SURVEY FUTURES

SURVEY DATA COLLECTION METHODS COLLABORATION

Working Paper 11:

Measurement Mode Effects in Live Video Interviewing: Evidence from a Mixed-Mode Quasi-Experimental Survey

Sebastian Kocar[1], Matt Brown[2], Carole Sanchez[2], George B. Ploubidis[2], Gabriele Durrant[3]

[1]University of Queensland, [2]University College London, [3]University of Southampton

December 2025

www.surveyfutures.net

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

The BCS70 Age 51 Survey was designed and managed by the UCL Centre for Longitudinal Studies and supported by the Economic and Social Research Council (ESRC) (Grant number ES/W013142/1). Fieldwork was conducted by NatCen and Kantar (now Verian).

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:

Kocar S, Brown M, Sanchez C, Ploubidis GB & Durrant G (2025) 'Measurement Mode Effects in Live Video Interviewing: Evidence from a Mixed-Mode Quasi-Experimental Survey', *Survey Futures Working Paper* no. 11. Colchester, UK: University of Essex. Available at https://surveyfutures.net/working-papers/.

Contents

Abstract

Live video interviewing (VI) emerged as a mode for conducting large-scale social surveys during the COVID-19 pandemic but there is limited existing evidence on measurement differences between VI and in-person interviewing. This paper addresses this evidence gap by investigating measurement mode effects in a large-scale birth cohort study with large samples of respondents participating via VI and in-person, and a wide range of survey items included in the analysis of distributional differences. Due to non-random allocation to survey mode, propensity score weighting (PSW) was used as a method for causal inference, with carefully selected confounders associated with the mode of participation (exposure) and key survey items (outcomes).

After PSW adjustment, the magnitude of measurement differences between VI and in-person interviewing is minimal and appear largely attributable to residual sample composition effects arising from unobserved confounders. We detected slightly more candid responses from VI participants (regarding political views and psychological stressors) and marginally higher cognition scores, likely due to the absence of a physically present interviewer. Overall, the evidence suggests that VI can be considered a suitable alternative to in-person interviewing from a measurement perspective, offering advantages such as cost-efficiency and the potential to increase overall response rates in mixed-mode surveys.

Keywords: live video interviewing, measurement mode effect, birth cohort study, propensity score weighting, mixed-mode survey, quasi-experimental study

1. Introduction

Live video interviewing (in the following abbreviated as VI), also described as "video-mediated face-to-face interviewing" (Schober, 2018) is a relatively new survey mode. Use of this mode accelerated greatly during the COVID-19 pandemic (see Conrad et al., 2023; Durrant et al., 2024; Endres et al., 2023; Pennay et al., 2024; Schober et al., 2023). However, investigations into the feasibility of VI to conduct social surveys had begun sooner, with methodological research exploring recruitment of sample members, interviewer setup, such as screen configuration and visual background (e.g. Schober et al., 2020), and technical issues, such as choice of video-mediated platforms and access to the internet and platforms by sample members (e.g. Schober, 2018). Other aspects explored were the potential impact of VI on representativeness (e.g. Anderson, 2008) and nonresponse bias (e.g. Jeannis et al., 2013). The literature generally suggests that the use of VI could, particularly in a mixed-mode setting, present advantages for survey practitioners, as offering an additional response mode could positively impact response rates and representativeness (Jeannis et al., 2013) and increase cost-efficiency (see Conrad et al., 2023). Durrant et al. (2024) reported that all the large-scale surveys in the UK they studied, which incorporated VI into their survey design, used VI either: (i) as a supplementary mode (i.e. an additional option for respondents) or (ii) as a temporary substitute for in-person interviewing during the pandemic. However, the introduction of VI has the potential to introduce non-ignorable measurement differences compared to other modes, which is an important aspect of this emerging survey mode that has not yet been extensively explored. Our study addresses this gap with its primary objective being to investigate the effects of measurement mode in a mixed-mode survey that includes both in-person and video interviews. We aim to assess whether responses collected via VI align closely with those obtained in person, thereby informing the extent to which VI can be deployed without compromising measurement quality. To date, measurement mode effects in mixed-mode surveys that include VI have been investigated in a small but growing body of studies (i.e. Asensio et al., under review; Conrad et al., 2023; Endres et al., 2023; West et al., 2022). These studies compared measurement differences between VI and pre-recorded video interviewing (West et al., 2022), VI and the online mode (Asensio et al., under review; Conrad et al., 2023; Endres et al. 2023), and VI and in-person interviewing (Asensio et al., under review; Conrad et al., 2023; Endres et al. 2023). Findings relating to the impact of VI on measurement have been largely encouraging. For example, Conrad et al. (2023) reported fewer instances of non-differentiation by VI respondents compared to web respondents when answering grid questions, and VI showed promising similarities to in-person face-to-face interviewing for several data quality metrics (Endres et al., 2023). Asensio et al. (under review) also found limited measurement differences between VI and in-person interviewing, though they observed more significant differences between VI and web responses, particularly for sensitive items. Much of the existing research has been conducted in the U.S. (e.g. Conrad et al., 2023; West et al., 2022), has relied exclusively on cross-sectional data, and has been based on data collected for small-scale, exploratory research and observational data (Centeno et al., 2024). Additional evidence is therefore needed.

This paper addresses these gaps and is based on large-scale research, in a non-US context, using data from a longitudinal study (birth cohort) employing a quasi-experimental design. We investigate VI measurement mode effects in a UK survey, using data from one of the largest-scale surveys to use VI in the world, the 1970 British Cohort Study (BCS70). The study incorporates large samples of respondents participating via VI and in-person from the latest wave of the study conducted between 2021 and 2024, providing an opportunity for a comprehensive analysis of measurement differences between the modes. More specifically, the study aims to identify relevant differences in the distribution of a wide range of more than 100 survey items across VI and in-person interviewing.[1] The evidence will also shed light on the potential effect of measurement differences between in-person interviewing and VI on the measurement of individual-level change over time, resulting from switching between modes in a birth cohort study. Often, the identification of measurement effects is made more complicated through their confounding with selection effects. Importantly, our study allows the differentiation between measurement and selection effects. Our analytical approach uses a method for causal inference to make the important distinction between the effects of non-random mode allocation and selection (i.e. sample composition effects) and measurement effects due to the survey mode, under the assumption that no unobserved confounders influence these relationships.

1.1. Research questions

While the main aim of our investigation is to identify measurement differences attributable to mode effects, the study design also calls for an assessment of the impact of initial survey mode allocation (VI or in-person) on nonresponse patterns, as well as an investigation into the characteristics of sample members who chose a particular mode. The following research questions are addressed:

- *RQ1) How do nonresponse patterns differ between respondents initially assigned to VI and in-person interviewing, given that participants could choose to respond via an alternative mode?*

[1] Different mechanisms through which measurement mode effects influence responses are presented in the 'Measurement mode effects in mixed-mode surveys' section.

We will compare refusal and noncontact rates among respondents, as well as the proportion of respondents who chose not to participate via their allocated mode but instead used an alternative mode (e.g. allocated to VI but responded in person).

- *RQ2) What are key characteristics of those who complete the survey via VI and those who complete in-person, accounting for the quasi-experimental nature of the design?*[2]

  We will assess how the compositions of the final VI and in-person samples differ due to imperfect random assignment to mode,[3] differential nonresponse (RQ1) and respondents' own self-selection to mode.[4]

- *RQ3) What measurement differences attributable to mode effects are observed between VI and in-person survey modes?*

  After adjusting for non-random allocation to mode we address our central research question by estimating measurement mode effects between VI and in-person modes.

Our evidence extends the existing published evidence from Conrad et al. (2023), Endres et al. (2023), and Asensio et al. (under review).

2. Literature review

2.1. Measurement mode effects in mixed-mode surveys

Differences between modes in survey data are typically categorized into three types: coverage effects, selection effects (including nonresponse effects and selection-to-mode effects), and measurement mode effects, which are systematic differences in measurement attributable to the mode of data collection. Such differences in measurement can be considered a form of measurement error, one of two main types of errors in the survey process in the Total Survey Error Framework (Groves et al., 2011). As such, these errors differ from coverage errors, which are selection effects associated with survey mode (e.g. Beulens et al., 2012; Schouten et al., 2013), and nonresponse errors, both of which are errors of representation (Groves et al., 2011). The largest differences in measurement are typically found between interviewer administered modes (such as in-person face-to-face or telephone) and self-administered modes (such as web or paper).

Typically, interviewer administered modes are more prone to social desirability, positivity bias and acquiescence response bias (Dillman et al. 2014; Groves et al., 2011). On the other hand, self-completion modes tend to result in more unit and item nonresponse. Another source of measurement differences is how questions are presented – interviewer administered modes typically present questions orally, whereas self-completion modes present questions visually. This can result in

[2] A quasi-experimental design is a type of research design used to estimate the effect of a treatment or intervention. In contrast to randomized experiments, such as randomized controlled trials, the treatment conditions are not assigned at random. There are both practical and ethical constraints that make randomisation either impossible or inadvisable in practice (Reichardt, 2009). PSW or matching methods as methods for causal inference are approaches typically used in quasi-experimental designs.

[3] In principle, the respondents were randomly allocated to mode, but a small proportion of respondents were targeted for earlier interview due to their previous response patterns, making them more likely to be assigned to the VI mode. More information is provided in the Methods section.

[4] Respondents had the opportunity to respond via an alternative survey mode to which they were not allocated to. More information is provided in the Methods section.

presentation and order effects, including recency bias in interviewer administered modes and primacy bias in self-completion modes (Dillman et al. 2014; Krosnick et al. 1996).

Additionally, complex questions involving detailed instructions and/or definitions can be more challenging to administer in self-completion modes. These outlined measurement differences between modes are a potential drawback in mixed-mode surveys (Dillman et al., 2009), as analyses using variables affected by mode of completion may lead to biased results. Evaluating mode differences that affect measurement is even more important in the case of mode changes between waves of a longitudinal study as measurement mode effects may lead to biased estimates of change (Cernat, 2015; Dillman, 2009; Dillman & Christian, 2005; Kocar & Biddle, 2020).

