Understanding Society
Working Paper Series
No. 2024 – 04
March 2024

Assessing bias prevention and bias adjustment in a sub-annual online panel survey

Jamie C. Moore[1], Jonathan Burton[1], Thomas F. Crossley[2], Paul Fisher[1], Colin Gardiner[3], Annette Jäckle[1] and Michaela Benzeval[1]

[1] Institute of Social and Economic Research, University of Essex,

[2] Institute for Social Research, University of Michigan

[3] Ipsos MORI Social Research Institute

Please note: this working paper replaces working paper 2021-03: High frequency online data collection in an annual household panel study: some evidence on bias prevention and bias adjustment.

Non-technical summary

Understanding Society is the UK's largest household panel survey. From April 2020, participants in *Understanding Society* were additionally invited to partake in a series of web surveys (the *Understanding Society* COVID-19 Study) designed to capture higher frequency information during the COVID-19 pandemic. These data allow researchers and policy makers to explore how the pandemic impacted individuals and their families across the UK. A key feature of the *Understanding Society* COVID-19 Study is that it is designed to be "representative" of the UK population, in the sense of allowing analysts to make unbiased estimates of averages, prevalences and other statistics for that population.

In any survey, biases can arise if some segments of the population are not adequately covered by the survey design, or if particular segments of the population are less likely to respond to the survey. To prevent bias, the COVID-19 Study invited the full range of Understanding Society participants – both those who regularly use the internet and those who do not – to take part. Furthermore, periodically subsets of web non-respondents were invited to a telephone follow-up survey. Among web non-respondents, non-regular internet users were particularly targeted with this second mode. To further adjust for any bias arising from nonresponse, a weighting strategy was implemented taking advantage of the rich background information available from past annual interviews of the same individuals in the main *Understanding Society* survey.

In this paper we examine the efficacy of these bias reduction and bias adjustment measures, and along the way we develop some new statistical methods to aid in this assessment. We find that both the telephone follow-ups and weighting helped to reduce bias, but that inviting those who do not regularly use the internet to the web survey appears to have been of little benefit. We conclude by drawing broader lessons for future survey design.

Assessing bias prevention and bias adjustment in a sub-annual online panel survey

Jamie Moore[1], Jonathan Burton[1], Thomas F. Crossley[2#], Paul Fisher[1], Colin Gardiner[3], Annette Jäckle[1] and Michaela Benzeval[1]

[1] Institute of Social and Economic Research, University of Essex,
[2] Institute for Social Research, University of Michigan
[3] Ipsos MORI Social Research Institute

Corresponding Author: Thomas Crossley, Survey Research Centre, Institute for Social Research, 426 Thompson St., Ann Arbor, MI, 48106-1248, USA. Email: tfcross@umich.edu.
ORCID: 0000-0003-0952-7450

Abstract: To minimise non-response bias in survey-based estimates, both bias prevention measures during data collection and bias adjustment measures post-data collection are employed. We evaluate such measures using the UKHLS Covid-19 Study as a case study. The Covid-19 Study is a primarily web-based derivative of the annual UK Household Longitudinal Study. We find that telephone follow-ups of web non-respondents and non-response weighting helped to increase dataset quality, but inviting non-regular internet users to the web survey was of little benefit. We develop a statistical test of non-response weight performance. Inverse-probability non-response weights outperform simple calibration weights.

Keywords: mixed mode data collection, non-response population inferences, web survey, weighting.

Running title: Non-response bias: prevention and adjustment.

Acknowledgements and Funding: The COVID-19 Study was funded by the Economic and Social Research Council (ES/K005146/1) and the Health Foundation (2076161). Fieldwork for the survey was carried out by Ipsos MORI and Kantar Public (now Verian). Understanding Society is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex. The research data are distributed by the UK Data Service.

1. Introduction

Non-response is a pervasive issue in survey design (Groves et al. 2001). Not interviewing all units selected for interview (henceforth, the eligible set) reduces dataset size, causing survey estimate precision loss. If non-respondents differ from respondents, estimates may also suffer from non-response biases, causing invalid inferences. Response rates are declining, so the risk of these difficulties is increasing (Williams & Brick, 2018; Luiten et al. 2020). Hence, survey producers expend considerable resources to ameliorate such biases. Measures can be divided into those that seek to: 1) increase response rates and / or reduce differences in non-response between sub-groups, for instance by offering multiple contact modes and / or following up non-respondents ("bias prevention": see, for example, Groves et al. 2001; Groves & Heeringa 2006; Wagner 2008; Peytchev et al. 2010); and 2) reduce impacts post-data collection, for instance by computing weights to correct for (differential) non-response ("bias adjustment": see Valliant et al. 2013 for the use of weights, and Carpenter & Kenward 2012; Little & Rubin 2014 for other methods).

The issues caused by non-response and measures taken to ameliorate its impacts became even more critical during the COVID-19 pandemic. Producers had to meet demands from decision makers for in-depth, up to date information about its impacts on populations. However, the speed with which it affected society and the restrictions it placed on human interactions meant that new data collection methods had to be implemented at unprecedented pace. In response, many producers fielded surveys interviewing participants more often than before (see, for the UK: Brown et al. 2020; Institute for Social and Economic Research 2020; Germany: Blom et al. 2020; EU: European Parliament, Directorate-General for Communication, Public Opinion Monitoring Unit 2021a, b, c; Switzerland: FORS 2021). Moreover, face-to-face (F2F) interviews, a key component of most previous designs, were not possible. Hence, these surveys were conducted with limited information on how fieldwork and mode changes would affect levels and patterns of non-response, and dataset quality.

