National Centre for Social Research

Skills and Employment Survey 2024: The impact of mode of data collection on survey estimates

Authors: Sarah Butt, Curtis Jessop, Simon Moss, Joanna D'Ardenne, David Hussey, Toby Li and Martin Wood

Prepared for: Economic and Social Research Council

JULY 2025

We believe that social research has the power to make life better.

By really understanding the complexity of people's lives and what they think about the issues that affect them, we give the public a powerful and influential role in shaping decisions and services that can make a difference to everyone. And as an independent, not for profit organisation we're able to put all our time and energy into delivering social research that works for society.

National Centre
for Social Research
35 Northampton Square
London EC1V 0AX
020 7250 1866

www.natcen.ac.uk

A Company Limited by Guarantee

Registered in England No.4392418

A Charity registered in England and Wales (1091768) and Scotland (SC038454)

This project was carried out in compliance with ISO20252

designbysoapbox.com

Contents

Acknowledgements

Conducting fieldwork for the 2024 Skills and Employment Survey was very much a team effort. The core research team at the National Centre for Social Research (NatCen) consisted of Martin Wood, Sarah Butt, Simon Moss and Noemie Bourguignon. Work on the NatCen Opinion Panel was overseen by Curtis Jessop and led day to day by Marta Mezzanzanica. Throughout the process the research team received invaluable support from Project Manager Jan Read. The specialist skills of many departments contributed to the success of the project including statisticians Fritz Anker, Alina Carabat, David Hussey, Anna Keys and Toby Li; programmers Nafiis Boodhumeah and Richard Weeber; Data Managers Jessica Messling and Grace Everitt; members of NatCen's Questionnaire Design and Testing Hub led by Joanna D'Ardenne, and the NatCen Telephone Unit.

The research team at NatCen worked closely with the SES academic team throughout the project. Thank you to Alan Felstead from Cardiff University and Wales Institute for Social and Economic Research and Data (WISERD)) Rhys Davies (also Cardiff University and WISERD), Golo Henseke (Institute of Education, UCL), Ying Zhou (Surrey University), Duncan Gallie (University of Oxford) and Francis Green (Institute of Education, UCL) for unfailing enthusiasm for and support of the project at every stage.

We would like to extend particular thanks to the large number of Field interviewers, supervisors and Field Performance Managers who were integral to delivering the face-to-face fieldwork. Last, but most importantly, thanks to all of the respondents who gave up their time to take part in the survey.

The Skills and Employment Survey 2024 is funded by the Economic and Social Research Council, the Department for Education, and the Advisory and Conciliation and Arbitration Service with additional funding from the Department for the Economy to extend the survey to Northern Ireland

Executive Summary

The Skills and Employment Survey series (SES) is a long-running series of cross-sectional surveys of people aged 20-60 (20-65 since 2006) in paid employment that provides the means to chart and explain the changing pattern of job quality and job skills over time. The 2024 SES was conducted between September 2023 and July 2024.

In 2024 for the first time the survey was conducted not only face to face but also, in a data collection exercise run in parallel to the main SES, via a web-telephone survey using sample from the NatCen Opinion Panel. The parallel run provides a way to explore the potential impact of the change in sampling design and mode on survey estimates and to provide a baseline for a new SES timeseries should the decision be taken to move away from face-to-face data collection.

This report summarises the magnitude and direction of differences observed in SES survey estimates between the two data collection designs. Estimates were compared for 312 categorical variables and a total of 1,242 survey estimates or proportions. The report also attempts to identify whether any differences are more likely to be the result of differences in sample composition between the two surveys – that is differences in the types of people who took part - or the result of measurement effects between face to face and online data collections – that is differences in how similar respondents choose to respond to questions depending on how they are asked.

The report is accompanied by a series of annex tables which provide a variable-by-variable comparison of the estimates across the two surveys.

The main findings are:

- There were differences between the two surveys in the distribution of responses to survey questions. An average difference across response categories (ignoring directionality) of 3 percentage points or more between the Face to Face and Panel surveys was observed for 161 out of 312 variables; 67 variables exhibited a difference of 5 percentage points or more, and 7 variables a difference of 10 percentage points or more.

- Differences in sample composition appear to account for a limited number of the differences observed. After applying calibration weights, which align the age, sex, education, occupation and regional profiles of the Face to Face and Panel samples, few differences were observed in the demographic and job characteristics

of respondents to the two surveys. Controlling for these differences in sample composition between the two surveys reduced the magnitude of some of the differences. However, the majority persisted.

- There is some evidence to suggest that measurement effects were produced by the difference between the two survey designs. Amongst the survey questions identified as being at risk of mode effects, 70% to 80% produced estimates that met the threshold for being different, after controlling for differences in sample composition. This compared to 49% of survey questions identified as not being at risk of measurement effects which met this threshold.

- Further evidence of measurement effects being present was found when comparing the direction of the differences observed between the two surveys. For variables considered to be at risk of social desirability/positivity bias, it is possible to predict which response categories should be more/less prevalent among respondents completing the survey with an interviewer versus self-completion. For the majority of these variables (81%), differences between modes were observed in the expected direction, that is Panel respondents were less likely to provide the positive or socially desirable response. The difference between modes was statistically significant in 63% of cases.

- Evidence from SES suggests that it is equally possible to collect meaningful data on occupation and industry in online and face-to-face surveys using open questions and coding responses to standard code frames such as SOC and SIC. The proportion of cases that could not be assigned a SOC or SIC code was very low in both the Face to Face and Panel studies. A comparison of the confidence scores assigned to those codes shows little difference between the two surveys. There were differences in the 2-digit SOC and 1-digit SIC codes which respondents selected when asked to self-code their response from a drop-down menu, compared with the codes assigned based on their responses to the open questions. For a survey such as SES where there is particular interest in detailed information on occupation and industry, it may be advantageous to collect job details via open code questions, even if the data collection moves away from face to face.

- It should be borne in mind that the choice of threshold for identifying a difference is arbitrary to some extent. SES data users are encouraged to decide on a case by case basis, depending on the variables used and the purpose of the analysis, whether the data from the two surveys are sufficiently similar to enable the Face to Face and Panel data to be combined. The annex tables showing descriptive findings for each SES variable which accompany this report provide a useful starting point for this.

1. Background

The Skills and Employment Survey series is a long running series of cross-sectional surveys of people aged 20-60 (20-65 since 2006) in paid employment that provides the means to chart and explain the changing pattern of job quality and job skills over time. The survey has been carried out approximately every five years over the last three decades – in 1986, 1992, 1997, 2001, 2006, 2012, 2017 and 2024. The survey has traditionally been conducted face to face with a random sample of individuals across Great Britain (and occasionally Northern Ireland).

In 2024, for the first time, the survey was conducted not only face to face (F2F) but also, in a data collection exercise run in parallel to the main SES, via a web-telephone survey. In light of rising costs and falling response rates experienced by face to face surveys in recent years, a trend exacerbated by the COVID-19 pandemic, the UK's large scale probability surveys have increasingly been looking into alternative modes of data collection, including online data collection. There is, however, a need for caution given that a change in survey mode could lead to difference in the survey estimates obtained, either as a result of differences in the types of people who answer F2F versus online surveys or the way in which they answer surveys which are presented visually and in the absence of an interviewer. One of the strengths of SES is the almost 40 year timeseries it provides, allowing analysts to explore trends in people's experiences of work back to the 1980s. The parallel run was, therefore, particularly important as a way to explore the potential impact of the change in mode on survey estimates and to provide a baseline for a new timeseries should the decision be taken to move away from face to face data collection.

This report summarises the magnitude and direction of differences observed in SES survey estimates between the two survey designs. The report also attempts to identify whether any differences observed are most likely to be the result of differences in sample composition between the two surveys – that is differences in the types of people who took part - or the result of measurement effects between face to face and online data collections – that is differences in how similar respondents choose to respond to questions depending on how they are asked. The analysis presented here provides a high-level overview of mode differences occurring between the two SES surveys across a large number and variety of variables. Attempts are made to identify variables at risk of different types of measurement effects, and the extent to which these differences are observed in SES, However, the report is primarily descriptive in nature looking variable by variable. It does not provide in-depth analysis of which types of variables are most prone to measurement effects or conclusions which can necessarily be generalised to other surveys.

