	1.147
Highest qualifications	Level 1 (Reference)	-	-	-
	Level 2	0.333*	0.140	1.395
	Level 3	0.330*	0.150	1.390
	Level 4	0.317*	0.128	1.373
	No qualifications	0.112	0.139	1.247
	Others	0.283*	0.140	1.327
Ethnicity	White British	-	-	-
	Other (Reference)	-0.570***	0.131	0.565
Marital status	Single (Reference)	-	-	-
	Married/Civil partnership	0.188	0.087	1.121
	Separated	0.097	0.166	0.759
	Divorced	0.913	0.099	1.011
	Widowed	0.630	0.115	1.057
Household size	1	-	-	-
	2	0.885	0.084	1.012
	3	0.958	0.102	0.995
	4+	0.408	0.109	0.914
Tenure	Owned (Reference)	-	-	-
	Private rented	0.510	0.261	1.187
	Other private rented/Rent free	-0.258**	0.085	0.772
	Rented from council/LA	0.261	0.098	0.896
	Other social rented	0.266	0.126	0.869
Wales region	Blaneau Gwent (Reference)	-	-	-
	Bridgend	0.586	0.207	0.893
	Caerphilly	0.497	0.204	1.149
	Cardiff	0.775	0.184	1.054
	Carmarthenshire	0.273	0.197	0.805
	Ceredigion	0.656	0.241	1.113
	Conwy	0.099	0.209	0.709
	Denbighshire	0.350	0.220	0.814
	Flintshire	0.270	0.203	0.799
	Gwynedd	0.651	0.219	1.104
	Isle of Anglesey	0.206	0.218	0.759
	Merthyr Tydfil	0.470	0.218	0.855
	Monmouthshire	0.094	0.240	1.496
	Neath Port Talbot	0.149	0.193	0.757
	Newport	0.314	0.212	1.238
	Pembrokeshire	0.129	0.212	0.725
	Powys	0.877	0.193	1.030
	Rhondda Cynon Taf	-0.449*	0.183	0.638
	Swansea	0.589	0.193	1.110
	Torfaen	0.739	0.227	1.079
	Vale of Glamorgan	0.112	0.204	0.723
	Wrexham	0.240	0.209	0.782
Urban/Rural	Rural (Reference)	-	-	-
	Urban	0.753	0.069	0.979
Intercept		1.921***	0.262	6.826
Log-likelihood	-4611.696 ($df = 46$)			
AIC	9315.4			

Note: OR – Odds Ratio, SE – Standard Error.

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

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 | Université de Lausanne
Ipsos
verian Formerly Kantar Public

www.surveyfutures.net