Understanding Society: The UK Household Longitudinal Study (UKHLS) is a large, annual, multi-domain panel survey of people living in the UK (see University of Essex & Institute for Social and Economic Research 2022 and Section 2). It is based on probability samples (Lynn 2009; Lynn et al. 2017) and supports high quality population inferences (Benzeval et al. 2020), and so is widely used by decision-makers and researchers. A mixed-mode design is employed, combining initial invitations to web and F2F

interviews with follow-up in alternative modes including telephone. However, given the need for more frequent data collection and restrictions in place, in April of 2020 the UKHLS team developed and fielded the UKHLS COVID-19 Study, a series of mainly web-based surveys designed to capture information from UKHLS main survey participants more often than normal (University of Essex, Institute of Social and Economic Research 2021). Nine surveys were conducted, the first in April 2020 and the last in September 2021. For full details of the COVID-19 Study, see the Study User Guide (Institute for Social and Economic Research 2021) and website: https://www.understandingsociety.ac.uk/topic/covid-19.

The COVID-19 Study deployed both bias prevention and bias adjustment strategies. The full range of UKHLS main survey participants (both regular and non-regular internet users) were invited to complete the web surveys. For the first wave, some non-regular internet user non-respondents were also followed-up by telephone. This telephone sampling could not be repeated at all waves due to its cost (a telephone response cost 21.7 times that of a web response) but was repeated for a second time in November 2020 at Study wave 6. Deriving the COVID-19 Study eligible set from the main survey offered methodological advantages. Beginning from a probability sample-based survey provided a strong starting point for population inferences, and main survey information could be used in bias minimisation efforts. Concerning the latter, inverse propensity non-response (IP-NR) weights that adjusted main survey design and non-response weights to account for COVID-19 Study non-response were developed for both web and, where applicable, web plus telephone survey datasets.

In this paper, we use the UKHLS COVID-19 Study as a case study to evaluate the effectiveness of bias prevention and bias adjustment measures in terms of dataset quality. To date, the COVID-19 Study has been the basis of more than 213 published journal articles studying the impact of the pandemic. However, the utility of the COVID-19 Study for making population inferences has not been evaluated. The quality of the COVID-19 Study dataset cannot be assumed on the basis of the main survey, as it different in a number of key aspects. Fieldwork was greatly compressed (one week instead of 22), and non-response was greater than in the main survey. Response modes were also reduced, which may have led to greater differential non-response across sub-groups (for COVID-19 related examples, see Rothbaum & Bee 2021; Schaurer and Weiß 2020). In addition, the performance of the Study IP-NR weights has not been assessed, or compared to that of other methods such as calibration weighting (e.g. Valliant et al. 2013).

When evaluating weight performance, both bias reduction and precision loss (the over-estimation of estimate variances caused by unequal weights: Little & Vartivarian 2005; Solon et al. 2015) should be considered. Moreover, there may be interactions between bias adjustment and prevention measures (Schouten et al. 2016). Understanding these issues is also of broader relevance. Many surveys are moving to web sampling, a shift that, due to lower costs and greater participant choice, was occurring pre-pandemic (e.g. Couper et al. 2007; Schonlau et al. 2009; Baker et al. 2010). Information on (comparative) calibration weight performance is of interest because a number of other surveys use non-probability samples combined with such weights (e.g. Adams-Prassl et al. 2020) or quota samples defined by similar benchmarks (e.g. Belot et al. 2021). The challenges faced by rapidly developed and deployed surveys during the pandemic in general, and the bias prevention and adjustment measures undertaken in the COVID-19 Study in particular, make the latter an interesting case study to examine.

We focus on the first wave of the COVID-19 Study. To evaluate bias prevention measures, we quantify the main-survey measured socio-demographic characteristics in three component datasets: regular internet user web respondents, non-regular internet user web respondents, and telephone respondents. Dividing respondents this way enables us to also assess the value of inviting non-regular internet users to a web survey, and the value of telephone follow-up.

To evaluate bias adjustment measures, we compare unweighted and weighted estimates with alternative non-response weights. We develop a test of the performance of non-response weights. This test complements existing tests for non-random attrition (Fitzgerald et al. 1998; Becketti et al. 1988), focusing instead on the ability of non-response weights to overcome such non-random attrition. To add to the IP-NR weights released for the web and web plus telephone survey datasets, we construct new IP-NR weights for the data set containing only regular internet users responding by web. We also construct calibration weights for each of the three datasets. These use eligible set population totals computed from main survey information. For each set of weights, we then consider: a) bias reduction, by using our new statistical test to compare COVID-19 Study weighted mean estimates of main survey wave 9 measured variables (socio-demographic characteristics, economic outcomes, and health outcomes) to main survey wave 9 weighted benchmarks; and b) precision loss, by computing the coefficient of variation (CVs) of weights, and Kish's

DEFF (Kish 1965). We also report on how weighted estimates of COVID-19 Study measured variables differ across datasets.

The paper proceeds as follows. Section 2 describes the UKHLS main survey and COVID-19 Study. Section 3 describes our evaluation methods, including weight construction and our proposed test of weight performance. Section 4 presents our results, and Section 5 discusses their implications, both for COVID-19 Study and for survey design more generally.