The descriptive findings, together with the accompanying annex tables, can help to inform decisions about the feasibility of moving the SES online and any potential risks to the survey timeseries of such a move. The findings will also be valuable as a reference point for users of the SES 2024 data who may be interested in conducting analysis using the combined F2F and Panel datasets to maximise sample sizes. The report and tables will serve as a guide to which variables exhibit differences to a lesser or greater degree and which might therefore be more

or less suitable for analysing with a combined dataset. While this report examines the differences between the two surveys it does not draw any conclusions as to which is better. This may vary depending on the variables of interest and the purpose of the analysis. Data users are advised to conduct their own sensitivity analysis regarding the effect on their results of including data from the online survey in their analysis, especially if interested in analysing change over time.

The rest of the report proceeds as follows. The next chapter gives an overview of how data collection for the two SES 2024 surveys was carried out and how the data have been analysed. Chapter 3 presents headline descriptive results from the two surveys and summarises the extent to which differences between the two were observed. Chapters 4 and 5 are concerned with trying to understand what might explain those differences described in Chapter 3. Chapter 4 looks at whether there are differences in sample composition between the F2F and Panel surveys and the extent to which, if at all, controlling for these differences in sample composition accounts for differences observed. Chapter 5 goes on to consider the extent to which any differences which persist after controlling for differences in sample composition might be due to the presence of measurement differences between the two surveys. Chapter 6 looks in more detail at some specific variables which are central to the SES survey - variables recording pay and the occupation and industry in which the respondent works - and how these varied between the two survey designs.

2. Methodology

This section summarises the data collection methodology used for the Skills and Employment Survey (SES) 2024 face to face (F2F) survey and the online data collection via the NatCen Opinion Panel. Further details of the survey methodology can be found in the SES 2024 Technical Report. As discussed below, careful attempts were made throughout the process to harmonise the data collection methodology as far as possible, including using a single unimode questionnaire. However, differences necessarily remain.

This section also summarises some key features of the analysis conducted to compare results from the two surveys.

The 2024 survey was funded by the Economic and Social Research Council (ESRC), the Advisory, Conciliation and Arbitration Service (Acas) and the Department for Education. Additional funding was provided by the Department for the Economy in Northern Ireland to facilitate the online data collection being extended to Northern Ireland.

The project was hosted by Cardiff University (Principal Investigator Alan Felstead from Cardiff University and Wales Institute for Social and Economic Research and Data (WISERD)) with Co-Investigators Rhys Davies (also Cardiff University and WISERD), Duncan Gallie (University of Oxford), Francis Green and Golo Henseke (Institute of Education, UCL), Ying Zhou (Surrey University), and Martin Wood, Curtis Jessop and Joanna D'Ardenne (National Centre for Social Research). The National Centre for Social Research (NatCen) carried out the 2024 fieldwork.

2.1 Data collection

2.1.1 Sampling

Both the F2F and Panel surveys used a random probability sample of individuals aged 20 to 65 and in paid work. However, the sample was drawn from different sources.

The F2F survey started with a random sample of addresses in Great Britain drawn from the Royal Mail Postcode Address File. Interviewers visited the addresses and attempted to determine whether anyone meeting the SES eligibility criteria (aged 20-65 and having done one hour or more of paid work in the last seven days) was resident at the address. If one or more eligible adults was found at an address, up to two were selected (at random) for the survey.

The web-telephone survey was conducted using the NatCen Opinion Panel. The Panel is a probability-based online panel run by NatCen and consists of a nationally representative sample of individuals in Great Britain recruited from the British Social Attitudes Survey, who are invited to compete short web surveys on a range of different topics at regular intervals. For SES, a random sample of 4,000 individual Panel respondents aged 20-

65 was invited to participate. These individuals were screened for their current employment status at the start of the interview and only those in paid work invited to continue. Online data collection was also carried out in Northern Ireland. However, this report focuses on Great Britain only for comparability with the F2F survey.

2.1.2 Questionnaire

The same questionnaire was fielded F2F and via the NatCen Opinion Panel. The survey covered the following topics in addition to checking the respondent's eligibility and collecting basic demographics:

- BLOCK B: Broad questions about the current job including hours worked, workplace location, etc.
- BLOCK C: Detailed job analysis questions, including the skills necessary to do the job
- BLOCK D: Qualifications
- BLOCK F: Work attitudes
- BLOCK E: Attitudes towards the organisation worked for
- BLOCK G: Pay questions
- BLOCK J: Recent skill changes and future perspectives
- BLOCK I: Well-being at work

Two of the modules (C and I) were completed self-completion in the F2F survey and may, therefore, be less prone to differences due to measurement effects relative to the Panel. Most respondents (83%) did complete these modules themselves using Computer Assisted Self Interviewing (CASI). The remainder of respondents completed one or both with the interviewer.

Prior to fieldwork, the questionnaire was reviewed by NatCen's Questionnaire Design and Testing Hub who identified questions considered to be particularly at risk of measurement differences between modes. Where possible, adaptations were made to the layout and presentation of questions to try and minimise any differences and ensure consistency in measurement across modes. However, any changes were light touch so as to avoid disrupting the timeseries. The main adjustments made were:

- Shortening long question stems/introductions where this would not materially affect the question.
- Rewording interviewer instructions to be respondent facing in the web survey and using help screens to provide additional information/definitions.
- Agreeing a format for presenting batteries of questions – and split these batteries into shorter sets of questions – for ease of administration face to face and online.
- The qualifications showcard/response options were reorganised and split up with the use of subheading to make them easier to navigate.

Some of the key questions, on qualifications and job type for example, underwent online usability testing prior to the launch of mainstage fieldwork.

The main difference between the F2F and online administration of the survey was that the online questionnaire was split up into two shorter surveys which respondents were invited to complete a few weeks apart rather than a single interview. Wave 1 of the survey covered Modules B-D and Wave 2 covered Modules F-I. This was to try and avoid break offs occurring or fatigue resulting in low data quality during a long online interview.

2.1.3 Fieldwork and response rates

F2F fieldwork was conducted between September 2023 and July 2024 by trained NatCen interviewers who had been briefed by the research team. In total 2,824 individual productive interviews were obtained in 2,287 households from a starting sample of 21,475 addresses (Table 2.1).

The 2,824 individual interviews were obtained from a pool of 8,799 eligible individuals identified. That represents an individual response rate of 32% (Table 2.2). The response rate was considerably lower than in previous waves of SES and certain groups, for example young people and those with no qualifications, were underrepresented. Weighting was used to correct for this underrepresentation as far as possible (see below).

Table 2.1 Face to Face household response rate

Outcome category	**Number**	**%**	**%**	**%**	**%**
Issued addresses	21,475	100			
Deadwood:	1,274	6			
In scope addresses	20,201	94	100		
Not screened:	7,709		38		
Screened	12,492		62	100	
No eligible individuals	6,746			54	
At least one eligible individual	5,746			46	100
At least one productive interview in household	2,287				**40**

Table 2.2 Face to face individual response rate

	Number	% of eligible
Eligible households	5,746	
Eligible individuals	8,799	
Productive households	2,287	40%
Productive individual interviews	2,824	32%

Panel fieldwork took place between 29th September and 26th November 2023 (Wave 1) and 19th October and 11th December 2023 (Wave 2). Panellists were given the option to complete the survey over the telephone if they preferred. Only a small number (103) did so.

Table 2.3 shows the proportion of individuals who were invited to take part in SES2024 who did so at Wave 1 and Wave 2. Overall, the Panel survey delivered 1,892 fully productive interviews, that is people who completed both waves of the survey, in Great Britain. 64% of people cooperated with the initial request to complete SES Wave 1, of whom 91% then went on to complete Wave 2

Table 2.3 Panel survey SES response rate

	Wave 1	**Wave 2**

Issued	4,000	2,069
Deadwood[1]	2	0
Achieved	2,069	1,892
Screened out	495	0
Survey response rate [2]	64%	91%

Looking just at the rate of cooperation with the SES survey request (64% in GB at Wave 1), the online survey performed better than the face to face survey (where the cooperation rate among eligible individuals was 32%). However, these two figures are not directly comparable. It should be borne in mind that there are other stages of non-response among the panellist that are not represented in these response rates. This includes non-response to the initial recruitment survey and subsequent attrition from the Panel before the SES sample was drawn. The cumulative response rate for SES online, taking the original recruitment sample as the starting point, is estimated to be around 5% (see Table 2.4 below). That is considerably lower than the F2F survey (32%).