In studies of measurement mode effects, differences in data quality indicators across the modes, such as completeness, accuracy, reliability, and response distributions across survey items, are typically examined (Beulens et al., 2012; De Leeuw & Zouwen, 1988; Schouten et al., 2013). While existing research offers extensive evidence of mode effects for a range of survey modes, including both interviewer-administered and self-completion modes (e.g. Dillman et al., 2009), there is much less evidence regarding the impact of VI, which has only recently emerged as a survey mode (Centeno et al., 2024; Durrant et al., 2024).

2.2. Measurement mode effects in surveys using video interviewing

The review of literature on VI identified a few recent studies that have explored the effect of VI mode on measurement in mixed-mode designs – in comparison to web, pre-recorded video interviewing and in-person interviewing. For instance, in a study comparing web surveys and VI, Conrad et al. (2023) identify several measurement differences. Respondents in VI exhibited higher data quality in terms of non-differentiation, i.e. selecting a single response option for all statements in a battery, compared to web survey participants. However, VI respondents also rounded numerical answers more, disclosed less sensitive information, and left more sensitive questions unanswered compared to the web respondents. The authors conclude that interviewer-administered modes, whether in-person or via video, encourage more conscientious responses by reducing non-differentiation, likely due to increased attention and engagement from the respondent, in parts due to the presence of the interviewer (Conrad et al., 2023). Moreover, in a study comparing VI and pre-recorded video interviewing, measurement differences resulting from these two approaches were relatively rare (West et al., 2022); though for a small number of survey items, including the frequency of volunteering, there was more interviewer variance in VI than in pre-recorded ones. This finding can be explained by the interactive nature of VI (West et al., 2022).

Evidence also suggests that VI can differ from in-person interviewing. On the one hand, VI can attract participants who may be unwilling to take part in in-person interviews and offer benefits such as enabling sign language for deaf respondents and allowing discreet volume adjustments (Schober et al., 2020). On the other hand, VI can present additional challenges compared to the in-person mode, including the difficulty interviewers face in avoiding nonverbal cues during sensitive questions (e.g. trying to keep a "poker face"), and the potential effects of the interviewer's gaze and image size on the respondent's screen (Schober et al., 2020; West et al., 2022). Additionally, Conrad et al. (2023) argue that VI may also introduce additional time pressure in comparison to in-person interviews, as evidenced by more rounding of numerical answers in the VI interviews. Despite these challenges, Endres et al. (2023) report that video interviews show promising similarities to in-person interviews in terms of data quality metrics such as non-differentiation, item nonresponse, and depth of responses to free-text questions. More recently, Asensio et al. (under review) provide the largest-scale experimental evidence to date on VI mode effects using a randomised study comparing video, in-person and web interviewing. Their findings reinforce the conclusion that VI most closely resembles in-person interviewing on key indicators of measurement quality. Across 25 survey items, the authors identify only very small differences between VI and in-person modes in both item nonresponse and response distributions, even for sensitive measures administered via a self-completion module. VI respondents also showed slightly lower item nonresponse than in-person respondents overall, though levels of nonresponse were low across all modes. In contrast, substantially more differences were observed between VI and web respondents, particularly on sensitive items such as mental wellbeing, health, and financial difficulty (Asensio et al., under review). In summary, the evidence so far suggests that video interviews can provide data quality comparable to in-person methods, but that measurement differences between VI and online surveys may be more pronounced. This can be explained by the fact that VI as an interviewer-administered survey mode tends to resemble in-person interviews more closely than self-completion modes.

2.3. Methods for causal inference in the absence of random assignment to modes

Fully experimental research designs, such as randomised controlled trial-type experiments (Rubin, 1974) with little to no unit nonresponse, are theoretically the most suitable for investigating measurement mode effects (see West et al., 2022). However, such designs are not always feasible or cost-effective (D'Agostino Jr, 1998). One of the alternatives is a quasi-experimental design, where treatment conditions are not assigned (completely) at random (Reichardt, 2009). In cases where respondents are not randomly allocated to survey modes, two potentially competing sources of response differences can emerge: measurement mode effects and mode (self-)selection effects; non-random differences in sample composition across survey modes, which can largely explain variations in survey estimates, make isolating the potential impact of measurement mode effects more challenging due to the potential confounding of these effects and self-selection effects across modes (Suzer-Gurtekin et al., 2018). This is similar to the way treatment effects in quasi-randomised experiments such as medical trials can be confounded (Reichardt, 2009). Therefore, to accurately identify net measurement mode effects in quasi-randomised survey experiments, it is crucial to implement a careful post-survey design that accounts for factors like survey mode coverage (i.e. whether respondents are able to respond via a particular mode) and respondent mode preferences (Beulens et al., 2012; Jäckle et al., 2010; Schouten et al., 2013).

Methods for causal inference that account for non-randomised assignments have the potential to provide an effective solution to separate out the confounding effects (see King et al., 2011, for comparative analysis). These methods are referred to as quasi-experimental approaches, and they are based on relatively strong assumptions, such as correct model specification and no unmeasured confounders[5] in the case of propensity score weighting (PSW) (Freedman & Berk, 2008). In the study of measurement mode effects, these methods attempt to balance samples collected in different modes (i.e. treatment and control groups) across a vector of measured characteristics that are relevant in mitigating selection biases and isolating the effects of different survey modes on data quality (see

[5] Confounders can be defined as variables which are associated with both the outcome variables of interest (e.g. well-being items in our study) and a choice of treatment (i.e. mode of participation). In causal analysis, they are a source of selection bias (Mercer et al. 2017), and are often used with methods for causal inference.

Kibuchi et al., 2024; Kocar et al., 2021; Suzer-Gurtekin et al., 2018). Vannieuwenhuyze and Loosvelt (2013) explain that such methods can be used as part of mixed-method calibration, one of three approaches to disentangle selection effects from measurement effects in mode effect analysis. This approach is based on the ignorable treatment assignment assumption/conditional exchangeability, which states that, given a set of observed covariates, treatment assignment is independent of potential outcomes (Rosenbaum & Rubin, 1983). The more plausible this assumption is, the less likely it is that the remaining mode differences are explained by self-selection effects. Mixed-method calibration is particularly useful in the absence of comparable single-mode data (preferably with representative benchmarks), which can be incorporated into the extended mixed-method comparison approach, with or without calibration, for studying mode effects (Vannieuwenhuyze & Loosvelt, 2013).[6]

The literature describes and empirically evaluates a range of methods for causal inference that have been used in disciplines such as medicine, epidemiology, and behavioural economics, and less commonly in survey methodology (e.g. Kibuchi et al., 2024; Kocar et al., 2021; Li, 2024). Methods for causal inference include matching techniques, which can be further classified into modelling methods and stratification methods (Sizemore & Alkurdi, 2019), as well as propensity score-based methods such as propensity score weighting (Austin, 2011; Rosenbaum & Rubin, 1983). Some of the most commonly used matching methods include propensity score matching, exact matching, coarsened exact matching, Mahalanobis distance matching, and Euclidean distance matching (Austin, 2011; King et al., 2011; Rosenbaum & Rubin, 1983). While some authors suggest that the choice of causal inference method may not result in substantial differences in the balance of the adjusted samples (rather than the selection of confounders), other authors are less confident in the robustness of methods like propensity score matching (see King & Nielsen, 2019; Kibuchi et al., 2024). This is due to a paradox which arises from the way propensity scores interact with matching methods, and while propensity scores have many other productive applications, such as regression adjustment, inverse weighting, and stratification, they are not always suitable for use in matching (King & Nielsen, 2019).

Often, selecting appropriate confounders[7] for adjustment is more critical than choosing a statistical method for mitigating bias from survey mode allocation. Researchers using causal inference methods should prioritize variables based on their theoretical impact on the outcome, as both imbalance in confounder distribution and the relative importance of these variables can introduce bias (King & Nielsen, 2019). The optimal variable selection in, for example, a propensity score model should include variables that are related to the substantive outcome (e.g. cognitive ability scores or health outcomes) and, ideally, to the exposure (i.e. the mode of participation in our case). Including variables unrelated to the outcome but related to the exposure may increase the variance of the estimated exposure effect without reducing bias (Brookhart et al., 2006; Pearl, 2000).

[6] An alternative approach to studying measurement mode effects in randomised experiments where some respondents do not comply with their allocated treatment is presented in the study from Goodman et al. (2022).

[7] These are measured characteristics, such as demographic variables, that are crucial for mitigating selection biases in the absence of random assignment (to group or to mode). In practice, they are commonly used as covariates in propensity score models.

3. Methods

3.1. Data

BCS70 Birth cohort. In this study, we use data from BCS70, specifically from the Age 51 survey, conducted between 2021 and 2024, when the cohort was aged between 51 and 53 years.[8] BCS70 is a national longitudinal birth cohort study that began in 1970, collecting data on births and families from the parents of an initial sample of approximately 17,196 babies born in the UK during the same week. The initial survey was followed by a total of ten follow-up surveys which tracked respondents who continued residing in the UK. The Age 51 Survey was conducted by the National Centre for Social Research (NatCen) and Verian (formerly Kantar Public) on behalf of the Centre for Longitudinal Studies (CLS). The survey included a range of topics, such as household composition, family and relationship details, housing, employment, income, cognition, lifelong learning, physical and mental health, COVID-19, political attitudes, life satisfaction, and diet.

BCS70 Age 51 Fieldwork. The BCS70 Age 51 Survey was planned to commence in June 2020 and was intended to be conducted in-person as had been the case with the majority of previous follow-up surveys. However, the launch of fieldwork was delayed by the COVID-19 pandemic and the suspension of in-home interviewing. In 2020, it was unclear for how long in-home interviewing would be unfeasible and so the survey was adapted for administration via VI. This approach was successfully trialled in a small-scale feasibility test conducted in late 2020 before being launched at scale in June 2021.

Fieldwork for the Age 51 Survey was split into seven "waves' or 'batches'. The first four waves of fieldwork, between June 2021 and June 2022, were conducted only by VI, due to the pandemic, whilst the final three waves, between May 2022 and November 2023, were conducted primarily in-person as pandemic restrictions on in-home interviewing had been lifted. Participants approached during the video-only waves, who opted not to respond, were re-contacted once in-person interviewing had resumed and invited to take part with an interviewer. Participants contacted during the predominately in-person waves (May 2022 to November 2023) were able to take part via VI if they expressed a preference to do so (for participation rates, see below). A small number of participants, contacted during the in-person waves, were only invited to take part via VI if no in-person interviewer was available in their area.