2. Data

2.1. UKHLS main survey

The UKHLS main survey is a multi-domain longitudinal study of people living in the UK (University of Essex & Institute for Social and Economic Research 2019). An attempt is made to interview all adult members of participating households (HHs) annually. The UKHLS began in 2009 but carries on from the earlier British Household Panel Survey, which began in 1991. It employs a sequential mixed-mode design, with some eligible set members initially allocated to a web interview and others to a face-to-face interview, with follow-up in the other mode. The UKHLS eligible set was constructed from probability samples, and non-response has been carefully modelled (Lynn & Kaminska 2010; Lynn et al. 2012). Research indicates that the survey continues to support valid population inferences (Benzeval et al. 2020).

This mixed mode infrastructure enabled the main survey to transition instantly to web and telephone interviewing in March 2020, and hence for fieldwork to continue despite the pandemic (see Burton et al. 2020). However, the UKHLS, like other large longitudinal surveys, is not set up to field rapidly changing content or to provide data to decision makers and researchers at pace. Hence, in March 2020 the decision was taken to set up a new COVID-19 Study, which would complement the annual interviews with higher frequency web-based data collection.

2.2. UKHLS COVID-19 Study

Beginning in April 2020, UKHLS main survey participants were invited to complete additional web surveys to track how the pandemic and associated policy responses were affecting them. Specifically, the COVID-

19 Study eligible set was defined as all UKHLS main survey participants aged sixteen or over, residing in HHs that responded at main survey waves 8 or 9, who had not subsequently withdrawn from the survey or died or emigrated as of April 2020. Pre-notification letters introducing the COVID-19 Study were sent on 17 April. Invitations to each web survey were sent by email and/or SMS text message, or by post, depending on contact details available, offering a £2 incentive for completion (at later waves, this could alternatively be donated to a NHS charity). The design and implementation of the Study built on research by the UKHLS team on event-trigged data collection (see Jäckle et al. 2019). Each web survey had a 7-day fieldwork period with reminders sent on days 2, 3, and 6.

Each web questionnaire was designed so that it took about 20 minutes to complete. The first web survey was fielded on 24 April. Eight further surveys were fielded at the end of May, June, July, September, and November 2020, and in March and September 2021. More information can be found in the Study User Guide (Institute for Social and Economic Research, 2020).

To capture eligible set members unlikely to respond via web, a telephone survey was also conducted in May 2020l. This ran concurrently with wave 2 of the Study but was based on the wave 1 questionnaire (adjusted for telephone, and with interviewer instructions added), leading to small between survey differences in time periods covered for non-baseline items. This telephone survey was issued to 3,411 non-regular internet users selected from non-respondents to the web survey. Internet use was quantified using a main survey wave 9 variable, which asked how often respondents use the internet: non-regular users are those reporting less than once a week, or for whom there was no information. The invitation letter for the telephone survey emphasized that recipients were an important part of society and that this was why they had been invited to a telephone survey and offered them a £2 gift card for completion. Fieldwork ran from 28 May to 7 June.

3. Methods

3.1. Evaluation of bias prevention measures

To evaluate how effective bias prevention measures were in eliminating differential non-response to the COVID-19 Study, we quantify the socio-demographic characteristics, as measured at wave 9 of the Main

survey, of COVID-19 Study wave 1 web respondents who were regular internet users, wave 1 web respondents who were non-regular internet users, and wave 1 telephone respondents. We consider a range of characteristics. For comparison, we also present similar information on the entire eligible set, and on all respondents combined.

3.2. Evaluation of bias adjustment measures

3.2.1. IP-NR weight construction

Cross-sectional survey inverse probability (IP-NR) weights were produced for each COVID-19 Study wave, for the web and (where relevant) web plus telephone datasets. These weights were derived by adjusting the main survey wave 9 cross-sectional weights (which in turn adjust the design weights given non-response: Lynn & Kaminska 2010). They scale respondents to the target population at wave 9 (2017/18), updated for subsequent mortality and emigration but not immigration.

More specifically, the COVID-19 Study weights were computed as the product of the inverse of the probability of response to the relevant wave, conditional on responding to main survey at wave 9, and the main survey wave 9 cross-sectional (selection) weight. Hence, they were only computable for respondents with main survey wave 9 weights. To estimate conditional response probabilities, regression methods were used with predictors drawn from the main survey so that they are observed for all eligible set members. The availability of information on non-respondents is one advantage of launching the COVID-19 Study from the pre-existing panel. The predictor set includes basic demographics, household composition, and economic and health variables. In addition, as suggested by several authors (e.g. Moffitt et al. 1999; Nicoletti & Peracchi 2005; Durrant et al. 2017), it includes information on previous survey outcomes and survey design variables. Such variables are rarely incorporated into substantive social or health science models but may be predictors of response and also correlated with outcomes of interest: for example, web response to previous surveys may predict current survey response and also employment.

When many predictors exist, an issue for regression modelling is model overfitting (Harrell 2001; Burnham & Anderson 2002). Use of over-fitted models to compute weights will cause precision loss (Little & Vartivarian 2005), so predictors were selected using logistic regression with a Least Absolute Shrinkage and Selection Operator (Lasso) procedure (in the package 'lassopack' for Stata 16: Ahrens et al. 2020).

Lasso procedures shrink unstable coefficient estimates towards zero, enabling model selection without formal statistical tests (Tibshirani 1996; Steyerberg et al. 2001). These procedures are further detailed in section 1 of the Appendix, including: how they were used to select models (minimizing the Extended Bayesian Information Criterion to choose tuning parameters λ), how they were used to estimate response probabilities (post-Lasso Probit estimation), and the reasons for methods used.