Table 2.4 Cumulative response rate for SES GB (Panel)

Recruitment survey issued[3]	226,642
Recruitment survey in scope	204,254
Recruitment survey productive	35,608
Recruited to Panel	24,623
Recruitment survey response rate	17%
Panel recruitment rate	69%
Active panel members at time of SES (after accounting for attrition)	18,241
Panel recruitment rate (as proportion of recruitment survey sample in scope)[4]	9%
Wave 1 response rate	64%
Wave 2 response rate	91%
Cumulative response rate to SES (estimated)	5%

However, we note that there is not a linear or straightforward relationship between lower response rates and non-response bias.[5] Furthermore, where sampling was done via the Panel, it was possible to adjust for some degree bias at the sampling stage. Over-sampling those with characteristics known to be less likely to take part leads to lower survey response rates, while increasing the sample's representativeness on those measures, reducing the efficacy of the 'cumulative response rate' as a measure of bias further.

[1] 2 cases were found to be out of age range post-sampling
[2] Survey response rate counts cases screened out as productive. Survey response rate is provided as a measure of bias/propensity to respond to the survey request.
[3] British Social Attitudes 2015 - 2022
[4] All adults 18+
[5] See, for example, ttps://ojs.ub.uni-konstanz.de/srm/article/view/8475

These same data can also be used in non-response weighting post-fieldwork. Comparing the pre-calibration weighting distribution of the Panel sample against LFS estimates, there is less evidence of bias in terms of age, region and education than was found in the F2F survey, despite the lower response rates (see Chapter 4).

2.1.4 Questionnaire length

The time taken to complete the SES interview was similar in both surveys. Interview length for Panel respondents was calculated separately for each wave. The median length of Wave 1 was 26 minutes. The median length of Wave 2 was 20 minutes. The equivalent timing for F2F modules A to D (Wave 1 Panel) was 27 minutes and for Modules F to K (Wave 2 Panel plus some demographics not asked in the Panel) was 25 minutes.[6]

2.1.5 Weighting

Weights were computed to account for differential selection probabilities and, in the case of the Panel survey, non-response or attrition at previous stages in the Panel lifecycle. The resulting weights were calibrated so that the weighted profile of respondents matched that of working adults aged 20 to 65 according to Labour Force Estimates; the same approach was used for both surveys. The following calibration targets were used: 5-year age bands nested within sex, occupation (1-digit SOC 2020), Government Office Region, highest educational qualification and ethnicity.

2.2 Data analysis

The analysis compares the estimates obtained from the F2F survey with those obtained from NatCen Panel respondents in Great Britain. The Panel estimates include both online respondents and the minority of respondents (n=105) who completed the survey on the telephone. The purpose of the analysis is to compare data collected via the two alternative survey methodologies i.e. a fresh sample of respondents interviewed F2F vs an online panel with a telephone option. The number of telephone respondents is too small for this group to be analysed separately, and it would be particularly difficult to disentangle differences resulting from measurement effects and sample composition.

Estimates were compared for 312 substantive variables. All variables were categorical. In total 1,242 survey estimates or proportions were compared. This represents nearly all the SES survey variables. Exceptions are where variables were intended primarily for routing or survey administration or were used to derive demographic variables such as highest qualification held. SES contains a small number of continuous variables, for example to collect data on number of people in the workplace. These have been banded and treated as categorical for the purposes of this analysis. It was sometimes necessary to collapse and combine two categories where the number of respondents in a category was too small to enable analysis to be run. Where this has been done it is indicated in the accompanying tables by a variable name in the format “Varname_No”.

In summarising the number and type of variables where differences are observed, the presence of a difference is assessed using two criteria – whether the difference in estimates between surveys is statistically significant at the 5% level and whether the magnitude of the difference is above/below a certain threshold. Both criteria have their limitations. Whether or not a difference is statistically significant will, for example, be influenced by the sample size (which, on SES, was large at 2,000 plus cases for each survey) and the nature of the estimate (differences in percentages at the extremes of the distribution e.g. less than 10% or above 90% are more likely to be statistically significant than differences in percentages closer to 50%). Given the large number of variables

[6] The estimates exclude people who started and finished the survey on different days.

and response categories compared and tested for statistical significance, differences will be found to be statistically significant purely by chance.[7] Combining a test of statistical significance with looking at the substantive magnitude of any differences between surveys attempts to overcome these limitations. However, the choice of threshold for identifying a substantive difference is itself arbitrary and findings will be sensitive to the choice of threshold.

Setting thresholds for identifying an effect was necessary to allow us to summarise and discuss the results of testing differences across hundreds of variables. However, data users are encouraged to examine the data and make their own judgements regarding what they consider to be an acceptable level of variation between surveys depending on the type of analysis they wish to conduct.

All analysis is weighted by the final calibration weights. That means that some differences in sample composition, for example with respect to age and education, between the two surveys will already have been controlled for.

Analysis is focussed on 'point estimates' only. Exploring differences in the relationships between variables across the two survey designs is beyond the scope of this report.

While the parallel run provides an opportunity to explore differences in estimates between the two designs, SES did not have an experimental design and was not intended primarily as a methodological study. While attempts are made to provide explanations for differences observed (by, for example, controlling for differences in sample composition), it is not possible to fully isolate or explain the reasons for any differences observed. As well as differences in sample composition and/or measurement effects arising from self-completion vs interview modes for example, other differences in study design - for example the fact that one study uses a fresh sample and the other is a Panel, differences in fieldwork timings, or the fact that the Panel involved some telephone interviews – may also play a role in differences between the two studies.

[7] Significance tests have not been adjusted for multiple testing.

3. Differences between modes

The first stage of the analysis was to compare the estimates obtained via the SES face to face (F2F) and NatCen Opinion Panel surveys and summarise for which survey estimates differences were found. A full set of results is provided in the accompanying Excel file: “SES descriptives.xls”.

3.1 Don’t know and refusals

One way in which mode of survey administration may affect estimates is if the propensity for people to give a ‘Don’t know’ or ‘Refusal’ response – and therefore the level of missing data – varies by mode. The extent to which differences in the level of missing data are likely to arise between modes will depend in part on how similarly, or otherwise, missing data are treated. In both the SES F2F and Panel surveys ‘Don’t know” and “Refusal’ options were hidden – that is they were not read out by the interviewer (F2F/telephone), did not appear on showcards (F2F) or appear on screen. Respondents were able to say ‘Don’t know” or refuse at any question; this response would be coded by the interviewer or appear on screen when online respondents attempted to skip a question without answering. Given the similarity in the treatment of missing data across the two SES surveys any differences are expected to be minimal. Where they do occur, it is anticipated that the level of missing data would be lower in the Panel. This is because the online data collection avoids any potential reluctance over providing sensitive information to an interviewer.

The level of ‘Don’t know’ and ‘Refusal’ responses was very low across both surveys. There were only 21 variables where the proportion of missing data in one or other of the surveys was above one percent (7% of all variables) and four variables where the proportion was above five percentage points. There were eight variables where the proportion of ‘Don’t Know’ responses differed by more than one percentage point between the two surveys. The face-to-face estimate was higher than the Panel estimate for seven of these eight variables.

Most of the variables where the proportion of ‘Don’t know’ responses was relatively high were in Module E, in which respondents answered questions relating to the organisation they work for. For face-to-face respondents, 12 of 21 variables where the proportion of ‘Don’t Know’ responses exceeded one percent were in Module E. For Panel respondents, this was the case for eight of 11 variables affected. This suggests that respondents may have found it more difficult to answer questions about their employer or organisation, compared with answering questions related to their own role in the workplace.

Table 1: Level of missing data by survey mode

	F2F	Panel
% Don't Know (Mean)	0.2%	0.1%
% Refusal (Mean)	0%	0%
Number of variables where DK>=1 pp	21	11
Number of variables where DK>=5 pp	4	1
Number of variables where difference in DK between modes >=1 pp	8	

Given the relatively low level of missing data across the two surveys, it is reasonable to ignore missing values in all subsequent analysis and focus on comparing substantive responses only.

3.2 Effect on substantive survey estimates: Overview by questionnaire module

To explore the extent of differences between survey estimates at the variable level, an average of the differences observed in the estimates for all responses categories for that variable was calculated for each variable. The average absolute difference across the variable – that is ignoring the directionality of any differences - was calculated. So, for example, if a question had 3 categories and the difference between the F2F and the Panel Survey was +2pp for the first category, -4pp for the second and +2pp for the section, the average absolute difference would be 8/3 or 2.7pp.