Of the 12,165 eligible study members, 7,339 (60.3%) participated in the mainstage fieldwork. Although not officially permitted, these include 29 interviews that were conducted by telephone. Additionally, nonrespondents were invited to take part in a 'mop-up' self-completion survey, where they were asked to complete a short web version of the questionnaire after the mainstage fieldwork. Mop-up fieldwork took place between December 2023 and January 2024, and an additional 5.6% of eligible study members completed the questionnaire. Importantly, for the purpose of our measurement mode effects analyses, interviews completed by web[9] or telephone are excluded.

Mode of data collection. The allocation of individuals to Wave 1 to 7 was conducted at the outset of the project in 2020, prior to the introduction of the VI mode. Allocation to waves was based on the intention to create batches of cases across Britain which were sufficiently close to each other geographically to form workable 'assignments' for in-person interviewers. In addition, those who were more challenging to contact and persuade to take part in previous BCS70 surveys were

[8] For the BCS70 cohort profile, see Sullivan et al. (2022).

[9] The mop-up respondents completed a notably shorter version of the questionnaire compared to VI and in-person respondents.

disproportionately allocated to the earlier waves of the BCS70 Age 51 survey to extend the time during which these individuals could be contacted and interviewed. In addition, participants initially allocated to waves which became VI-first but who had not previously provided their telephone or email addresses had to be re-allocated to one of the later in-person-first waves to facilitate contact with these individuals as all contact in VI-first waves was made remotely. As such, although allocation to wave, and subsequently therefore allocation to mode was not fully random, the cases allocated to the first four waves (VI-first approach) and the cases allocated to the later waves (in-person-first) were broadly equivalent in terms of key socio-demographic characteristics.[10] Furthermore, since the initial selection to survey mode did not require participants to respond via the mode they were allocated to, self-selection to survey mode further impacted the final distribution of the sample by mode of completion. Consequently, the design of our measurement mode effect study is quasi-experimental.

Both in-person and video interviews included a self-completion section which included more sensitive questions on topics including mental health, loneliness and politics including voting (see variable examples in Table A1). During in-person interviews, respondents completed this section privately on the interviewers' laptops. Interviewers could also read the questions aloud upon request. For VI, participants accessed a web-based version, using a unique link shared via Microsoft Teams. Interviewers remained on the call to assist with any technical issues and if necessary, could share their screen so that the participant could read the questions and tell the interviewer their answer. In situations where the participant was unable or unwilling to complete the self-completion section during the interview a link to the web survey was emailed so that the questions could be answered later. Both self-completion approaches were carefully designed to mirror each other to minimize measurement mode effects, such as avoiding grid layouts for repeated question items. Despite these efforts to maintain consistency, the differing contexts of data collection required analysis of potential measurement mode effects for those survey items. Given the contextual differences in how the surveys were administered, measurement differences between in-person interviewing and VI might be more pronounced for interviewer-administered questions than for self-completed questions, where the mode of administration was fundamentally the same across both contexts (computer-assisted self-interviewing and web).

3.2. Data analysis methods

Method for causal inference: propensity score weighting (PSW). To investigate measurement mode effects using a robust approach, we first implement PSW as a method for causal inference, given the quasi-experimental design of our study. We use this approach to address measurement mode effects by attempting to mitigate (self-)selection effects. We decided to use this method due to its widespread application, as well as recommendations from previous research (e.g. Austin & Shuster, 2016; Austin & Stuart, 2017; Desai & Franklin, 2019; Kibuchi et al., 2024; King & Nielsen, 2019). PSW is one of the methods that uses the propensity score (the probability of receiving treatment given observed covariates) to address biases in non-randomised studies. In this study, we use the inverse probability of treatment weighting (IPTW), which derives weights for adjustment from the inverses of the model-estimated probabilities of treatment (Czajka et al., 1992; Chesnaye et al., 2022).[11] One of

[10] These were reported in the previous Age 46 surveys. In the results section, we examine the socio-demographic differences among respondents in the Age 51 survey.

[11] There are other propensity score weighting methods, including the method of matching weights (Li & Greene, 2013), fine stratification approach (Desai et al., 2017), and overlap weights (Li et al., 2019).

the main advantages of PSW, including IPTW as one of its approaches, is that it is more efficient than matching in terms of sample size retention; in contrast to propensity score matching and other matching methods, it preserves a large majority of the study sample. Additionally, a wider range of confounders can be included compared to (coarsened) exact matching, where matching may result in a small proportion of matches when a large number of matching variables are included (Desai & Franklin, 2019; King et al., 2011).

Using logistic regression, we calculate propensity score weights as the inverse of the predicted response propensities for each respondent in the participating sample to mitigate the compositional differences between respondents who were asked to or chose to participate via VI (as a treatment group) and those who participated via in-person interviewing (as a control group). This creates a pseudo-population in which assignment to treatment is independent of the selected confounders (Chesnaye et al., 2022). In practice, VI participants who are more similar to in-person participants will receive larger weights, while VI participants who differ significantly from in-person participants will receive smaller weights, reflecting their reduced similarity.[12] After applying PSW, the weighted subsamples (treated and untreated) are balanced on the selected confounders, which means that the adjusted sample compositions are similar. This enables a more robust identification of measurement mode effects than without PSW adjustments. To ensure that the ignorable treatment assignment assumption (Rosenbaum & Rubin, 1983) holds as strongly as possible, we must identify confounders that effectively separate the samples based on characteristics associated with non-random mode allocation and self-selection. The confounders are presented in the section below.

Confounders in propensity score weighting analysis. After carefully reviewing the data and incorporating insights from previous BCS70 research, the following 34 confounders[13] from six BCS70 surveys conducted between 1970 and 2016 were identified and selected for inclusion in the propensity score model (with the purpose to derive propensity score weights). As indicated, we selected these as previously proven to be effective for nonresponse adjustment and known to be associated with the substantive outcome variables of interest:

- Socio-demographic variables: *father* and *mother staying at school after minimum school leaving age*, *parental social class at birth* (from Birth Survey data), *early life housing tenure* (from Age 10 data), *sex at birth, highest qualification, economic status, marital status, cohabiting as a couple, number of children, age of oldest child, household size*, *housing tenure, residence in London or Southeast England, socio-economic classification*, and *index of multiple deprivation* (from Age 46 Survey data)[14];

[12] For the treatment group, the weights are calculated as '1/propensity score', and for the control group, the weights are calculated as '1/(1−propensity score)'. The logistic regression model includes a binary variable 'group' (0=control, 1=treatment) and confounders as predictors.

[13] Missing values in the confounders were imputed to be able to calculate propensity score weights for all respondents. Given the number of carefully selected covariates, previously identified as effective for nonresponse adjustment and known to be associated with the outcome variables of interest, it can be assumed that the missing data mechanism is Missing At Random (MAR). Therefore, we employed random forest imputation using the missForest R package (Stekhoven & Bühlmann, 2012), which is suitable for handling different types of variables. The approach is consistent with the recommendations outlined in the User guide for handling missing data from the BCS70 data custodians (Silverwood et al., 2024). Importantly, these imputed data were not used for estimation, which was conducted to investigate measurement mode effects using the most recent Age 51 Survey data.

[14] Most of them were from the Age 46 Survey. Notably, most of the same socio-demographic variables from the Age 51 Survey, along with *time at current address, ethnicity* and p*articipating in the previous Age 46 BCS70 Survey*, will also be used as predictors in regression models to address RQ2.

- Topical variables: *breastfeeding* (from Age 5 data), *early life cognition* (from Age 10 data), *early life mental health* (from Age 16 data), *self-efficacy* and *having fixed place of work* (from Age 42 data), *total amount of savings*, *longstanding illness*, *debt, trust in other people, voting in previous general elections, how well managing financially, participant's earnings, total household income, number of jobs held,* and *cognitive assessment scores* (from Age 46 data).

Importantly, these survey items, which arguably distinguish between those assigned to/responded via VI and in-person interviewing, cover a range of topics that are also included in the analysis of mode effects. Based on the authors' previous experience with BCS70 data, the selected confounders are expected to be associated with nonresponse, as well as the outcome variables of interest (see Table 1 below). Notably, the confounders were not selected with the intention of mitigating selection bias for specific outcome variables, as the selection of items for the measurement mode effect analysis presented in the next section was conducted separately.

Survey items for measurement mode effect analysis. The selection of survey items included in our mode effect analysis (see Table 1 and Table A1 in the Appendix) was based on their relevance to longitudinal analysis in the BCS70 study, particularly considering how these items might be influenced by the introduction of a new survey mode during the COVID-19 pandemic. To minimize selection bias in the survey items included in the mode-effect analysis, and to ensure that our findings are not limited to specific question topics or other variable characteristics, we include a broad range of survey items in our analysis. An important criterion for inclusion of many of these items was the expectation that the distributions of the selected variables could be affected by the specific characteristics of VI, aligning with existing literature on measurement mode effects.

Given the limited evidence on such effects in mixed-mode studies incorporating VI (e.g. Asensio et al., under review; Conrad et al., 2023; Endres et al., 2023; West et al., 2022), we included a broad range of survey items to expand the evidence base. This approach also increases the generalizability of our findings compared to a more homogeneous selection of variables from a limited range of topics. The items selected for the measurement mode effect analysis cover a variety of topics, including cognition, education, employment, family, physical and mental health, wellbeing, housing, finances, and political attitudes & behaviour. For most items, and based on existing literature, we do not have specific theoretical expectations regarding measurement mode effects. Instead, we aim to explore general distributional differences across the two modes and explain them using phenomena such as social desirability bias, satisficing, answer order effects (such as primacy and recency effects), and acquiescence response bias. This makes our investigation exploratory rather than confirmatory in nature.