After computation the weights were trimmed. Trimming is applied to restrict precision loss, the over-inflation of estimate variances caused by unequal weights (Valliant et al. 2013). Extreme values are replaced with values that impact less on precision. In the COVID-19 Study, values above a threshold were replaced with the threshold value. However, trimming may also reduce the ability of non-response weights to reduce non-response bias. Lacking any published guidance on acceptable levels of precision loss, weights were created using different thresholds, then the set to release was selected by evaluating weight performance using the methods described in the following section.

For this paper IP-NR weights were also computed for the regular internet user web respondent dataset. In addition, for comparison we computed calibration weights scaling respondents to eligible set main survey wave 9 measured and weighted sex * age * education cell frequencies.

3.2.2. Non-response weight evaluation

The aim of weighting adjustments is to eliminate bias, but unequal weights are generally inefficient, so non-response bias reduction may come with a loss of precision (Little & Vartivarian 2005; Solon et al. 2015). Hence, any evaluation of weight performance must consider both bias reduction and precision loss.

To evaluate bias reduction, we propose a test that compares UKHLS main survey wave 9 measured, COVID-19 Study IP-NR weighted mean estimates of respondent characteristics to main survey wave 9 weighted benchmarks. To formalize this test, consider a "quasi-randomization" setup (Valliant and Dever, 2018). Let $I_i = 1$ indicate that individual i is selected into the eligible set for UKHLS, and $I_i = 0$ if not. Let $R_i^t = 1$ indicate that individual i responds to wave t (here, wave 9 of the main survey), and $R_i^t = 0$ if not, conditional on being in the eligible set. Denote the probability that individual i is in the eligible set by $\Pr(I_i = 1) = \pi^i$ and probability that individual i responds at wave t, given they are in the eligible set,

by $\Pr(R_i^t = 1 | I_i = 1) = \phi_i^t$. Let U be the set of individuals in the population and r^t be the set of respondents at wave t (that is, the set of individuals for whom $R_i^t I_i = 1$).

A1. Assume that $\pi_i > 0 \; \forall i, \;\; \phi_i^t > 0 \; \forall i,$ and weights for wave t, w_i^t, are available such that $w_i^t = (\pi_i \phi_i^t)^{-1}$.

For a quantity, y_i^t, observed in wave t, an estimator of the population total is:

$$\hat{T}(y^t) = \sum_{i \in r^t} w_i^t y_i^t = \sum_{i \in U} R_i^t I_i w_i^t y_i^t \qquad (4)$$

Again, in our application wave *t* is wave 9 of the main study, so this is just the weighted total using wave 9 respondents and the associated wave 9 weights. It is a standard result that $\hat{T}(y^t)$ is unbiased under **A1** (see, e.g., Valliant and Dever, 2018, Chapter 3). To see this take expectations over both the sampling and response processes:

$$E_I E_{R^t}\left[\sum_{i \in U} R_i^t I_i w_i^t y_i^t\right] = \sum_{i \in U} w_i^t y_i^t E_I E_{R^t}\left[R_i^t I_i\right] = \sum_{i \in U} y_i^t \qquad (5)$$

The last equality uses the fact that $E_I E_{R^t}\left[R_i^t I_i\right] = E_I\left[I_i\left[E_{R^t}\left[R_i^t | I_i\right]\right]\right] = \pi_i \phi_i^t$, and **A1**.

Now consider wave 1 of the COVID-19 Study, which we treat simply as a subsequent wave, $t + k$ of the panel. Let $R_i^{t+k} = 1$ indicate that individual i responds to panel wave $t + k$, and $R_i^{t+k} = 0$ if not, conditional on being in the eligible set *and* responding to wave t. We refer to this as *retention*. Let r^{t+k} be the set of respondents retained at wave $t + k$ (in our application, from wave 9 of the main survey, retained into Wave 1 of the COVID-19 Study).

The probability that individual i responds at wave $t + k$, given they are in the eligible set and responded at time t (that is they are retained), is $\Pr(R_i^{t+k} = 1 | I_i = 1, R_i^t = 1) = \theta_i^{t+k}$. Thus, the probability that they respond to wave $t + k$ is $\pi_i \phi_i^t \theta_i^{t+k}$.

A2. Assume that $\pi_i > 0 \; \forall i, \;\; \phi_i^t > 0 \; \forall i, \; \theta_i^{t+k} > 0 \; \forall i,$ and weights for wave $t + k$, w_i^{t+k}, are available, such that $w_i^{t+k} = (\pi_i \phi_i^t \theta_i^{t+k})^{-1}$

Consider an alternative estimator of population total of y^t, the quantity of interest *at wave* t:

$$\tilde{T}(y^t) = \sum_{i \in r^{t+k}} w_i^{t+k} y_i^t = \sum_{i \in U} R_i^{t+k} R_i^t I_i w_i^{t+k} y_i^t \qquad (5)$$

By similar arguments to those above, $\tilde{T}(y^t)$ is unbiased under **A2**. To see this take expectations of the sampling, response *and* retention processes.

$$E_I E_{R^t} E_{R^{T+k}}\left[\sum_{i\in U} R_i^{t+k} R_i^t I_i w_i^t y_i^t\right] = \sum_{i\in U} w_i^t y_i^t E_I E_{R^t} E_{R^{T+k}}\left[R_i^{t+k} R_i^t I_i\right] = \sum_{i\in U} y_i^t \quad (6)$$

The last equality uses the fact that $E_I E_{R^t} E_{R^{T+k}}\left[R_i^{t+k} R_i^t I_i\right] = E_I\left[I_i\left[E_{R^t}\left[R_i^t E[R_i^{t+k}|R_i^t, I_i]|I_i\right]\right]\right] = \pi_i \phi_i^t \theta_i^{t+k}$, and **A2**.