Table 3.2 summarises the average of those average absolute differences (in percentage points) by questionnaire module. The average difference per variable was relatively low at 3.5 percentage points. There were relatively few variables which exhibited large differences between modes, with 21% showing an average difference of 5 percentage points or more and only 2% an average difference of 10 percentage points or more.

By way of summarising the information, the table also shows a breakdown of results separately for each questionnaire module. All questionnaire modules show evidence of mode differences. It is difficult to draw firm conclusions about the nature and extent of mode differences comparing across modules as, while the modules cover different topics, each module contains a wide mix of different types of questions – including for example a mix of behavioural and attitudinal questions and simple and more complex items – which we might expect to be more or less prone to measurement effects between modes (see Chapter 5 for more on this). One thing to note from Table 3.2 however is that, with the exception of Module G (which has a very small number of items), the two modules with the lowest level of differences of 3 percentage points or more between the surveys are the two self-completion modules, C and I. With these modules being self-completion across both surveys, we would anticipate fewer measurement differences here compared with other modules (again see Chapter 5 for a more detailed discussion).

Table 3.2 Summary of mode differences at variable level

	Number of variables	Average difference by mode (percentage points)	N/% variables with average difference >3pp	N/% variables with average difference >5pp	N/% variables with average difference >10pp
Overall	312	3.5	161 (52%)	67 (21%)	7 (2%)
B – Broad Job Characteristics	66	4.0	39 (59%)	18 (27%)	2 (3%)

C – Job Analysis (self-completion)	58	2.4	17 (29%)	4 (7%)	0 (-)
D – Skills Use and Qualifications	8	3.7	4 (50%)	3 (38%)	0 (-)
F – Work Attitudes	20	3.2	11 (55%)	2 (10%)	1 (5%)
E – Organisation	49	4.7	35 (71%)	23 (47%)	2 (4%)
G – Pay	4	2.3	1 (25%)	0 (-)	0 (-)
J – Recent Changes to the Job and Training Experiences	51	3.7	29 (57%)	13 (25%)	2 (4%)
I[8] - Well-being at Work (self-completion)	56	2.9	25 (45%)	4 (7%)	0 (-)

Table 3.3 repeats the analysis from Table 3.2 but looking at the absolute difference in responses to individual response categories rather than variable by variable. Again, figures are presented for each module separately. At the category level, as well as identifying mode differences of various magnitudes, categories are counted as displaying a mode effect if the difference between modes was statistically significant at the 5% level.

We would expect to find statistically significant differences in around 5% of estimates, by definition. In fact, we find statistically significant differences – along with a substantive difference of 3 percentage points or more - in 35% of estimates overall. This suggests that there are mode differences in SES.

However, as discussed in Chapter 2, it should be noted that the choice of threshold and significance level for identifying a mode effect as notable is arbitrary to some extent. The presence of these relatively small differences by mode need not necessarily be cause for concern. SES data users are encouraged to decide on a case by case basis, depending on the variables used and the purpose of the analysis, whether estimates from the two surveys are sufficiently similar to allow the F2F and Panel data to be combined.

For full details of the results of the mode comparisons see the “Mode differences.xls” tab in the accompanying annex.

[8]Modules are shown in the order in which they were asked in the questionnaire. Module A was just the initial screening questions. There was no module H. Apart from one question, Module K was demographic variables only.

Table 3.3 Summary of mode differences at category level

	Number of categories	% categories with statistically significant difference >3pp	% categories with statistically significant difference >5pp	% categories with statistically significant difference >10pp
Overall	1,242	35.1%	20.9%	4.5%
B – Broad Job Characteristics	250	34.8%	22.8%	6.4%
C – Job Analysis (self-completion)	261	23.0%	9.2%	0.8%
D – Skills Use and Qualifications	38	34.2%	23.7%	2.6%
F – Work Attitudes	74	36.5%	27.0%	2.7%
E – Organisation	164	49.3%	37.8%	13.4%
G – Pay	8	25%	-	-
J – Recent Changes to the Job and Training Experiences	130	33.1%	22.3%	6.2%
I[9] - Well-being at Work (self-completion)	317	38.8%	18.3%	0.2%

[9]Modules are shown in the order in which they were asked in the questionnaire. Module A was just the initial screening questions. There was no module H. Apart from one question, Module K was demographic variables only.

4. Effect of differences in sample composition

This section examines the extent to which the differences in estimates observed between the two surveys described in the previous section may be the result of differences in sample composition, that is differences in the types of people who responded to each survey.

4.1 Characteristics of Face to Face and NatCen Opinion Panel respondents

There were some notable differences observed in the composition of the Face to Face (F2F) and Panel achieved samples prior to the final calibration weighting being applied. Compared to the population of 20-65 year olds in paid work, the F2F survey attracted more female respondents and people educated degree level or above, but fewer younger respondents and people in certain regions, for example London and Scotland (see Table 4.1). These biases – which were less apparent in the Panel - reflect established patterns in differential non-response to surveys as well as specific challenges for SES 2024 in survey coverage.[10] However, also as shown in Table 4.1, the application of calibration weights all but eliminates the difference between the two surveys with respect to: age, sex, occupation, region, education and ethnicity.

[10] The fact that bigger biases were observed in the F2F estimates compared with the Panel estimates is likely the result of the fieldwork challenges experienced on the F2F SES 2024 as well as the fact that the sample design and pre-calibration weighting applied on the Panel takes account of known biases where possible (see Chapter 2 and SES 2024 Technical Report for details).

Table 4.1 Composition of achieved sample pre and post-calibration weighting, relative to LFS population estimates

		F2F		Panel Wave 1	
	Population	Pre-calibration	Post-calibration	Pre-calibration	Post-calibration
	%	%	%	%	%
Age by sex					
Males 20-29	9.9	6.1	9.8	8.9	9.9
Males 30-39	13.1	11.7	13.2	13.1	13.1
Males 40-49	11.8	12.2	11.8	12.3	11.8
Males 50-60	12.7	12.2	12.7	14.1	12.7
Males 61-65	3.6	4.0	3.6	3.9	3.6
Females 20-29	9.8	6.9	9.8	8.6	9.8
Females 30-39	12.6	13.7	12.6	11.4	12.6
Females 40-49	11.4	14.7	11.4	11.6	11.4
Females 50-60	11.9	14.1	12.0	12.6	11.9
Females 61-65	3.1	4.4	3.1	3.7	3.1
SOC 2020 Occupation Groups					
Managers, Directors And Senior Officials	11.1	11.0	11.1	9.6	11.1
Professional Occupations	27.6	30.0	27.6	26.2	27.6
Associate Professional occupations	15.7	15.8	15.7	16.5	15.7
Administrative And Secretarial Occupations	9.3	8.8	9.4	11.0	9.3
Skilled Trades Occupations	8.6	7.2	8.5	7.1	8.6
Caring, Leisure And Other Service Occupations	8.3	9.0	8.3	9.3	8.3
Sales And Customer Service Occupations	5.7	5.6	5.7	7.7	5.7
Process, Plant And Machine Operatives	5.5	4.7	5.5	5.2	5.5
Elementary Occupations	8.3	7.9	8.3	7.5	8.3
Region					
North East	3.8	2.7	3.8	3.8	3.8
North West	11.0	9.2	11.0	10.9	11.0
Yorkshire and The Humber	8.2	9.8	8.2	8.1	8.2
East Midlands	7.3	6.6	7.4	7.4	7.3
West Midlands	8.9	8.1	8.9	9.2	8.9
East of England	9.7	12.1	9.8	9.8	9.7
London	15.2	9.5	15.1	15.6	15.2
South East	14.6	21.2	14.6	14.6	14.6
South West	8.7	10.1	8.7	8.1	8.7
Scotland	8.2	5.7	8.2	8.2	8.2
Wales	4.3	4.9	4.4	4.2	4.3

		F2F		Panel Wave 1	
	Population	Pre-calibration	Post-calibration	Pre-calibration	Post-calibration
	%	%	%	%	%
Highest Qualification Held					
No Qualifications	3.5	3.3	3.5	7.4	3.5
Level 1	6.1	7.1	6.1	6.2	6.1
Level 2	15.4	13.3	15.4	13.7	15.4
Level 3	21.4	16.9	21.4	20.7	21.4
Level 4 and above	53.5	59.4	53.6	52.0	53.5
Ethnicity					
White	83.2	82.9	83.2	85.5	83.2
BAME	16.8	17.1	16.8	14.5	16.8

The focus here is on the potential for other differences in the demographic or job characteristics of respondents, not accounted for in the weighting, that could potentially affect the final weighted survey estimates. Table 4.2 shows that there was very little difference between the two surveys with respect to demographic characteristics such as marital status, the presence of children in the household or whether the respondent had a long-term limiting illness, factors which might reasonably be thought to influence people's attitudes towards and experiences of employment. Similarly, Table 4.3 shows only small differences in the reported job characteristics of respondents, such as whether someone is an employee or self-employed or works full- or part-time, across the two surveys. It is therefore unlikely that controlling for these variables will have a significant effect on any differences observed. Nevertheless, this hypothesis is tested below.