Table 1: Selected survey items for mode-effect analysis (103 variables)

Broad topic	Variables / measured concepts	
Finances	Number of benefits reported	Number of sources of savings reported
	Whether has pension	Total savings
	Whether contributing to pension	Number of debts reported
	Whether partner has pension	Received an inheritance or substantial gift
	Number of other sources of income reported	Total value of reported inheritances
	Income from other sources last month	Whether has made a will
	Total household income	
Employment	Work stress	Job security
	Economic activity status	Earnings
	Whether job SOC coded	Partner Economic Activity
	Hours	Whether partner job SOC coded
	Place of work	Partner net pay
	Physicality of work	Expected retirement age
	Job satisfaction	Economic Shocks
Health	Self-reported health	Alcohol
	How health has changed in last 12 months	Smoking
	Long-standing illness	Vaping
	Number of health problems reported	Weight in Kg
	Hospital visits	Whether trying to change weight
	Dental health	Self-assessment of weight
	Exercise	
Family	Whether in non-cohabiting relationship	Frequency of other contact with family
	Helping parents	Frequency of meeting friends
	Hours spent helping parents	Frequency of other contact with friends
	Frequency of meeting family	
Housing	Number of moves	Amount owed on mortgage
	Completeness of info– dates and postcodes	Damp
	Number of rooms	Car ownership
	Value of property	
Mental health & wellbeing	Mental health	Self-efficacy
	Loneliness	Worry about retirement
	Relationship satisfaction	Trust
	Life satisfaction	
Political attitudes & behaviour	Political interest	EU Referendum vote
	Whether voted in 2019 GE	Believing Britain will be better off outside EU
	Party voted for	Attitude to migration
	Who would vote for tomorrow	
Household	Domestic chores (cooking, shopping, etc.)	Whether pays someone to do domestic tasks
	Number of children reported	No. of other household members reported
Cognition	Cognitive test scores (e.g. word recall)	
Education	Number of qualifications reported	
Other topics	Social Provisions	Use of social media
	Interpersonal violence	

Regression modelling for measurement mode effect analysis. To address RQ3, the 103 survey variables listed in Table 1 - some of which were derived from a range of other survey items - will be included in the measurement mode effects analysis. These items/concepts will be treated as the dependent variables in the regression models, with survey mode indicator (categories: in-person, VI) as the exposure. Before and after applying the PSW-derived weights to both samples, i.e. in-person and VI, we apply a range of regressions in Stata 17, depending on the outcome variables investigated. These include binary, ordinal and multinomial logistic regression with categorical outcome variables, negative binomial and Poisson regression with count variables, Generalized Linear Models (GLM) with a gamma distribution and log link, as well as linear regression for continuous outcome variables. It is expected that statistically significant differences will be identified in the unadjusted data across a range of variables, most of which can be attributed to compositional differences. However, a statistically significant impact of survey mode on outcomes in the PSW-adjusted analysis will indicate the possibility of differences which may be genuinely attributable to measurement mode effects. We report significance at three levels: $p<0.05$, $p<0.01$, and $p<0.001$.

Additionally, as a robustness check, we will also apply an instrumental variable approach for linear regression and probit modelling (specifically, two-stage least squares, 2SLS). Our preliminary analysis showed a strong association between 'mode of allocation' and 'mode of participation', and the evidence also suggested that 'mode of participation' is not associated with socio-demographic characteristics. This allows us to use 'mode of allocation' as an instrument to estimate the causal effect of 'mode of participation' on topical outcome variables.

3.3. Assumptions and principles for interpreting measurement mode effects

Given the study's quasi-experimental design and the need to use a method for causal inference, and given available confounders from previous surveys, our approach requires making certain assumptions and applying specific principles and criteria for interpreting results when identifying measurement mode effects. First, our approach is based on the assumption that, conditional on the selected covariates, there are no unobserved confounders influencing both mode of participation and the outcomes (conditional exchangeability). We also assume positivity, meaning that every covariate pattern has a non-zero chance of each survey mode, and consistency, meaning that the observed outcome under the actual mode equals the potential outcome for that mode (see Imbens & Rubin, 2015; Rosenbaum & Rubin, 1983). Second, our interpretation is guided by the expectation that measurement mode effects for the majority of outcomes will be minimal, consistent with existing literature on live video interviews compared to in-person (see Asensio et al., under review; Conrad et al., 2023; Endres et al., 2023; West et al., 2022), unless compelling evidence suggests otherwise. Third, to identify and interpret mode effects, we consider the statistical significance of differences between modes (VI and in-person), the magnitude and direction of associations (pre- and post-PSW), theoretical plausibility and interpretability (i.e. consistency with existing literature on measurement mode effects), and evidence from robustness checks (i.e. instrumental variable approach).

Importantly, interpretation may need to account for whether the limited effect of PSW in mitigating differences between VI and in-person groups is due to the absence of one or more confounders sufficiently associated with the outcome variable of interest, which could manifest in a number of items from a particular topic showing minimal adjustment. For example, if there are notable compositional differences between VI and in-person groups in health outcomes, and only limited health-related confounders are available from previous waves, PSW may have limited ability to mitigate selection-related differences between modes. In such cases, remaining differences should be interpreted with caution rather than assumed to reflect measurement effects, especially when items unlikely to be affected by measurement mode are grouped with others, all of which continue to show meaningful differences after adjustment.

4. Results

In this section, we will address our three research questions. First, we will discuss nonresponse patterns based on mode of allocation (RQ1). Second, we will examine the key socio-demographic characteristics of respondents who participated via VI and those who participated in person, and compare these characteristics with those of the groups originally allocated to each mode (RQ2). Lastly, we will investigate the differences in measurement outcomes between VI and in-person interviewing as the central part of our empirical investigation (RQ3).

4.1. Nonresponse patterns and mode-switching by mode of allocation

As previously discussed, BCS70 study members were allocated to either VI-first or in-person-first groups, with this allocation being relatively close to random, although respondents could complete the survey via a different mode to that which they were initially allocated (see the 'Methods' section for more information). The nonresponse pattern analysis results are presented in Table 2 to address RQ1. Of the 11,934 eligible study members with mode allocation information 6,870 were allocated to VI-first in waves 1-4, and 5,064 were allocated to in-person-first in waves 5-7. The final refusal rate was higher in the in-person-first group (23.5%) than in the VI-first group (20.7%) (χ^2=13.27, p<0.001), but the final noncontact rate was higher in the VI-first group (in-person-first: 4.7%, VI-first: 6.3%, χ^2=14.48, p<0.001). Additionally, the proportion of in-person-allocated study members with other nonresponse survey outcome (10.9%) was substantially larger than for VI (4.5%) (χ^2=173.8, p<0.001). Notably, Chi-Square testing confirmed statistically significant differences in the mode of response, with VI-allocated respondents being about 10 times more likely to switch modes (VI: 14.5%, in-person: 1.4%, χ^2=616.29, p<0.001).

Table 2: Nonresponse and mode-switching analysis, by mode of first allocation, n=11,934

Mode of first allocation		Refusal	Non-contact	Other nonresponse	Response via alternative mode	Total
In-person	***n***	1,191	237	549	69	5,064
	%	23.5%	4.7%	10.9%	1.4%	100.0%
VI	***n***	1,424	433	311	993	6,870
	%	20.7%	6.3%	4.5%	14.5%	100.0%
Total	***n***	*2,615*	*670*	*860*	1,062	*11,934*
	%	*21.9%*	*5.6%*	*7.2%*	8.9%	*100.0%*

These results confirm the effects of allocation mode on both nonresponse rates and participation mode, which is expected to have some influence on the composition of the samples if nonresponse and mode-switching are differential. In subsequent analysis, we will focus on those other effects of mode by comparing in-person interviewing and VI. The final sample, as our analytical sample for measurement mode effects analysis, included 3,498 (47.9%) survey respondents who participated via VI and 3,812 (52.2%) who participated in-person.

4.2. Socio-demographic characteristics of respondents by modes

To answer the second research question on the key characteristics of respondents participating via VI and in-person, we carried out binary logistic regression analysis with the following outcome variables in two models:

- Model 1: mode of allocation (VI=1 versus in-person=0) and
- Model 2: mode of participation (VI=1 versus in-person=0).

Table 3 presents the results. Overall, the socio-demographic profiles of participants allocated to VI and in-person interviewing were very similar, indicating that allocation to mode was close to random as expected.[15] The only notable exceptions are variables associated with the ease of contacting participants, namely housing tenure and time at current address. Renters were less likely to be assigned to VI, while those who had lived at their address for 10 years or less were more likely to be assigned to VI.

On the other hand, more substantial socio-demographic differences emerged when comparing the final mode of completion. This was expected given mode-switching and potential differential nonresponse. Firstly, those who had lived at their address for longer were more likely to be allocated to in-person interviewing and complete the interview via that mode. Secondly, several socio-demographic groups were significantly less likely to participate via VI, even though they were not less likely to be assigned to VI. These include men, the less educated (Level 1 or no qualification), the long-term sick/disabled (compared to full-time employees), single/never married individuals (compared to those with a partner), and renters (compared to owners with a mortgage). In addition, participation in the previous BCS70 Age 46 sweep, was positively associated with completing the Age 51 interview by VI. Taken together, these patterns indicate clear socio-demographic differences between those who ultimately participated via VI versus in-person. On this basis, it was appropriate to include these variables as confounders in the propensity score weighting procedure used to adjust for non-random mode allocation and self-selection into mode.

[15] It is important to note that the most recent socio-demographic data were only available for BCS70 Age 51 participants.

Table 3: Binary logistic regression analyses with survey mode (of allocation and of participation) as outcome variables (VI=1 versus in-person=0) and socio-demographic explanatory variables (n=7,122).