This result simply says that under **A2** we can alternatively estimate the population total of y^t using the subset of wave t respondents who are retained at wave $t+k$, and the appropriate wave $t+k$ weights. Note that $\hat{T}(y^t)$ is unaffected by the retention process, so that $E_I E_{R^t} E_{R^{T+k}}\left[\hat{T}(y^t)\right] = E_I E_{R^t}\left[\hat{T}(y^t)\right] = \sum_{i\in U} y_i^t$, and together these results imply:

$$E_I E_{R^t} E_{R^{T+k}}\left[\hat{T}(y^t) - \tilde{T}(y^t)\right] = 0 \quad (7)$$

This is the joint implication of **A1** and **A2** that we test.

Note that:

$$\tilde{T}(y^t) = \sum_{i\in r^{t+k}} w_i^{t+k} y_i^t = \sum_{i\in s^t} R_i^{t+k} w_i^{t+k} y_i^t. \quad (8)$$

This allows us to proceed as follows:

$$\hat{T}(y^t) - \tilde{T}(y^t) = \left(\sum_{i\in s^t} w_i^t y_i^t - \sum_{i\in s^t} R_i^{t+k} w_i^{t+k} y_i^t\right) = \sum_{i\in s^t} y_i^t\left(w_i^t - R_i^{t+k} w_i^{t+k}\right)$$

$$= \sum_{i\in s^t} y_i^t \omega_i \quad (9)$$

Where the composite weight ω_i is observed for all $i \in s^t$ because $R_i^{t+k} w_i^{t+k} = 0$ when $R_i^{t+k} = 0$. This means that we don't need to observe w_i^{t+k} for attritors (those not retained from wave t to $t+k$), although in practice we often do.

This formulation of $\hat{T}(y^t) - \tilde{T}(y^t)$ takes advantage of the fact that each retained individual (wave *t+k* respondent) is also a wave t respondent and so their weights can be "paired." Working with $\hat{T}(y^t) - \tilde{T}(y^t) = \sum_{i\in s^t} y_i^t \omega_i$ means that we need only make inferences about a weighted total, which we do using standard methods for inference with complex survey samples. We test the null that $\hat{T}(y^t) - \tilde{T}(y^t) = 0$. A rejection of the null would suggest either A1 or A2 (or both) do not hold. As the main survey weights have been extensively evaluated in previous work, a rejection of this null would lead us to doubt A2, that is, the adequacy of the COVID-19 Study weights.

In practice, we implement a version of this test based on weighted means. This test compares

$$\widehat{M}(y^t) = \sum_{i \in s^t} w_i^t y_i^t \Big/ \sum_{i \in s^t} w_i^t \quad , \tag{10}$$

And:

$$\widetilde{M}(y^t) = \sum_{i \in r^t} w_i^{t+k} y_i^t \Big/ \sum_{i \in r^t} w_i^{t+k} \quad . \tag{11}$$

These ratio estimators are not generally unbiased, but the bias is typically small (Cochran 1977) and given **A1** and **A2,** they are consistent estimators of the population mean. Thus, $\widehat{M}(y^t) - \widetilde{M}(y^t)$ will converge to zero with probability one as the sample size goes to infinity (formally this requires either that the size of the population goes to infinity, as in, for example, DuMouchel and Duncan, (1983); or that the sample size goes to infinity via sampling with replacement, as in, for example, Deaton (1997, Chapter 1)). The advantage of working in means is that the magnitude of departures from the null are often easier to interpret. For example, it may be easier to assess the importance of differences in weighted estimates of mean age or mean income across the two alternative estimates, than it is to assess differences in totals. A version of this test can also be derived in a model-based framework, see Crossley et al. (2021), who draw on results for model-based inverse probability-weighted estimators in Wooldridge (2002, 2007).

Here we use this test to assess whether alternative COVID-19 Study weights deal adequately with differential retention from the main survey to the COVID-19 wave of interest. More generally, this test could be used to assess whether IP-NR weights deal adequately with differential retention/attrition from one wave of longitudinal study to another (in cases, such as a cohort study, where the target eligible set does not change wave to wave). It complements existing tests for non-random attrition, such as the attrition Probit of Fitzgerald et al. (1998), and the pooling test of Becketti et al (1988). However, those tests focus on biases in the underlying retention/attrition process, rather than the adequacy of bias adjustments.

Turning to quantifying precision loss, we employ two measures commonly employed in survey design (e.g. Groves et al. 2001; Valliant et al. 2013). The first is the coefficient of variation (CV) of the weights i.e. $SD(weights)/mean(weights)$, where $SD(weights)$ is the standard deviation of the weights). With the CV, a larger value indicates greater variability in the weights, and potentially greater precision loss. The second is Kish's DEFF (Kish 1965), which is a measure of actual variance inflation in that it quantifies the extent to which the survey sampling error can be expected to depart from that expected under simple random sampling with a 100% response rate. The DEFF is a nonlinear function of the CV:

$$DEFF = 1 + (SD(weights)/mean(weights))^2 . \quad (4)$$

Again, a larger value implies a greater loss of precision.