Table 4.2 Demographic characteristics of weighted F2F and Panel respondents

	F2F	Panel
Marital status		
Married/Civil partnership	51	52
Living with partner	19	21
Single	23	21
Widowed/separated/divorced	7	6
Children under 16		
Yes	37	37
No	63	63
Children under 5		
Yes	14	14
No	86	86
Long term limiting health condition		
No health condition	76	75
Yes, limits a lot	3	3
Yes, limits a little*	12	14
Yes, does not limit	9	7
Minimum N	2,819	1,891

*Difference between surveys significant at p<0.05

Table 4.3 Job characteristics of weighted F2F and Panel respondents

	F2F	Panel
Employment status		
Permanent employee	84	84
Temporary employee	6	7
Self-employed *	10	8
Full time vs part time		
Full time	79	76
Part time	21	24
Whether work from home		
Yes	27	27
No	73	73
Time with current employer		
< a year	13	12
1-2 years	23	24
3-5 years	19	19
6-10 years	18	19
11+ years	27	25
Importance of computer in job		
Essential	67	69
Very important	10	10
Fairly important	9	10
Not very important	6	5
Not at all important/Does not apply	8	7
Whether manage others (Employees only)		
Yes supervise others	24	26
Yes have managerial duties	15	17
Neither	61	57
Number of people at workplace		
1 to 2	12	10
3 to 24	24	23
25 to 99 *	22	26
100 to 499	21	24
500+*	22	17
Minimum N	2,445	1,869

*Difference between surveys significant at p<0.05

4.2 Differences after controlling for differences in sample composition

Multiple regression analysis was used to isolate the effect of mode on survey estimates after controlling for differences in observed demographic and job characteristics. Multinomial logistic regression was used given the categorical nature of the survey variables. This allowed the effect of mode on the proportion of respondents selecting a specific response category to be estimated. Separate models were run for each survey variable in turn with mode (1= F2F 0=Panel) as a predictor. The models were run first without and then with controls and

the size and statistical significance of the mode coefficient compared for the two models. The following control variables were included (categories as per the tables above):

- Age (nested with sex)
- One digit SOC
- Highest qualification held
- Region
- Ethnicity
- Marital status
- Whether have children under 16 (yes/no)
- Whether have children under 5 (yes/no)
- Whether have long-term limiting illness
- Whether employee on permanent/temporary contract or self-employed
- Whether work full-time or part-time
- Whether work at home or away from home
- How long worked for current employer
- How important computer is in work
- Whether respondent completed the self-completion modules themselves or via an interviewer.[11]

The default reference category was the first response option given to respondents. In a few cases, where the proportion of responses in the first category was low, the reference category was changed to be the modal category. The first category (usually, though not always, the most positive response where a scale was used) was chosen as the default given that one predicted difference in response by mode is a tendency towards positivity bias in F2F respondents (see Chapter 5). Comparing the proportion of responses to other categories relative to the first category by mode provides a test of this positivity bias. In some cases, it was necessary to collapse response categories with small numbers of cases to enable the regression models to run. Where this was necessary it is indicated on the results spreadsheet.

For full details of the results of the regression results see the “Controlling for sample composition.xls” tab in the accompanying annex.

Table 4.4 shows the number of variables for which differences are observed, first in the absence of any controls for differences in sample composition and, second, after controlling for differences in demographics and job characteristics between the F2F and Panel samples. A mode effect for a variable was judged to be present if both conditions below were met when comparing at least one of the response categories to the reference category.

- The coefficient for the mode variable was statistically significant at the 5% level
- The odds of a Panel respondent selecting category A over category B was substantively higher or lower than the odds of a F2F respondent doing so.[12] The table indicates the number of variables for which the odds ratio was less than or equal to 0.8 or greater than or equal to 1.2, that is where the difference in the

[11] Panel respondents completing over the telephone were treated as “not self-completion”. Sensitivity analysis was run to ensure that the inclusion of this variable (which is correlated with the mode variable) did not have a large impact on the size of the mode effect.
[12] An odds ratio of less (more) than 1 means that the odds of a Panel respondent picking Category A over Category B are lower (higherr) than the odds of a F2F respondent picking Category A over Category B. For example, an odds ratio of 1.1 when modelling the likelihood of saying “agree” rather than “disagree” means that Panel respondents were 10% less likely to select “agree” over “strongly agree” compared with F2F respondents. That does not necessarily mean that the proportion of Panel respondents selecting “agree” was lower than the proportion of F2F respondents selecting “agree”, just that the propensity of Panel respondents to pick “agree” over “strongly agree” waslower.

odds of a Panel respondent compared with a F2F respondent selecting category A over category B was 20% or more.

More variables are flagged as having notable mode effects using this methodology and thresholds than was the case in Chapter 3 when comparing differences in percentages. The important thing to focus on here is not the total number of variables/categories where a mode effect is flagged, but how this number changes after controlling for differences in sample composition.

Table 4.4 Number of variables for which mode effects present, with and without controls for sample composition

	Without controls for sample composition		With controls for sample composition	
	N	%	N	%
Total variables	312		312	
1+ category comparison where mode effect statistically significant	247	79%	242	78%
1+ category comparison where odds ratio <=0.8 / >=1.2	249	80%	254	81%
1+ category comparison where mode effect significant and odds ratio <= 0.8/>=1.2	233	75%	231	74%

Table 4.5 Number of category comparisons for which mode effects present, with and without controls for sample composition

	Without controls for sample composition		With controls for sample composition	
	N	%	N	%
Total category comparisons	943		943	
Category comparisons where mode effect statistically significant	580	62%	584	62%
Category comparisons where odds ratio <=0.8 / >=1.2	622	66%	636	67%
Category comparisons where mode effect significant and odds ratio <=0.8/>=1.2	545	58%	551	58%

The figures presented in Table 4.4 and 4.5 – whether at variable or category level – suggest that controlling for differences in sample composition makes little difference to the total number of variables displaying evidence of differences by survey mode. This suggests that (observed) differences in sample composition do not fully account for the differences observed between modes.

At the same time, however, we cannot necessarily conclude that sample composition has no effect on differences observed between the two survey designs. While the total number of variables displaying mode effects may be very similar under both models, it may not necessarily be the same variables which are identified

in both cases. Controlling for differences in sample composition may lead to some variables no longer displaying mode effects but may lead to others exhibiting mode effects (if, for example, the effect of sample composition and measurement effects were previously cancelling one another out and – after controlling for sample differences - the measurement effects become apparent). There may, therefore, be evidence of a sample composition effect even if the total number of variables exhibiting mode effects after controlling for sample differences remains the same.

Table 4.6 presents data on the number of variables where the extent of any mode effect observed changes after controlling for differences in sample composition. It identifies variables where at least one odds ratio changed by 0.1 or more – that is the odds of respondents in one mode versus the other choosing Category A over Category B changed by 10% or more – as well as variables where the mode effect changed from being statistically significant to not significant (or vice versa) with the addition of controls.

Around a quarter of variables saw a change in the significance or size of the mode effect after controlling for differences in sample composition. Again, this suggests that sample composition is, at best, only a partial explanation for the differences – or lack of difference for some variables - observed by mode. This is consistent with the lack of evidence for difference in key demographic and job variables when comparing between modes using weighted data (see Tables 4.2 and 4.3).