Explanatory variables	**Model 1: Mode of allocation (n=7,122)**	**Model 2: Mode of participation (n=7,122)**
	Coef. *(SE)*	Coef. *(SE)*
Sex (ref. group: *Female*)		
Male	-0.034 *(0.055)*	-0.154** *(0.055)*
Highest qualification (ref. group: *Level 4: Graduate degree*)		
No qualification	-0.156 *(0.099)*	-0.430*** *(0.103)*
Level 1 - GCSE lower grades or equivalent	-0.073 *(0.102)*	-0.282** *(0.104)*
Level 2 - GCSE higher grades or equivalent	0.059 *(0.067)*	-0.113 *(0.066)*
Level 3 - A-level or equivalent	0.101 *(0.076)*	0.074 *(0.075)*
Level 5 - Postgraduate degree	-0.123 *(0.087)*	0.003 *(0.086)*
Economic status (ref. group: *Full-time paid employee*)		
Part-time paid employee	-0.003 *(0.077)*	-0.034 *(0.076)*
Full-time self-employed	-0.097 *(0.078)*	-0.099 *(0.077)*
Part-time self-employed	0.034 *(0.127)*	0.056 *(0.125)*
Unemployed and seeking work	-0.125 *(0.220)*	-0.443 *(0.234)*
Long term sick/disabled	-0.094 *(0.128)*	-0.604*** *(0.141)*
Looking after home or family	0.118 *(0.125)*	-0.077 *(0.124)*
Wholly retired	-0.031 *(0.248)*	-0.274 *(0.247)*
Something else	-0.490* *(0.217)*	-0.773** *(0.234)*
Marital status (ref. group: *Without a partner*)		
Single (never married)	-0.055 *(0.076)*	-0.217** *(0.076)*
Married	-0.128 *(0.106)*	-0.217* *(0.107)*
Divorced	0.051 *(0.096)*	-0.055 *(0.095)*
Legally separated	0.190 *(0.220)*	0.103 *(0.217)*
Other	0.172 *(0.252)*	0.353 *(0.249)*
Cohabiting as a couple (ref. group: *Yes*)		
No	-0.049 *(0.086)*	-0.069 *(0.087)*
Having children (ref. group: Yes)		
No	0.019 *(0.098)*	0.033 *(0.099)*
Age of oldest child (ref. group: *18-25 years*)		
Less than 10 years	-0.055 *(0.163)*	0.097 *(0.161)*
10-17 years	-0.044 *(0.070)*	0.109 *(0.069)*
More than 25	-0.043 *(0.088)*	-0.095 *(0.089)*
Household size (ref. group: *4 persons*)		
1 person	0.025 *(0.156)*	0.001 (*0.161)*
2 persons	0.017 *(0.102)*	-0.103 *(0.103)*
3 persons	-0.051 *(0.069)*	-0.167* *(0.069)*
5 or more persons	-0.071 *(0.079)*	-0.180* *(0.079)*
Housing tenure (ref. group: *Owning dwelling with mortgage*)		
Owning without mortgage	-0.058 *(0.060)*	-0.047 *(0.059)*
Rent it	-0.186* *(0.078)*	-0.455*** *(0.079)*
Live rent free	-0.144 *(0.187)*	-0.022 *(0.190)*
Other	-0.146 *(0.198)*	-0.164 *(0.198)*
Time at current address (ref. group: *>15 and 20 years*)		
5 years or less	0.658 *(0.445)*	0.885* *(0.415)*
Between >5 and 10 years	0.494* *(0.150)*	0.812*** *(0.147)*
Between >10 and 15 years	0.036 *(0.066)*	0.036 *(0.065)*
Between >20 and 25 years	-0.131 *(0.067)*	-0.152* *(0.067)*
More than 25 years	-0.217* *(0.079)*	-0.362*** *(0.080)*
Ethnic group (ref. group: *White*)		
Mixed	-0.191 *(0.365)*	-0.082 *(0.369)*
Asian	-0.085 *(0.213)*	0.297 *(0.215)*
Black	0.422 *(0.410)*	0.546 *(0.402)*
Other	-0.088 *(0.079)*	-0.093 *(0.081)*
Participating in the previous Age 46 BCS70 Survey (ref. group: *Yes*)		
No	0.065 *(0.088)*	0.374*** *(0.091)*
Constant	0.557*** *(0.117)*	0.064*** (*0.091)*
Pseudo R-Square (McFadden R^2)	*0.008*	*0.035*

**p<0.05, **p<0.01, ***p<0.001; Coef.=coefficient, SE=standard error*

4.3. Measurement mode effects analysis

To address the central research question, RQ3, on the measurement differences attributable to mode effects, we conducted various types of regressions with the selected substantive variables of interest from the BCS70 Age 51 survey based on their type and distribution. The evidence, presented in Table A1 in Appendix A, is based on regression modelling, with ordinal logistic (n=26), binary logistic (n=21), multinomial logistic (n=21), linear (n=17), negative binomial (n=8), GLM gamma with log link (n=7), and Poisson (n=3) regressions. Importantly, we also compared the differences in distributions of target variables between unweighted models (no causal inference models used) and weighted models with propensity score weights applied.

First, we can observe compositional differences in the unadjusted data (i.e. before PSW-adjustments) for a large proportion of outcome variables (see Table A1, unweighted data analysis). This initial finding is consistent with the evidence presented to address RQ2 (see Table 2), which also confirmed socio-demographic differences in the unadjusted data. Such differences between VI and in-person samples, most of which can arguably be explained by self-selection to mode in the form of mode switching (quasi-experimental design), also exist for most of the 103 target variables presented in Table A1. For 32.0% of all variables, we observed no differences between VI and in-person respondents, but for 42.7% of all variables, the differences were significant at the p<0.001 level. The magnitude of these differences was mostly small.

After adjusting the compositions using PSW as a method for causal inference, these differences were largely mitigated for the majority of these variables. After PSW adjustment, there were no statistically significant differences between VI and in-person samples for 56.3% of variables, and differences were significant at the p<0.001 level for 17.5% of all variables. This indicates that the adjustment was effective, suggesting that PSW weighting successfully controlled for differences between the two groups. This is also evidenced by the fact that even when differences exist in the fully adjusted models, PSW based on the 34 carefully selected confounders mostly mitigated the magnitude of the differences, even though they often remained statistically significant at the p<0.05 level.[16]

Second, after PSW adjustment, we observe that the topics exhibiting the largest number of items with significant differences between VI and in-person respondents are *cognition, political attitudes & behaviour*, and *finances*. The other topics with notable differences between the modes are *social capital wellbeing*, *health* (particularly healthy living survey items), and *family-orientation*. It is noteworthy that *mental health & wellbeing*, *political attitudes & behaviour*, and *social provisions* were administered using two different types of self-completion – in-person respondents completed it on a

[16] The results from our sensitivity analysis using the instrumental variable approach indicated that using ‘mode of allocation’ as an instrumental variable, which appeared to be exogenous and not correlated with confounders, mitigated the effect of ‘mode of participation’ on the outcomes, similar to the results obtained through PSW (with conclusions regarding measurement mode effects remaining largely the same). We conclude that the near-random or exogenous nature of ‘mode of allocation’ successfully isolated the variation in ‘mode of participation’ that was not confounded, enabling a more accurate estimation of its causal effect. This mitigated effect suggests that some of the original associations, interpreted as measurement mode effects, were likely driven by confounding. Unlike PSW, which adjusts for observed confounding, the instrumental variable approach also addresses unobserved confounding. Overall, the findings on the presence of measurement mode effects were similar to those drawn from PSW adjusted data. For example, we identified statistically significant differences in cognition scores between VI and in-person participants for two out of four outcome variables, including the letter cancellation test.

laptop handed to them by the interviewer, and VI respondents accessed a web-based version.[17] For the listed topics, we observed the following differences, with particular focus on those that may be attributed to measurement mode effects:

- ***Political attitudes & behaviour:*** For the topic showing the greatest differences between VI and in-person respondents, we generally identified differences that may reflect higher levels of conservatism among those who participated via VI. This includes a greater support for the Conservative Party (compared to Labour) and the sentiment that Great Britain would be better off outside the EU.[18] Some of the expressed views may have been less popular at the time, which suggests that some VI respondents might be more candid in their responses compared to in-person respondents.
- ***Cognition***: Differences were identified for all four cognition score variables, with scores being higher for VI participants. These differences were observed even after PSW adjustment, despite the inclusion of the same four scores from the Age 46 Survey as confounders. Higher scores in the letter cancellation task, which was the score with the greatest difference between VI and in-person respondents,[19] could potentially be explained by respondents opening the envelope and completing the assessment prior to the virtual interview which was not possible during in-person interviews. However, slightly higher scores in other assessments are more difficult to explain through procedural differences alone.
- ***Social capital wellbeing***: VI respondents reported lower levels of trust in other people and life satisfaction.[20] This appears to be another example of more candid responding in a virtual setting when disclosing information about psychosocial stressors (a self-completion module). This interpretation is further supported by VI respondents reporting higher levels of work stress and lower perceived job security (an interviewer-administered section).
- ***Finances***: VI respondents were more likely to report additional sources of income and pensions,[21] and express less worry about retirement and associated savings, despite no observed differences in total household income or savings. We argue that these differences are difficult to explain with "traditional" measurement mode effect theories discussed in the

[17] We did not find evidence that the differences between the self-completion modes (completion on a laptop in the presence of an interviewer vs. web completion) were more or less common than those observed between the interviewer-administered modes (in-person vs. video).
[18] We report predictive margins as average predicted probabilities for the groups from a previously fit to provide examples of relatively large differences between VI and in-person respondents. For *whether respondent believes Britain will be better off outside EU* variable, the following predictive margins were calculated with a logistic regression model with socio-demographic controls: unadjusted data, 'Yes': VI = 35.5%, in-person = 41.3%; PSW adjusted data, 'Yes': VI = 34.9%, in-person = 41.5%.
[19] Predictive margins with *letter cancellation score* (range=5-65 points) as the outcome variable were calculated to provide another example of relatively large differences between the group: unadjusted data: VI = 29.2 points, in-person = 27.7 points; PSW adjusted data: VI = 29.1 points, in-person = 27.8 points.
[20] These two items stood out, showing differences between VI and in-person respondents that increased after PSW adjustment, in contrast to the vast majority of other items where differences were mitigated. As an example, we calculated the following predictive margins for *life satisfaction* (range 0-10): unadjusted data: VI = 7.43, in-person = 7.23, PSW adjusted data: VI = 7.46, in-person = 7.24. This provides supporting evidence for the presence of mode effects when measuring psychosocial stressors.
[21] As another example of a relatively large difference between the VI and in-person groups compared to other outcome variables, we calculated predictive margins (with socio-demographic controls) with *having a pension* as the outcome variable: unadjusted data: VI = 89.3%, in-person = 85.6%; PSW adjusted data: VI = 89.2%, in-person = 86.5%.

existing literature, such as satisficing, social desirability, acquiescence response bias or order effects.

- ***Healthy living:*** There were a range of items across different topics for which we identified differences, with VI respondents consistently reporting better overall and dental health, as well as generally healthier lifestyle choices (less alcohol consumption and smoking/vaping, and efforts to lose weight). However, these differences appear to reflect compositional differences between in-person and VI respondents, which may be attributed to the health topic being less represented among the selected confounders (see 'Confounders in propensity score weighting analysis' section). Similar to the *finances* topic, it would be challenging to attribute the remaining differences to "traditional" measurement mode effect theories.
- ***Family-orientation:*** We also identified particular items from the *family, household*, and *social provision* topics that fundamentally reflect how family-oriented respondents are. In addition to reporting having more children on average, VI respondents were more likely to help their parents, had more frequent contact with family members, and reported stronger networks of close individuals who provide safety, security, and advice. Similar to the *finances* and *healthy living* topics, the differences between the modes appear to reflect compositional differences rather than measurement differences. While PSW adjustment notably reduced the magnitude of these differences, it did not fully eliminate them; they appear to be attributable to residual sample composition effects arising from unobserved confounders.