When developing the IP-NR weights, these methods were used to evaluate performance before and after trimming, and to choose a level of trimming. Biases were quantified for a range of characteristics, both those included in response probability regressions and those not. This evaluation led to a choice to trim weights above 25 times the normalized median, which provides acceptable bias levels while keeping DEFFs below 3. In Section 3 of the Appendix we show how biases, weight CVs and DEFFs vary with trimming level for the wave 1 web survey dataset. In this paper we use the same methods to evaluate weight performance with the datasets under consideration. The first key question here is whether there is a benefit to bias prevention measures (a broader invitation strategy and / or mixed mode data collection) or to a sophisticated bias adjustment strategy. The second is whether there is an interaction between bias prevention and adjustment. For example, it could be that less weight trimming is required to control DEFFs when bias prevention strategies are employed.

In addition to the methods just described, we also examine a set of COVID-19 Study health and economic outcome variables. These variables are the survey targets, but non-respondent information is not available, so we cannot compute biases as we do for main survey measured subject characteristics. However, we can document the sensitivity of these weighted estimates to bias-prevention measures.

4. Results

4.1. Evaluation of bias prevention measures

Table 1 reports the prevalence of UKHLS main survey wave 9 information and response rates for wave 1 of the COVID-19 Study and the three components defined above. Column (i) shows that 43,981 main survey participants were eligible for the Study (the eligible set). This number is correct at the time of writing but may drop slightly if it is subsequently discovered that some subjects were actually ineligible (for example, had died by the time of the study): the released weights are updated as this kind of information arrives. Recall that the COVID-19 Study weighting strategy begins with the main study wave 9 weight. Thus COVID-19 Study weights are available for and much of our analysis is limited those with main survey wave 9 information. Of the Study eligible set, 35,352 had main survey wave 9 information.

The next three columns, (ii) through (iv), describe the three component datasets. Column (ii) focuses on 29,723 regular internet users in the eligible set (67.6% of the total). These are eligible set members that, at wave 9, reported using the internet 1-2 times per week or more. 15,514 responded to the first COVID-19 web survey, giving a response rate of 52%. Column (iii) focuses on the 14,258 non-regular internet users in the eligible set (32.4% of the total). This includes both those using the internet less than once per week (5,629) at wave 9 of the main study, and those for whom wave 9 information is missing (8,629). Since only the former (39.4 of the total) have wave 9 information, they represent just 15.9% of the eligible set with such information. The overall response rate for this group was 16%, but among those with wave 9 information it was just 13% (747/5629). This confirms the expectation that issuing non-regular internet users to the web survey is not very productive in terms of respondent numbers. Column (iv) focuses on the 3,398 non-regular internet web non-respondents issued to the telephone follow-up. Of these, 2,944 had wave 9 information. The overall response rate to the telephone survey was 21%, but among those with wave 9 information it was 23% (674/2,944).

Finally, Column (v) considers the combined dataset. That is, it considers the entire eligible set, whether regular internet user or not, and defines a respondent as one who responded to the web or telephone survey (recall that those issued to the telephone follow-up had previously been invited to the web survey). Eligible set size is the same as in Column (i), 43,981, of whom 35,352 have main survey wave 9 information. The number of respondents is 18,479 (which is the sum across Columns (ii), (iii) and (iv): 15,514+2,247+718). Of these, 16,935 had wave 9 information. The overall response rate is 42% and the response rate of those with wave 9 information is 48%.

Table 2 addresses how "selected" respondents from different components of the eligible set are, relative to the full eligible set. Main survey wave 9 measured subject characteristics are considered, requiring that we focus on eligible set members with such information. The left-most column, labeled (i), gives descriptive statistics for the eligible set members. This is the 'target' group, as the IP-NR weights released with the COVID-19 Study are derived given the wave 9 weight (see section 3.2.1). Put another way, if we were able to obtain a response from all these eligible set members, no adjustment to the wave 9 weights would be needed.

The narrowest data collection strategy we consider is issuing to regular internet users only (column (ii)). Relative to the eligible set, the resulting respondents are more likely to be aged below seventy, have higher education and have higher incomes, and less likely to be ethnic minorities or live in social housing. These imbalances would need to be addressed by bias adjustment. Comparing Column (iii), non-regular internet users, to Column (ii) reinforces this point. Relative to regular internet users, these respondents are older, less educated (60% have GCSE (General Certificate of Secondary Education) or lower, compared to 28% of regular internet users), and more likely to have a longstanding health issue. Thus, issuing to all eligible set members, not just regular internet users, leads to a more balanced dataset.

Column (iv) shows that telephone respondents are even older than non-regular internet user web respondents, even more likely to be less educated, and more likely to be ethnic minorities. They also have much lower household incomes and are more likely to be in social housing. Hence, adding these respondents to the dataset should further improve it in terms of replicating the eligible set. Column (v) then illustrates the combined effect of our bias prevention methods (a broader invitation and mixed-mode data collection). The resulting dataset is more similar to the eligible set in terms of observed variables than the component datasets. Nonetheless, this dataset is still better educated than the eligible set (with 48% having a degree, versus 39% in the eligible set, for example), less likely to be in social housing, or to be ethnic minorities. Hence, even this dataset is unlikely to enable satisfactory population inferences.