It should be noted, however, that this analysis controls only for a limited number of observed characteristics. We cannot rule out that mode differences may be being driven by unobserved differences in the characteristics and profile of respondents to the F2F and Panel surveys These differences could, in turn, account for any mode differences observed. In particular, the fact that the sample for the online/telephone data collection was drawn from a panel study could mean that panel conditioning - that is the idea that repeated survey participation may affect how people respond in surveys - has an effect.[13]

Table 4.6: Effect of controlling for differences in sample composition on any mode effect observed

	N	% of total
All variables	312	
1+ category comparison where mode significant - > not significant with controls (at 5% level)	34	11%
1+ category comparison where mode not significant - > significant with controls (at 5% level)	33	11%
1+ category comparison where odds ratio changed by +/- 0.1	80	26%
1+ category comparison where odds ratio changed by +/- 0.2	17	5%

[13] The number of prior waves of the Panel people had been invited to participate in/had completed varied (as different data collections have different eligibility requirements). The median number of waves of the Panel people had been invited to complete was four, while the median number they had completed was three.

5. Exploring potential measurement effects

Where differences in survey estimates remain after controlling for differences in sample composition, it may be the result of measurement effects, that is due to differences in the way in which people react to the same question when presented in different modes. This section assesses the extent to which the pattern of differences in survey estimates observed between modes - and which continue to be observed after controlling for differences in sample composition - is consistent with the measurement effects that might be anticipated.

Measurement effects due to difference in mode can occur for many reasons. For example, in the absence of an interviewer, people responding online may feel less pressure to give the more socially desirable response or respond positively to a question. Complex questions, including those which contain additional instructions on how to answer, may elicit different responses online if, in the absence of an interviewer, web respondents are more prone to satisficing behaviour. Finally, the way in which questions are presented – either visually or orally – may affect how people respond.

Ahead of fieldwork, NatCen researchers reviewed the SES questions and categorised them according to the risk of different measurement effects. The list of measurement effects coded for is listed below. Not all of these risks (e.g. those related to ranking or open questions) were present in the SES questions assessed here. Furthermore, where possible, attempts were made to mitigate against these risks when developing the questionnaire for both face to face and online administration. Nevertheless, it is anticipated that measurement effects would still be present. This is, in part, because the extent to which the questionnaire was adapted for web administration was limited given the importance of retaining the face to face timeseries. In addition, even with mitigating measures in place, some measurement effects are likely to remain. For example, to mitigate against the risk surrounding complex questions, instructions or definitions the original questionnaire were rephrased and simplified to make them more respondent friendly - and made available to web respondents via help buttons. However, any mitigation of measurement effects is likely to be partial at best because not all web respondents will read the instructions. The types of measurement effects most likely to affect SES, even after work to try and reduce measurement effects when developing the questionnaire, are shown in bold in the list below.

Figure 5.1:Potential measurement effects for which the SES questionnaire was considered

Social Desirability

- **A1. Fear of disclosure and socially desirable reporting**
- **A2. Positivity bias on rating scales**

Satisficing

- **B1. Complex question stem or clarifications**
- **B2. Extra information**
- **B3. Computation**
- B4. Open questions

Presentation effects

- **C1. High number of response options**
- **C2. Batteries of scale**
- C3. Hidden codes and interviewer-coded items
- C4. Ranking tasks
- C5. Non-standard question format or visual aid

For full details of how the SES variables were coded for measurement effects see “Risk of measurement effects.xls” tab in the accompanying annex.

To evaluate the extent to which differences in survey estimates observed are likely to be the result of measurement effects, a comparison was made between a) variables where, after controlling for differences in sample composition, differences continued to be observed b) variables where measurement effects were anticipated prior to fieldwork. The higher the degree of correspondence between those variables where measurement effects were (or were not) expected and variables where differences were (or were not) observed, the stronger the evidence for measurement effects being present.

5.1 Were differences by mode observed where expected?

Table 5.1 below summarises the number of SES items considered to be at risk from different types of measurement effect and the number of “at risk” variables for which sizeable measurement effects were observed. It also shows the number of variables for which mode effects were not anticipated but were, nevertheless, observed. A variable was counted as showing evidence of a mode difference if, in the multinomial regression controlling for demographic and job characteristics (see Chapter 4), there was at least one category comparison for which a statistically significant mode effect remained at the 5% level, AND the difference between modes in the likelihood of selecting that category over the reference was at least 20% (that is the odds ratio was less than or equal to 0.8 or more than or equal to 1.2) .

Table 5.1: Expected vs. observed measurement effects, by type of measurement effect

	Number of variables identified	Number where mode effect observed	Percentage where mode effect observed
Risk of social desirability	185	149	81%
Risk of satisficing	55	39	71%
Risk of presentation effects	164	131	80%
No risk	86	42	49%

NB: A variable may have been assessed as at risk from more than one type of measurement effect.

Notable mode effects were found for a high proportion of variables where mode effects were expected. We would not necessarily expect to observe differences for all variables where measurement effects were predicted. First, the predictions only identify a potential risk of measurement effects. Second, while we have controlled for observed differences in sample composition there may be other unobserved differences in same composition that are offsetting potential measurement effects - or multiple measurement effects acting in different directions - leading to no overall difference by mode. What is apparent is that it was more likely to observe measurement effects in variables where they were anticipated than where they weren't. The proportion of false positives – that is variables where a measurement effect was observed but no risk was predicted – is lower (49%) than the proportion of observed differences (71 – 81%) where differences were predicted.[14] This suggests that measurement effects are present and can explain at least some of the differences observed by mode.

Mode effects were observed in the two questionnaire modules which were self-completed by face to face respondents and which we might, therefore, expect to be less prone to measurement effect than questions in other modules.(The risk of measurement effects was predicted assuming one survey would be completed F2F and the other (the Panel) self-completion). Of the 107 variables across the two self-completion modules (C and I) which were identified as being at risk of measurement effects, 84 (79%) continued to exhibit notable mode effects even after controlling for differences in sample composition. It is possible that, even though the majority of F2F respondents (83%) completed the questions in Modules C and I themselves, the presence of an interviewer in the room nevertheless had an effect on their responses relative to a survey that was fully self-completion. It may also be the case that the measurement effects present for the 20% of the sample who did complete the questions with the interviewer are of sufficient magnitude to be present overall (the questions were considered sufficiently at risk of social desirability bias to be incorporated into self-completion modules).

It is possible to break down the high-level measurement effects described above and consider them at a more granular level. This is done in Table 5.2. However, there is little evidence of certain mode effects being more or less common in practice. The proportion of variables where a mode effect was expected and one was observed is very similar across all types of mode effect (the proportions were lower for B2 and B3 effects, but the number of variables identified as at risk of these measurement effects was small). One challenge with interpreting the results for different types of measurement effect is that the same variable may be vulnerable to more than one

[14] The false positives are themselves not entirely inconsistent with measurement effects. For example, A number of the false positives occurred for yes/no questions in Module E which asked about people's attitudes towards the organisation for which they worked (e.g. whether the respondent feels like they can express their views on different topics in management meetings). The questions were relatively simple and – because the response category was yes/no rather than an attitudinal scale – they were not flagged as at risk of measurement effect. However, the fact that the questions are indeed attitudinal (despite the yes/no response scale could still leave them prone to the risk of social desirability bias.

type of measurement effect. Where a difference by mode is observed in that variable, it is not always possible to determine which type or types of measurement effect is driving the difference. The next section does, however, attempt to explore the evidence for a specific measurement effect – socially desirable reporting and/or positivity bias – in more detail.

Table 5.2: Expected vs. observed measurement effects, by specific type of measurement effect

	Number of variables identified	Number where mode effect observed	Percentage where mode effect observed
A1 Socially desirable reporting	126	102	81%
A2 Positivity bias on rating scale	167	139	83%
B1 Complex stem and clarifications	23	19	83%
B2 Extra information	28	18	64%
B3 Computation required	7	4	57%
C1 Higher number of response options	114	90	79%
C2 Battery of questions/grid	138	114	83%

NB: A variable may have been assessed as at risk from more than one type of measurement effect.

5.2 Were differences observed between modes in the expected direction?

Identifying that a difference between modes was observed for a variable identified as “at risk” is one thing. A more stringent test of a possible measurement effect is whether or not the difference observed is of the type or in the direction that we would expect given the predicted measurement effect. For variables identified as at risk of social desirability effects, it is possible to predict which responses we would expect to be more or less prevalent in the self-completion web mode compared with F2F, that is which are the most socially desirable and/or positive responses. We can then compare the distribution of responses across modes and assess how consistently a difference between modes is observed in the expected direction. To the extent that differences between modes in the expected direction are consistently observed, even if not statistically significant, this provides further evidence of a potential measurement effect.