5. Discussion and conclusion

VI is an emerging survey mode that still requires extensive methodological research to assess its feasibility and to develop strategies for improving data quality. Our study is among the first to investigate VI-specific mode effects (e.g. Conrad et al., 2023; West et al., 2022). We focused primarily on identifying measurement differences between VI and in-person interviewing by comparing distributions across more than 100 survey variables covering a wide range of topics and measurement formats. To our knowledge, only four other studies compared measurement outcomes between VI and in-person interviews (see Asensio et al., under review; Conrad et al., 2023; Endres et al., 2023; West et al., 2022). We argue that our study represents the most comprehensive effort to date, drawing on large samples of VI and in-person respondents and taking advantage of a natural experiment created by the COVID-19 pandemic, during which VI temporarily replaced face-to-face fieldwork in a major cohort study. As such, our findings offer valuable evidence on the suitability of using VI alongside in-person interviewing without compromising measurement comparability. The key conclusion of our study is that, from a measurement perspective alone, VI should be considered as viable alternative for in-person interviewing. Across the vast majority of survey items, differences between the two modes are minimal and most of the differences are largely attributable to socio-demographic differences caused by non-random (self) selection into mode. By applying a method of causal inference (i.e. PSW) to adjust for these compositional differences, we are able to identify the effect of mode that could be attributed to measurement, which constitutes the other notable contribution of our research. We provide more detail and discuss the relevance and practical implications of these findings in the following paragraphs.

Our analysis highlights that mode differences between VI and in-person interviewing are greater for nonresponse and mode-switching patterns (RQ1) and associated sample composition (RQ2) than for measurement (RQ3). Specifically, refusal and noncontact rates differ between the two modes, and we observe a notably higher proportion of respondents switching from VI allocation to completion in-person than in the opposite direction. The group allocated to VI and the group allocated to in-person were nearly identical in terms of socio-demographic composition, but we still find compositional differences between VI and in-person participants which are attributable to mode-switching. The socio-demographic groups more likely to switch modes include men, individuals who are single and have never been married, those with lower levels of education, renters, and people who are not employed due to their long-term sickness or disability. The associations between education and marital status with VI participation are generally consistent with findings from existing research (e.g. Conrad et al., 2023; Dulaney et al., 2023; Durrant et al., 2024; Guggenheim & Howell, 2021), while the evidence regarding the effect of gender is more mixed (e.g. Dulaney et al., 2023; Durrant et al., 2024; Guggenheim & Howell, 2021; Schober et al., 2023). These compositional differences necessitated the use of a causal inference method to more robustly isolate the impact of mode on measurement.

After using PSW to adjust for non-random allocation to modes combined with mode self-selection, we observed certain measurement differences between VI and in-person interviewing in the form of differing distributions of outcome variables. Below, we highlight the key findings on measurement mode effects and relate them to existing literature:

1. The differences between VI and in-person administration (i.e. for sections that were interviewer administered) were similar to those between the two self-completion modes (i.e. computer-assisted self-interviewing, as part of in-person interviews, and web, as part of VI interviews). This was somewhat unexpected as we anticipated greater differences between the interviewer-administered modes.
2. Respondents appeared to be more candid in their responses when participating virtually compared to in-person, particularly during the self-completion portion which contains topics of a more sensitive nature. We consider this a positive aspect of virtual administration. VI respondents appeared to be more willing to report less popular views, such as conservative political opinions, and to disclose greater psychological stressors, such as lower life satisfaction. This may be explained by the absence of a physically present interviewer in the case of video administration. It aligns with the nature of web-based self-completion, where the lack of physical interviewer presence may encourage more honest reporting compared to self-completion conducted in the interviewer's presence, as previously reported for psychological functioning, including life satisfaction (Zager Kocjan et al., 2023).
3. The higher cognitive scores observed among VI respondents could be attributed to reduced pressure when the interviewer was not physically present, potentially leading to better performance. Ofstedal et al. (2021), who reported differences in cognitive scores between in-person and web administration, discussed similar reasons. Furthermore, one of the scores in our study may also have been influenced by procedural differences, as VI respondents received the envelope containing the test prior to the virtual interview and test completion.
4. However, we argue that most of the remaining differences, such as those for *finances* and *healthy living* topics, cannot be easily attributed to methodological and/or technical variations in survey administration. We found no evidence that these factors increased the propensity for social desirability bias, acquiescence response bias, satisficing, or order effects that could lead to distributional differences between the modes. What is more, findings on the better health

outcomes of respondents participating virtually, even after adjustment, are somewhat consistent with previous research that compared health status of members of an online panel with push-to-web respondents selected from an address-based sample frame (see Sherr et al., 2024). Online or virtual participation can be explained by factors such as internet access and literacy, and potentially interest in health.

Generally speaking, using a method for causal inference proved to be quite effective in controlling for the differences between the groups, as it substantially reduced the number of variables with statistically significant differences between VI and in-person to a total of 18 outcome variables (at $p<0.001$). It also decreased the magnitude of those differences for most included items. We believe that incorporating further confounders from the same domains (if available), particularly those associated with other survey items showing the effects of mode, would likely mitigate most of the remaining differences in distributions between VI and in-person interviewing. Even so, the existing evidence from our study strongly suggests that the magnitude of measurement differences between VI and in-person interviewing is minimal and can largely be explained by socio-demographic compositional differences (associated with topical differences) inherent in a quasi-experimental design. This is not surprising, given that both of the investigated survey modes are interviewer-administered and conducted face-to-face, and similar findings on measurement mode effects were reported by Endres et al. (2023). We agree with the conclusion from Brown et al. (2025), who argue that while adjustments might need to be considered for specific tasks (e.g. cognitive assessment), substantial adaptation of in-person questionnaires for VI is unlikely to be required.

One of the important findings, albeit not central to our empirical investigation, is the effectiveness of a method for causal inference in investigating measurement mode effects, and more broadly, survey errors in quasi-experimental survey experimental designs. These types of methods have not been commonly used, and when they were applied in practice, some of these methods received certain criticism. This predominantly applies to matching methods like propensity score matching or Mahalanobis distance matching (e.g. Kibuchi et al., 2024; Kocar et al., 2021). In turn, we followed recommendations from previous research on methods for causal inference (e.g. Austin & Stuart, 2017; Desai & Franklin, 2019; King & Nielsen, 2019) and used PSW, more specifically the inverse probability of treatment weighting (Czajka et al., 1992). This method proved to be quite efficient in mitigating non-random allocation combined with self-selection bias, and our assessment of efficiency is based on the following: (i) PSW decreased the distributional differences between VI and in-person interviewing that could not be explained with existing theory on measurement mode effects, (ii) the change was in the direction of a smaller magnitude for the vast majority of items with statistically significant differences before applying PSW, (iii) supplementary analysis using an instrumental variable approach – an alternative quasi-experimental method when the random mode of allocation largely determines the mode of participation and adjusts for both observed and unobserved confounding – mitigated bias similarly to PSW. Therefore, the overall evidence from this study suggests that PSW should be employed more widely in quasi-experimental designs used to explore survey errors, provided that the necessary confounders are available.

Lastly, we need to discuss further opportunities for methodological research in this area and acknowledge certain limitations of our study. The limitations are largely associated with the quasi-experimental design and the selection of variables included in our models, both as confounders used for PSW and as outcome variables used to explore measurement mode effects. Below, we provide more details on specific limitations:

1. Quasi-experimental designs inevitably have shortcomings compared to fully experimental survey designs based on the principles of randomised control trials. Randomly allocating birth cohort participants into mode was not feasible during the peak of a pandemic when in-person interviews could not take place. Although the quasi-experimental design allows us to explore a wider range of phenomena related to survey errors and data quality, we cannot rule out unobserved confounding or be entirely certain that the adjustments made by using PSW can fully mitigate all bias. This remains true despite using theory to confirm whether the PSW-introduced changes are consistent with previous research on such effects of mode. Therefore, future research could include both testing for measurement differences between VI and in-person interviewing using a fully experimental design and with no mode-switching possible (see also Asensio et al., under review).
2. The selection of confounders is key, and different specifications might have produced somewhat different estimates. Although we aimed to select outcome variables (and therefore confounders) across a broad range of topics to maximise generalisability, we were constrained by the variables available in BCS70. Some potentially important confounders could not be included, and additional outcome domains would likely strengthen the robustness of future studies.
3. Additionally, our investigation does not examine measurement mode effects using panel data analysis methods, such as fixed- and random-effects models, in a longitudinal context across time/waves, which is particularly relevant for longitudinal and birth cohort studies. This remains an important avenue for future research.
4. A further limitation concerns the age range of participants included in our study. All respondents were members of the BCS70 cohort and were therefore aged 51 at the time of data collection. This is an important contextual factor, as the feasibility, acceptability, and behavioural implications of VI may differ substantially across age groups. For example, younger adults may be more accustomed to video-mediated communication, potentially leading to different patterns of disclosure, engagement, or mode switching, while older adults may face greater technological barriers or feel less comfortable completing interviews via video. As such, our findings may not be fully generalisable to younger cohorts, older populations, or age-diverse general population samples. Future studies should therefore examine VI mode effects across a broader range of age groups to determine whether the patterns observed here, particularly the minimal measurement differences, hold consistently for other age groups.

Nevertheless, the implications of our findings for survey practice are important and have wide ranging relevance at a time where many longitudinal studies are considering different, additional or alternative modes to in-person data collection. The novel evidence from our study clearly demonstrates that while certain types of measurement mode effects are present, measurement differences between VI and in-person interviewing are generally minor. We are confident in suggesting that longitudinal studies which have predominantly used in-person interviewing can supplement it with VI, at least for a proportion of the sample, without a notable threat of measurement mode effects affecting their data quality in measuring change over time.

References

Anderson, A. H. (2008). Video-mediated interactions and surveys. Envisioning the survey interview of the future, 95-118.