4.2. Evaluation of bias adjustment measures

To account for imbalances in COVID-19 Study datasets that remained despite bias prevention measures, bias-adjustment measures were taken. The main survey wave 9 weights were adjusted to account for non-response to the COVID-19 Study as well as to attrition from panel membership between wave 9 and the initiation of the COVID-19 Study. The following tables explore the application of such measures in our three datasets: regular internet user web respondents, regular and non-regular web respondents (the full web survey dataset), and the web plus telephone survey dataset. The idea is to compare the effect of bias adjustment in datasets arise from sampling strategies that differ in the degree to which bias prevention measures are applied (regular internet users only, broadening the invitation, broadening the mode to include telephone follow up). Thus, the tables show how bias prevention and adjustment interact.

The adjustments to the main survey weights were based on estimated probabilities of response to the COVID-19 Study (as described section 3.2.1). These probabilities were estimated by a Probit model including main survey wave 9 predictors that are selected by *a priori* logistic regression with a Lasso procedure. A large number of predictors are selected. In Table 3, we report the estimated (partial) effect on response probability for the 20 predictors with the largest impacts on web survey response (both regular and non-regular internet user). We present these as average marginal effects (computed using the Stata command 'mfx'). All are for dummy variable predictors, so this is the average change across the eligible set in the estimated response probability as the predictor changes from 0 to 1, when other subject predictor values are at their observed values. The two largest effect sizes involve eligible set members previous interactions with the main survey. Those who responded to wave 9 via the web were 23% more likely to respond to the first COVID-19 Study web survey. Those for whom the UKHLS team held an email address (and so were invited to the study by an email containing a web link) were 28% more likely to respond.

Table 4 examines bias reduction through weighting adjustment. We assess the ability of weighted datasets to recover main survey wave 9 population estimates for a range of variables, using the statistical test developed in section 3.2.2. That is, we use the main survey wave 9 reports for the relevant COVID-19 Study respondents, plus the weights computed for the Study wave, and compare resulting mean estimates to those computed using the eligible set and wave 9 weight. In addition to estimates using the IP-NR weights, we also report analogous estimates and tests computed using simple calibration weights designed to scale respondents to eligible set (main survey wave 9 measured and weighted) sex * age * education cell frequencies (see section 3.2.1).

The first columns of Table 4 report the eligible set estimates, and subsequent columns the biases (differences between main study benchmarks and COVID-19 Study estimates). Variables (predictors) included in the IP-NR weight response probability models are considered, as well as some that are not included (nor in calibration weight computation). All variables are binary, so that estimates are prevalences, and biases reported can be multiplied by 100 to give percentage point differences. The first result to note is that the calibration weights do not perform well. Even with the widest invitation strategy and multiple data collection modes (the web plus telephone surveys), there are statistically significant biases in most variables. For example, for the incidence of living in homes owned outright, the bias is 5 percentage points

on a baseline of 34%. Biases in the incidences of capital income, smoking, and social housing are similar in magnitude, and all are statistically significant at $p<0.05$. With the component datasets, the biases remaining after calibration weighting are a little worse.

The second key finding from Table 4 is that the IP-NR weights do perform well. Even with regular internet user web respondents only (the most limited sampling strategy), few prevalences exhibit a statistically significant bias, and biases are modest, at most 2 percentage points (for tenure: mortgage).

Table 4 also contains results on bias prevention. Including non-regular internet users in the web survey (the column headed "all web") is of little value in reducing bias. The bias for tenure: mortgage is still statistically significant. The bias for tenure: owned is no longer so, but in this instance the bias for smoking prevalence is significant. In contrast, including telephone respondents is of value: in the column headed "web and telephone" no biases are larger than 1 percentage point, and none are statistically significant (when IP-NR weights are employed).

Table 5 reports variability (CV) and variance inflation (DEFF) measures for the untrimmed and trimmed weights computed for the three datasets defined above (calibration weights are not trimmed as they do not take extreme values.) IP-NR weight variability measures for the full web survey dataset is not reduced compared to the regular internet user web respondent dataset, indicating that a broader web survey invitation strategy (inviting non-regular internet users) is unlikely to bring precision gains (for a fixed eligible set size). However, including telephone respondents substantially reduces weight variability and variance inflation, especially for the untrimmed weights. This implies that it should improve precision. It also suggests that the mechanism for the observed bias reductions in the dataset is that fewer weights are trimmed. This is an important finding about the interaction between bias prevention and bias adjustment. CVs and DEFFs for the calibration weights, which as noted previously perform poorly in terms of bias reduction, exhibit a broadly similar pattern across the three datasets, but are smaller than IP-NR weight values.

Finally, Table 6 shows how the estimated prevalences of health and economic conditions vary between sampling strategies / datasets. These estimates are the true survey targets, but we do not have benchmarks to compare them to. In this case, all bar one of the considered variables are binary, so estimates are prevalences (the exception is household income, which is continuous so that its estimates are means).

Reported prevalences/means are calculated using the IP-NR weights. Modest but potentially important differences exist between datasets. The dataset generated by the broadest sampling strategy (web with telephone follow-up) estimates a lower employment rate than the dataset generated by web sampling alone, which in turn estimates a lower rate than the dataset generated by web sampling only regular internet users. Similar results exist for mean household income and the prevalence of caring for others. Asthma incidence is approximately the same in the two web datasets, but slightly lower in the web plus telephone survey dataset. In contrast, the reverse occurs for the prevalence of being advised to shield by the NHS, and the prevalences of suffering from arthritis or cancer are approximately the same in the two web datasets, but higher in the web plus telephone dataset. The prevalence of being on benefits is similar in the three datasets. These differences across data sets reinforce that survey design decisions (about whom to invite and the number of modes) are important even when a sophisticated bias adjustment strategy is implemented.