To examine the directionality of any remaining mode effects after controlling for known differences in sample composition, binary logistic regression was used. The dependent variable was coded such that a value of 1 indicated response categories predicted to be more prevalent among Panel respondents and where we would therefore expect the binary mode variable (coded F2F=1 Panel=0) to have an odds ratio of less than 1. The regression was run including controls for demographic and job characteristics (see Chapter 4) alongside the binary indicator of survey mode. Table 5.3 summarises the number of variables where the coefficient of the mode variable was in the expected direction, that is pointing towards Panel respondents giving the less positive/less socially desirable response.

For results from the binary logistics regressions see “Testing for direction of effect.xls” tab in the accompanying annex.

Table 5.3: Expected vs. observed measurement effects: results of binary regression to test for direction of mode difference where socially desirable response/positivity bias predicted

	Number of variables identified	% of variables
Mode difference in expected direction (Socially desirable/more positive response **less** common for Panel respondents): Statistically significant	105	63
Mode difference in expected direction (Socially desirable/more positive response **less** common for Panel respondents): Not statistically significant	30	18
Mode difference **not** in expected direction (Socially desirable/more positive response **more** common for Panel respondents): Statistically significant	17	10
Mode difference **not** in expected direction (Socially desirable/more positive response **more** common for Panel respondents): Not statistically significant	15	9

For the majority of variables (81%) differences between modes were observed in the expected direction, that is Panel respondents were less likely to provide a positive or socially desirable response. The difference between modes was statistically significant in 63% of cases. This suggests that social desirability effects may well be behind the differences we observe.

6. Quality of occupation and pay data collected

This chapter looks more closely at some specific variables which are central to the SES survey - variables recording pay and the occupation and industry in which the respondent works - and how these varied between the Face to Face (F2F) and Panel surveys.

6.1 Occupation and industry by mode

Information on occupation and industry is traditionally collected via open-ended, free text questions which ask “what does the firm you work for mainly make or do “ and “what kind of work do you do most of the time“ The resulting text strings are coded by researchers/expert coders into established code frames – the Standard Occupational Classification (SOC) and the Standard Industrial Classification of economic activities (SIC). Such open questions, requiring the entry of detailed and specific free text, pose a potential challenge for self-completion surveys (such as web) with concerns that respondents will provide insufficient information to allow SOC and SIC to be coded. As an alternative to free text entry, self-completion surveys sometimes ask respondents to self-select their occupation and/or industry from a list of pre-coded options. However, such lists are necessarily less detailed than full 4 digit SOC/SIC coding and respondent’s self-assignment may not necessarily be accurate.

This section looks in more detail at the information collected on industry and occupation in the SES Face to Face (F2F) and NatCen Opinion Panel surveys. It compares the quality of the free text responses and resulting SOC/SIC codes across the two. The Panel asked for information on occupation/industry using both the traditional open questions and a a closed list of pre-defined response options , allowing responses obtained using the two methods to be compared.

6.1.1 Quality of open code responses on occupation and industry

Several measures were used to assess the quality of the free text responses people gave when asked about their occupation and industry.

- The proportion of cases for which the information provided was sufficient to assign a 4 digit SOC or SIC code.

- The confidence score associated with the SOC/SIC code assigned. SOC and SIC were dual coded in SES, by a trained member of NatCen's Data Operations Team and by CASCOT.[15]-
- The length of the text string provided. All other things being equal it is anticipated that longer text strings should result in more accurate codes.

As show in Table 6.1, the proportion of cases which could not be assigned a 4-digit SOC 2020 or SIC 2007 code was very low for both the F2F and Panel surveys. There was no difference between modes in the proportion of cases coded for occupation, which was negligible at 0.6%. However, the proportion of cases which could not be assigned a SIC code was twice as large in the Panel as for the F2F survey (3.1% compared with 1.5%).

Panel respondents tended, on average, to give longer answers than F2F respondents when asked to describe what their firm made or do or what they themselves did most of the time. These longer answers did not necessarily result in better quality coding. There was in fact a weak negative correlation between coder confidence score and number of characters provided for SIC (p=-0.17) and no correlation for SOC (p=-0.02). Some of the information recorded by Panel respondents may have been unnecessary - especially for SIC where people tended to describe the type of work rather than the industry - with interviews trained or having the experience to record the key information that will be helpful to coders.

The biggest differences in confidence scores assigned to the final codes were for CASCOT-coded SIC, where F2F scored higher than the Panel. Overall, however, there was relatively little difference between the two surveys in the average (mean and median) confidence score assigned to the codes selected by either CASCOT or the human coders. This suggests that where codes were assigned, we can be equally confident of those codes in both modes.

Table 6.1: Comparing the quality of the free text industry /occupation data used to code SOC and SIC

	SIC		SOC	
	F2F	**Panel**	**F2F**	**Panel**
Percentage of cases not coded	1.5%	3.1%	0.6%	0.6%
Number of characters in string*				
Mean	26.4	47.6	61.4	106.3
Median	19.0	33.0	45.0	86.0
Number of words in string				
Mean	3.6	7.0	8.8	16.0
Median	2.0	5.0	6.0	13.0
Coder confidence score				
Mean	44.4	44.0	63.0	62.7

[15] CACOT is software designed to make the coding of text information to standard classifications simpler, quicker and more reliable by assigning a code to a piece of text (i.e. an industry for SIC or a job title for SOC). CASCOT has been designed to perform a complicated analysis of the words in the text, comparing them to the words in the classification, in order to provide a list of recommendations for how to code to the UK standards developed by the Office for National Statistics (ONS).

Median	39.0	40.0	63.0	64.0
CASCOT confidence score				
Mean	47.8	44.0	61.8	63.1
Median	45.0	41.0	60.5	63.0

*"What does the firm you work for mainly make or do?" (SIC) / "What kind of work do you do most of the time?" (SOC)

6.1.2 Comparing responses from open versus closed questions

On some web surveys, information on industry and occupation is collected using closed questions - with the response options being equivalent to 2 digit SOC and 1 digit SIC categories – either as an alternative to, or in addition, to open coded questions. SES online respondents were asked both the open and closed version of the occupation and industry questions providing an opportunity to compare the two approaches.

The data indicate that allowing Panel respondents to self-code their industry and occupation did not provide equivalent data to office coding of open text responses. Table 6.2 shows that in 33% of cases there was a mismatch between the final 1 digit SIC code assigned using the responses given at the open question "What does the firm you work for mainly make or do" and the industry respondents chose themselves via a drop down menu. The biggest mismatch occurred for people classed as working in "Administrative support services" who tended to class themselves as working in "Other Services".

Table 6.2 Comparing researcher and respondent coded industry

Office coded SIC 2007	Self Coded match	Self Coded does not match	Unweighted N
Section A Agriculture, Forestry and Fishing	*	*	*
Section B Mining and Quarrying	*	*	*
Section C Manufacturing	65%	35%	183
Section D Electricity, Gas, Steam and Air Conditioning Supply	*	*	*
Section E Water Supply; Sewerage, Waste Management and Remediation Activities	*	*	*
Section F Construction	70%	30%	74
Section G Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles	67%	34%	194
Section H Transportation and Storage	61%	39%	107
Section I Accommodation and Food Service Activities	67%	33%	69
Section J Information and Communication	50%	50%	119
Section K Financial and Insurance Activities	91%	9%	109
Section L Real Estate Activities	*	*	*
Section M Professional, Scientific and Technical Activities	47%	53%	144
Section N Administrative and Support Service Activities	10%	90%	67
Section O Public Administration and Defence; Compulsory Social Security	64%	36%	176
Section P Education	91%	9%	277

Section Q Human Health and Social Work Activities	75%	25%	312
Section R Arts, Entertainment and Recreation	55%	45%	47
Section S Other Service Activities	46%	54%	50
All	**67%**	**33%**	**2,000**

*% based on N <30

The mismatch was even more pronounced for occupation. In 55% of cases there was a mismatch between the final 2 digit SOC code assigned using the responses given at the open question "What kind of work do you do most of the time" and the occupation respondents chose themselves via a drop down menu (Table 6.3). Looking at the largest occupations (with 100+ cases), there were particularly pronounced mismatches in business related occupations.