Asensio, M., Brown, M., Silverwood, R., Hanson, T., & Durrant, G. (under review). Assessing the impact of video interviewing on measurement quality: Evidence from an experimental study on mode effects.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, *46*(3), 399-424.

Austin, P. C., & Schuster, T. (2016). The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: a simulation study. *Statistical methods in medical research*, *25*(5), 2214-2237.

Austin, P. C., & Stuart, E. A. (2017). The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical methods in medical research*, *26*(4), 1654-1670.

Beulens, B., van der Laan, J., Schouten, B., van der Brakel, J., Burger, J., & Klausch, T. G. (2012). *Disentangling mode-specific selection and measurement bias in social surveys* (Discussion paper). Statistics Netherlands. The Hague.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Stürmer, T. (2006). Variable selection for propensity score models. *American journal of epidemiology*, *163*(12), 1149-1156.

Brown, M., Hanson, T., Asensio, M., Cornesse, C., Wood, M., Spencer, S., Sanchez, C., Koerber, H., & Durrant, G. (2025). Video interviewing, Survey Futures Survey Practice Guide No. 4. University of Essex. https://surveyfutures.net/practice-guides/

Centeno, L., Arrue, J., Good, C., Edwards, B., & Steiger, D. (2024). Practical considerations for developing and implementing a video mode for survey data collection. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *164*(1), 122-146.

Cernat, A. (2015). The impact of mixing modes on reliability in longitudinal studies. Sociological Methods & Research, 44(3), 427-457.

Chesnaye, N. C., Stel, V. S., Tripepi, G., Dekker, F. W., Fu, E. L., Zoccali, C., & Jager, K. J. (2022). An introduction to inverse probability of treatment weighting in observational research. *Clinical kidney journal*, *15*(1), 14-20.

Conrad, F. G., Schober, M. F., Hupp, A. L., West, B. T., Larsen, K. M., Ong, A. R., & Wang, T. (2023). Video in survey interviews: Effects on data quality and respondent experience. Methods, data, analyses, 17(2), 135-170.

Czajka, J. L., Hirabayashi, S. M., Little, R. J., & Rubin, D. B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business & Economic Statistics*, *10*(2), 117-131.

D'Agostino Jr, R. B. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine*, *17*(19), 2265-2281.

De Leeuw, E., & van der Zouwen, J. (1988). Data quality in telephone and face-to-face surveys: A comparative analysis. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 283–299). John Wiley & Sons.

Desai, R. J., & Franklin, J. M. (2019). Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. *bmj*, *367*.

Desai, R. J., Rothman, K. J., Bateman, B. T., Hernandez-Diaz, S., & Huybrechts, K. F. (2017). A propensity-score-based fine stratification approach for confounding adjustment when exposure is infrequent. *Epidemiology*, *28*(2), 249-257.

Dillman, D. A., & Christian, L. M. (2005). Survey mode as a source of instability in responses across surveys. Field Methods, 17(1), 30-52.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. L. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet. Social science research, 38(1), 1-18.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, phone, mail, and mixed-mode surveys: The tailored design method. *Indianapolis, Indiana*, *17*.

Dulaney, R., Kelley, J., Arrue, J., & Edwards, B. (2023, July 17-21). Who is Choosing Video Interviewing? A Look into Respondent Characteristics of Video Adopters [Conference presentation]. 10th Conference of the European Survey Research Association. Milan, Italy.

Durrant, G., Kocar, S., Brown, M., Hanson, T., Sanchez, C., Wood, M., Taylor, K., Tsantani, M., & Huskinson, T. (2024). *Live video interviewing: Evidence of opportunities and challenges across seven major UK social surveys*. University of Essex. https://www.iser.essex.ac.uk/wp-content/uploads/files/working-papers/survey-futures/2024-01.pdf

Endres, K., Hillygus, D. S., DeBell, M., & Iyengar, S. (2023). A randomized experiment evaluating survey mode effects for video interviewing. Political Science Research and Methods, 11(1), 144-159.

Freedman, D. A., & Berk, R. A. (2008). Weighting regressions by propensity scores. *Evaluation review,* 32(4), 392-409.

Goodman, A., Brown, M., Silverwood, R. J., Sakshaug, J. W., Calderwood, L., Williams, J., & Ploubidis, G. B. (2022). The Impact of Using the Web in a Mixed-Mode Follow-up of a Longitudinal Birth Cohort Study: Evidence from the National Child Development Study*. Journal of the Royal Statistical Society Series A: Statistics in Society, 185*(3), 822-850. https://doi.org/10.1111/rssa.12786

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). Survey methodology. John Wiley & Sons.

Guggenheim, L., & Howell, D. (2021, January 27). Live Video Interviewing in the 2020 ANES Time Series Study [Seminar presentation]. Joint Program in Survey Methodology Seminar Series. College Park, Maryland, United States.

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *International Statistical Review*, *78*(1), 3-20.

Jeannis, M., Terry, T., Heman-Ackah, R., & Price, M. (2013). Video Interviewing: An exploration of the feasibility as a mode of survey application. Survey Practice, 6(1).

Kibuchi, E., Sturgis, P., Durrant, G. B., & Maslovskaya, O. (2024). The efficacy of propensity score matching for separating selection and measurement effects across different survey modes. *Journal of Survey Statistics and Methodology*, *12*(3), 764-789.

King, G., Nielsen, R., Coberley, C., Pope, J. E., & Wells, A. (2011). Comparative effectiveness of matching methods for causal inference. Unpublished manuscript, Institute for Quantitative Social Science, Harvard University, Cambridge, MA.

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political analysis*, *27*(4), 435-454.

Kocar, S., & Biddle, N. (2020). Panel mixed-mode effects: does switching modes in probability-based online panels influence measurement error?. *ANU CSRM and SRC Methods Paper No.1/2020*. Australian National University, Canberra.

Kocar, S., Biddle, N., & Phillips, B. (2021). The Effects of Mode on Answers in Probability-Based Mixed Mode Online Panel Research: Evidence and Matching Methods for Controlling Self-Selection Effect in a Quasi-Experimental Design. *ANU CSRM and SRC Methods Paper No.1/2021*. Australian National University, Canberra.

Krosnick, J. A., Narayan, S., & Smith, W. R. (1996). Satisficing in surveys: Initial evidence. *New directions for evaluation*, *1996*(70), 29-44.

Li, F., Thomas, L. E., & Li, F. (2019). Addressing extreme propensity scores via the overlap weights. *American journal of epidemiology*, *188*(1), 250-257.

Li, L., & Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *The international journal of biostatistics*, *9*(2), 215-234.

Li, Y. (2024). Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples. *Survey Methodology, 50*(1), 37–55. Statistics Canada.

Mercer, A. W., Kreuter, F., Keeter, S., & Stuart, E. A. (2017). Theory and practice in nonprobability surveys: parallels between causal inference and survey inference. *Public Opinion Quarterly*, *81*(S1), 250-271.

Ofstedal, M. B., McClain, C. A., & Couper, M. P. (2021). Measuring cognition in a multi-mode context. Advances in longitudinal survey methodology, 250-271.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Reichardt, C. S. (2009). Quasi-experimental design. *The SAGE handbook of quantitative methods in psychology*, 46-71.

Pennay, D., Phillips, B., Neiger, D., Ward, A., Slamowicz, S., & Lethborg, A. Results from the 2022 Australian Comparative Study of Survey Methods (ACSSM). *ANU CSRM and SRC Methods Paper No.1/2024*. Australian National University, Canberra.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Schober, M. F. (2018). The future of face-to-face interviewing. Quality Assurance in Education, 26(2), 290-302.

Schober, M. F., Conrad, F. G., Hupp, A. L., Larsen, K. M., Ong, A. R., & West, B. T. (2020). Design considerations for live video survey interviews. Survey Practice, 13(1).

Schober, M. F., Okon, S., Conrad, F. G., Hupp, A. L., Ong, A. R., & Larsen, K. M. (2023). Predictors of willingness to participate in survey interviews conducted by live video. Technology, Mind, and Behavior. https://doi.org/10.1037/tmb0000100.

Schouten, B., van den Brakel, J., Beulens, B., van der Laan, J., & Klausch, T. (2013). Disentangling mode-specific selection and measurement bias in social surveys. *Social Science Research, 42*(6), 1555–1570.

Sherr, S. A., Panas, K., & Call, K. (2024, July 10). *The health of health panel data: Comparing survey outcomes between probability panelists and address-based sample respondents*. SSRS. https://ssrs.com/insights/the-health-of-health-panel-data-comparing-survey-outcomes-between-probability-panelists-and-address-based-sample-respondents/

Silverwood, R., Narayanan, M., Dodgeon, B., Katsoulis, M., & Ploubidis, G. (2024). *Handling missing data in the CLS cohort studies: User guide*. London: UCL Centre for Longitudinal Studies. https://cls.ucl.ac.uk/wp-content/uploads/2020/04/Handling-Missing-Data-User-Guide-2024.pdf

Sizemore, S., & Alkurdi, R. (2019). Matching methods for causal inference: A machine learning update. Retrieved from https://humboldt-wi.github.io/blog/research/applied_predictive_modeling_19/matching_methods/

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112-118.

Sullivan, A., Brown, M., Hamer, M., & Ploubidis, G. B. (2022). Cohort Profile Update: The 1970 British Cohort Study (BCS70). International Journal of Epidemiology. https://doi.org/10.1093/ije/dyac148

Suzer-Gurtekin, Z.T., Valliant, R., Heeringa, S.G., & de Leeuw, E.D. (2018). Mixed-Mode Surveys: An Overview of Design, Estimation and Adjustment Methods and Empirical Applications. In Johnson, T.P., Pennell, B.A., Stoop, I.A.L., & Dorer, B. (Eds.), *Advances in Comparative Survey Methods: Multinational, Multiregional and Multicultural Contexts (3MC)*, (pp. 409-430).

Vannieuwenhuyze, J. T., & Loosveldt, G. (2013). Evaluating relative mode effects in mixed-mode surveys: Three methods to disentangle selection and measurement effects. Sociological Methods & Research, 42(1), 82–104.

West, B. T., Ong, A. R., Conrad, F. G., Schober, M. F., Larsen, K. M., & Hupp, A. L. (2022). Interviewer effects in live video and prerecorded video interviewing. Journal of Survey Statistics and Methodology, 10(2), 317-336.