Table 6.3 Comparing researcher and respondent coded occupation

Office coded SOC 2020	Self Coded match	Self Coded does not match	Unweighted N
CORPORATE MANAGERS AND DIRECTORS	37%	63%	140
OTHER MANAGERS AND PROPRIETORS	48%	52%	63
SCIENCE, RESEARCH, ENGINEERING AND TECHNOLOGY PROFESSIONALS	64%	36%	154
HEALTH PROFESSIONALS	80%	20%	89
TEACHING AND OTHER EDUCATIONAL PROFESSIONALS	89%	11%	142
BUSINESS, MEDIA AND PUBLIC SERVICE PROFESSIONALS	31%	69%	197
SCIENCE, ENGINEERING AND TECHNOLOGY ASSOCIATE PROFESSIONALS	8%	92%	51
HEALTH AND SOCIAL CARE ASSOCIATE PROFESSIONALS	23%	77%	57
PROTECTIVE SERVICE OCCUPATIONS	*	*	*
CULTURE, MEDIA AND SPORTS OCCUPATIONS	*	*	*
BUSINESS AND PUBLIC SERVICE ASSOCIATE PROFESSIONALS	12%	88%	172
ADMINISTRATIVE OCCUPATIONS	60%	40%	198
SECRETARIAL AND RELATED OCCUPATIONS	34%	66%	38
SKILLED AGRICULTURAL AND RELATED TRADES	*	*	*
SKILLED METAL, ELECTRICAL AND ELECTRONIC TRADES	43%	57%	58
SKILLED CONSTRUCTION AND BUILDING TRADES	70%	30%	37
TEXTILES, PRINTING AND OTHER SKILLED TRADES	23%	77%	31
CARING PERSONAL SERVICE OCCUPATIONS	42%	58%	134
LEISURE, TRAVEL AND RELATED PERSONAL SERVICE OCCUPATIONS	39%	61%	41
COMMUNITY AND CIVIL ENFORCEMENT OCCUPATIONS	*	*	*
SALES OCCUPATIONS	51%	50%	97
CUSTOMER SERVICE OCCUPATIONS	72%	28%	39
PROCESS, PLANT AND MACHINE OPERATIVES	39%	61%	51
TRANSPORT AND MOBILE MACHINE DRIVERS AND OPERATIVES	61%	39%	49

ELEMENTARY TRADES AND RELATED OCCUPATIONS	*	*	*
ELEMENTARY ADMINISTRATION AND SERVICE OCCUPATIONS	12%	88%	116
All	**45%**	**55%**	**2,043**

*% based on N <30

These findings regarding a high level of mismatch with self-coded SOC and SIC suggest that, in particular for a survey such as SES where there is particular interest in detailed information on occupation and industry, it may be advantageous to collect job details via open code questions even if the data collection moves away from F2F.

6.2 Pay by mode

Data collected on pay may differ by mode of data collection for a number of reasons. On the one hand, people may feel more comfortable giving a response online in the absence of an interviewer. On the other, online respondents may be more tempted to satisfice and give an estimated rather than a precise figure. Finally, to avoid frustrating respondents, web surveys often include fewer consistency checks than interviewer administered surveys and so may be more prone to data entry errors. The latter issue is less of a concern for SES as the same checks on very high or low values were included in both the F2F and web surveys. Nevertheless, there is still the potential for differences by mode.

In this section, we first look at whether there is any evidence that the mode affected how people provided information about pay – for example, differences in the level of missing data or the time period for which people gave their pay – and then consider whether, before and after controlling for observed differences in sample composition (see Chapter 4) there were significant differences in reported pay across the F2F and Panel surveys.

Table 6.4 below looks at whether there were differences by mode in the proportion of people who provided sufficient information for an estimate of weekly pay to be produced.[16] As expected, the level of refusals was lower on the Panel compared with F2F. The level of "Don't know" responses was also lower for the Panel. In the F2F survey, "Don't knows" may be acting as polite refusals. Panel conditioning, and the fact that the NatCen panellists are used to being asked to provide information on their income from previous surveys may also be playing a role here.

Table 6.4 Level of missing data on pay questions by survey mode

	F2F: Don't know	F2F:Refusal	Panel: Don't know	Panel: Refusal
Employees: Gross weekly pay	3.6%	5.0%	1.1%	1.7%
Self-employed: Net weekly pay	7.9%	8.8%	5.3%	7.6%

[16] The level of missing data on hourly pay was very similar.

The figures below show that there were some differences between modes in the period – annual, monthly, weekly etc- which respondents used to report their pay (Table 6.5). Employees completing the survey via the Panel were less likely to give their pay as an annual amount compared with employees interviewed F2F for example. These differences may be due to differences in sample composition (people in certain occupations and those on lower incomes/more casual employment are less likely to think in terms of annual pay). However, given the lack of strong evidence for differences on job or demographic characteristics between the two samples (see Chapter 4) there is no clear reason to expect this to be the case. There may, therefore, be a measurement effect at work, for example because – for the online Panel respondents – the question asking them to select a period appeared on the same page as the question asking for an amount whereas F2F respondents were first asked to give an amount and then say what time period that related to. The former may mean that respondents feel freer to give their pay using a reference period that suits them, rather than feeling prompted to give an annual amount. Ultimately, however, it does not matter if respondents in different modes give their incomes for different periods. What is important is whether, once everyone's pay has been calculated for the same period, the pay distribution varies by mode.

Tabe 6.5 Period for which pay provided, by mode

	F2F: Employee	**Panel: Employee**	**F2F: Self-employed**	**Panel: Self-employed**
Hourly	2%	5%	9%	21%
Weekly	4%	6%	3%	10%
4 weeks	3%	9%	19%	36%
Calendar month	18%	28%	69%	29%
Yearly	73%	52%	1%	4%
Other	1%	1%	9%	21%
Unweighted N	2268	1693	290	168

Table 6.6 below shows the distribution of weekly pay data across the two survey modes for employees and the self-employed while Table 6.7 does the same for hourly pay data. There are differences by mode in mean pay with reported pay for employees lower among Panel respondents compared with F2F respondents. However, these differences are not statistically significant at the 5% level either before or after controlling for differences in demographic characteristics through regression. The mean estimates for the relatively small number of self-employed respondents are sensitive to extreme values but a comparison of median values again points to reported pay being lower for Panel respondents compared with F2F respondents.

Table 6.6 Weekly pay by mode

	F2F: Employee	Panel: Employee	F2F: Self-employed	Panel: Self-employed
Weekly pay: Mean	£818.68	£735.89	£800.15	£1182.44
Weekly pay: SD	£1184.54	£944.76	£1426.69	£6006.93
Weekly pay: Median	£615.38	£577.37	£541.80	£400.00
Unweighted N	2,254	1,662	288	159

Table 6.7 Hourly pay by mode

	F2F: Employee	Panel: Employee	F2F: Self-employed	Panel: Self-employed
Hourly pay: Mean	£22.08	£19.99	£23.78	£32.71
Hourly pay: SD	£29.55	£25.79	£40.81	£111.21
Hourly pay: Median	£16.83	£15.40	£15.38	£14.43
Unweighted N	2,258	1,662	282	159

There is mixed evidence on how survey differences affect pay estimates in SES. Given the complexities around calculating pay, for example how to treat outliers, data users are encouraged to carry out their own comparisons of relevant pay estimates before deciding whether or not to conduct analysis using the combined F2F and Panel data.

7. Considerations for data analysts

Analysts wishing to explore findings from the Skills and Employment Survey 2024 have access to data from both the Face to Face survey and the online survey fielded via the NatCen Opinion Panel. Data users may wish to combine data from the two surveys in order to maximise the number of cases available for analysis, especially if interested in small subgroups. However, users also need to bear in mind that, as described in this report, there are differences between the two surveys in the estimates obtained. It is likely that these differences are due to measurement effects across the different modes of data collection used. Not all variables are affected and the magnitude of any differences varies across variables.

SES data users are encouraged to decide on a case by case basis, depending on the variables used and the purpose of the analysis, whether or not the data from the two surveys are sufficiently similar to enable the Face to Face and Panel data to be combined. The annex tables showing descriptive findings for each SES variable which accompany this report provide a useful starting point for this.


National Centre
for Social Research