Zager Kocjan, G., Lavtar, D., & Sočan, G. (2023). The effects of survey mode on self-reported psychological functioning: Measurement invariance and latent mean comparison across face-to-face and web modes. Behavior research methods, 55(3), 1226-1243.

Appendices

Appendix A

Table A1: Measurement mode effects analysis with 103 survey items

Variable	Broad topic	Regression model	Difference, unweighted	Difference, weighted (PSW)	Potential mode effects (weighted data based on PSW)
Number of children reported	Household	Negative binomial	YES**	YES*	VI respondents more likely to have more children
Number of other household members reported	Household	Negative binomial	YES*	NO	
Whether in non-cohabiting relationship	Family	Binary logistic	NO	NO	
Helping parents	Family	Multinomial logistic	YES**	YES*	VI respondents more likely to help parents
Hours spent helping parents	Family	Multiple linear	YES**	NO	
Frequency of meeting family	Family	Ordinal logistic	NO	NO	
Frequency of other contact with family	Family	Ordinal logistic	YES**	YES*	VI respondents more frequently in contact with other family
Frequency of meeting friends	Family	Ordinal logistic	YES**	YES**	VI respondents less frequently meeting friends
Frequency of other contact with friends	Family	Ordinal logistic	NO	NO	
Number of rooms	Housing	Multiple linear	YES***	NO	
Value of property	Housing	Multiple linear	NO	NO	
Amount owed on mortgage	Housing	Multiple linear	NO	NO	
Problematic damp	Housing	Multinomial logistic	NO	NO	
Car ownership	Housing	Multiple linear	YES***	YES**	VI respondents more likely to own more cars
Economic activity status	Employment	Multinomial logistic	YES***	YES**	VI respondents more likely to be on a government scheme, etc.
Whether job SOC coded	Employment	Binary logistic	YES***	YES**	VI respondents more likely to have SOC code
Working hours - not including overtime	Employment	Multiple linear	NO	NO	
Working hours - with paid overtime	Employment	Negative binomial	YES**	NO	
Working hours - with unpaid overtime	Employment	Negative binomial	YES*	NO	
Place of work	Employment	Multinomial logistic	YES***	YES***	VI respondents more likely to work from home; less likely on the road
Physicality of work	Employment	Multinomial logistic	YES***	NO	
Job satisfaction	Employment	Ordinal logistic	NO	NO	
Work stress	Employment	Ordinal logistic	YES***	YES*	VI respondents reporting higher levels of work stress
Job security	Employment	Ordinal logistic	YES*	YES**	VI respondents reporting feeling less job secure
Self-employed earnings - take home	Employment	Multiple linear	NO	NO	

Self-employed earnings - gross pay	Employment	Multiple linear	YES**	NO	
Partner Economic Activity	Employment	Multinomial logistic	YES**	NO	
Whether partner job SOC coded	Employment	Binary logistic	YES***	NO	
Partner net pay	Employment	Multiple linear	NO	NO	
Expected retirement age	Employment	Multiple linear	YES*	NO	
Economic Shocks since coronavirus	Employment	Binary logistic	YES*	NO	
Number of benefits reported	Income	Poisson	NO	NO	
Whether has pension	Income	Binary logistic	YES***	YES***	VI respondents reporting more likely to have pension
Whether contributing to pension	Income	Binary logistic	YES*	NO	
Whether partner has pension	Income	Binary logistic	YES***	YES***	VI respondents reporting their partner more likely to have pension
Number of other sources of income reported	Income	Poisson	YES***	YES*	VI respondents more likely to report other sources of income (and larger number)
Total household income	Income	Ordinal logistic	YES***	NO	
Number of sources of savings reported	Income	Poisson	YES***	YES***	VI respondents more likely to report more sources of income
Total savings	Income	Negative binomial	YES**	NO	
Number of debts reported	Income	Negative binomial	YES**	YES**	VI respondents more likely to report fewer debts
Self-rated financial situation	Income	Ordinal logistic	YES***	YES***	VI respondents self-rated financial situation higher
Whether received an inheritance/substantial gift	Income	Binary logistic	YES***	YES*	VI respondents more likely to inherit
Total value of reported inheritances	Income	Negative binomial	NO	NO	
Whether has made a will	Income	Binary logistic	YES***	NO	
Cognition score - test 1 (Word recall: number)	Cognition	Multiple linear	YES***	YES**	VI respondents with a higher cognition score for this test
Cognition score - test 2 (Animal naming)	Cognition	Multiple linear	YES***	YES***	VI respondents with a higher cognition score for this test
Cognition score - test 3 (Letter cancellation)	Cognition	Multiple linear	YES***	YES***	VI respondents with a higher cognition score for this test
Cognition score - test 4 (Delayed recall words: number)	Cognition	Multiple linear	YES***	YES***	VI respondents with a higher cognition score for this test
Number of qualifications reported	Education	Negative binomial	NO	NO	
Self-reported health	Health	Ordinal logistic	YES***	YES**	VI respondents reporting better general health
How health has changed in last 12 months	Health	Ordinal logistic	NO	NO	
Long-standing illness	Health	Binary logistic	YES*	NO	
Number of health problems reported	Health	Negative binomial	YES***	YES**	VI respondents reporting fewer illnesses
Hospital visits - out-patient	Health	Binary logistic	NO	NO	
Hospital visit - in-patent	Health	Binary logistic	NO	NO	
Dental health	Health	Ordinal logistic	YES***	YES*	VI respondents reporting better dental health

Exercise - gets sweaty (days per week)	Health	Ordinal logistic	NO	NO	
Alcohol - normal beer	Health	Multiple linear	YES***	YES*	VI respondents reporting less drinking of beer
Alcohol - strong beer	Health	Multiple linear	NO	NO	
Alcohol - spirits	Health	Multiple linear	YES***	YES**	VI respondents reporting drinking less spirits
Smoking - whether smokes	Health	Multinomial logistic	YES***	YES**	VI respondents are less likely to smoke daily
Smoking - whether smokes	Health	Multiple linear	NO	NO	
Vaping - e-cigarette	Health	Multinomial logistic	YES***	YES*	VI respondents are less likely to use electronic cigarettes compared to those who never used them
Weight in Kg	Health	Multiple linear	YES*	NO	
Whether trying to change weight -	Health	Multinomial logistic	YES**	YES**	VI respondents are more likely to aim to lose weight than to stay about the same
Self-assessment of weight	Health	Multinomial logistic	YES**	NO	
Political interest	Political attitudes & behaviour	Ordinal logistic	YES***	NO	
Whether voted in 2019 GE	Political attitudes & behaviour	Binary logistic	YES***	YES**	VI respondents more likely to have voted in 2019 general election
Party voted for	Political attitudes & behaviour	Multinomial logistic	YES***	YES***	VI respondents were less likely to vote for Scottish National Party than Conservative Party
Who would vote for tomorrow	Political attitudes & behaviour	Multinomial logistic	YES***	YES***	VI respondents would be more likely to vote for Conservatives, Liberal Democrats and Green Party compared to Labour
EU Referendum vote	Political attitudes & behaviour	Binary logistic	YES***	YES***	VI respondents were more likely to vote in the EU referendum vote in 2016
EU Referendum vote	Political attitudes & behaviour	Binary logistic	YES*	NO	
Whether believes Britain will be better off outside EU	Political attitudes & behaviour	Binary logistic	YES**	YES***	VI respondents were more likely to believe that GB would be better off outside EU
Attitude to migration	Political attitudes & behaviour	Ordinal logistic	YES***	YES***	VI respondents were more likely to support immigration
Whether pays someone to do domestic tasks	Household	Binary logistic	NO	YES*	VI respondents were more likely to pay someone to do domestic tasks
Domestic chores - cooking	Household	Multinomial logistic	NO	NO	
Domestic chores - shopping	Household	Multinomial logistic	YES**	NO	
Domestic chores - cleaning	Household	Multinomial logistic	NO	YES*	VI respondents were more likely to share cleaning responsibilities than doing most of it
Domestic chores - washing	Household	Multinomial logistic	NO	NO	
Domestic chores - do it yourself	Household	Multinomial logistic	YES***	YES***	VI respondents were less likely to have someone else do it
Domestic chores - finances	Household	Multinomial logistic	YES*	NO	
Domestic chores - tending children	Household	Multinomial logistic	NO	NO	

Domestic chores - teaching children	Household	Multinomial logistic	NO	NO	
Domestic chores - rearing children	Household	Multinomial logistic	NO	NO	
Mental health	Mental health & wellbeing	Negative binomial	YES***	NO	
Mental health	Mental health & wellbeing	Negative binomial	NO	NO	
Mental health	Mental health & wellbeing	Negative binomial	YES*	NO	
Loneliness - feeling lack of companionship	Mental health & wellbeing	Ordinal logistic	NO	NO	
Loneliness - feeling left out	Mental health & wellbeing	Ordinal logistic	NO	NO	
Loneliness - isolated from other	Mental health & wellbeing	Ordinal logistic	NO	NO	
Relationship satisfaction	Mental health & wellbeing	Ordinal logistic	NO	NO	
Social Provisions	Other	Ordinal logistic	YES***	YES*	VI respondents more likely to have family and friends who help feel safe, secure and happy
Social Provisions	Other	Ordinal logistic	YES***	YES*	VI respondents more likely to have someone to turn to for advice
Social Provisions	Other	Ordinal logistic	YES***	YES*	VI respondents less likely to have no one close to
Interpersonal violence	Other	Binary logistic	YES*	NO	
Use of social media	Other	Ordinal logistic	NO	NO	
Self-efficacy - get what they want	Mental health & wellbeing	Binary logistic	YES***	NO	
Self-efficacy - determinism	Mental health & wellbeing	Binary logistic	YES***	NO	
Self-efficacy - outcome	Mental health & wellbeing	Binary logistic	YES***	NO	
Worry about retirement	Mental health & wellbeing	Ordinal logistic	YES***	YES***	VI respondents reported lower levels of worry about retirement
Cannot afford save for retirement	Mental health & wellbeing	Ordinal logistic	YES***	YES***	VI respondents reported lower levels of not being able to save for retirement
Trust	Mental health & wellbeing	Ordinal logistic	YES***	YES***	VI respondents reported lower levels of trust
Life satisfaction	Mental health & wellbeing	Multiple linear	NO	YES***	VI respondents reported lower levels of satisfaction

****p<0.001, ** p<0.01, *p<0.05*

www.surveyfutures.net