5. Discussion

We evaluated the performance of the measures used to minimise non-response biases in the UKHLS COVID-19 Study, a mainly web-based survey of the UK population fielded during the pandemic. Bias prevention measures employed during data collection consisted of seeking to capture the UK population effectively by inviting the full range of UKHLS main survey participants, both regular and non-regular internet users, to complete the Study questionnaires, and telephone follow-up of some non-regular internet user non-respondents. Bias adjustment measures employed after data collection consisted of producing inverse probability non-response (IP-NR) weights to map Study respondents to the UK population. These make use of the fact that the main survey is based on a probability sample (they adjust main survey design and non-response weights,) and also the rich information the main survey provides on Study eligible set members (to predict underlying Study response propensities).

Our evaluations suggested these measures perform well. Concerning bias prevention, the socio-demographic characteristics of non-regular internet user web respondents and telephone respondents differed from those of regular internet user web respondents, so that combined dataset values were often closest to eligible set values. Hence, broadening the sample, to include non-regular internet users and

particularly telephone respondents was effective in reducing differences between COVID-19 Study and main survey respondents.

Concerning bias adjustment, with the IP-NR weights, biases were almost always small and smaller for the web plus telephone dataset than the other datasets. A similar pattern was true for precision loss measures (CVs of the weights and their DEFFs: see Kish 1965) computed to assess the extent to which the weights cause estimate variance inflation. This confirms an important interaction between bias prevention and bias reduction. Adding the telephone follow-up enabled bias-adjustment to eliminate biases with less loss of precision. Interestingly though, adding non-regular internet user web respondents to regular internet users did not increase weighted dataset quality: changes in biases were limited, for a precision loss that was higher than that for the other datasets. The IP-NR weights, which take advantage of a rich set of variables to model non-response, performed much better than simple calibration weights.

Our research has several implications. Regarding the UKHLS COVID-19 Study, it provides empirical evidence of the utility of the survey for making inferences about the UK population. This should be reassuring to decision makers and researchers who have employed the study to understand the impact of the pandemic and developed policies on the basis of such evidence. We note that a potential limitation in this context is that considered characteristics were measured at main survey wave 9, and not in the COVID-19 Study (our target). This is because information on non-respondent values for the latter variables do not exist, preventing us from quantifying biases for characteristics measured in the COVID-19 Study. However, we did find that estimated means and prevalences of such characteristics were altered by bias-prevention steps, even when using IP-NR weights. This underlines the importance of the bias prevention strategies.

Our research also has implications for survey design more broadly. Among both new and long-standing surveys, the COVID-19 pandemic accelerated a shift towards web modes, and changes in the frequency of data collection and the duration of fieldwork. Our research shows that with an appropriate bias minimisation strategy, it is possible to produce a high-quality (weighted) dataset following such mode changes and in such conditions.

That said, this study also demonstrates that ideally one would employ a multi-mode survey (with a telephone as well as a web component). Telephone follow-up captured population sub-groups that were

less-well captured by the web survey, leading to a higher quality final dataset. A further advantage of these interviews is likely to be the provision of sufficient respondents to support sub-group analysis, a dimension of dataset quality (Benzeval et al. 2020) not explicitly considered here. However, an issue is the high cost of telephone interviewing. In the COVID-19 Study, the average cost of a telephone response was 21.7 times that of a web response, so that the total cost of the 718 telephone responses was almost as large as that of the 17,761 web responses. There are fixed costs associated with both modes so that changes in scale or other parameters may affect this cost ratio. Nevertheless, we would expect it to remain substantial, and be similar for other surveys.

Second, on a related note, non-regular internet user web respondents did not increase weighted dataset quality. This suggests that it may be better to not sample such eligible set members by web, and instead use the resources to expand telephone sampling. These questions can be considered within the framework of adaptive survey design (see, for example, Groves & Heeringa 2006; Wagner 2008). Our findings suggest that web sampling may have continued past phase capacity, the point beyond which using the same collection methods continued to increase dataset quality. We note though, several caveats. To begin with, re-allocating resources may similarly lead to telephone sampling continuing past phase capacity, and hence to no increase in dataset quality. In addition, given the high cost of telephone responses, relatively few will be obtained for the resources saved by not issuing non-regular internet users to the web survey. 2247 such respondents were interviewed in wave 1, which given the reported cost ratio would equate to slightly more than 100 extra telephone responses. Due to these extra complexities (full consideration would also require evaluation of regular internet user web respondent plus telephone respondent dataset quality, estimation of sampling errors, simulation of extra telephone responses, plus, as the survey is longitudinal, quantification of impacts on datasets in following waves), we do not further investigate such questions here. They will, however, be a focus of future research.

Third, our research shows that calibration weights can perform poorly at reducing non-response biases, both in themselves and compared to IP-NR weights. Our findings suggest that without explicit evidence of acceptable performance, inferences about populations made from surveys employing such methods should be treated with caution (see also Couper 2007). The evidence presented here suggests that the most effective way to obtain high quality data quickly in times of national emergency is to launched

surveys from existing studies that are: a) based on probability samples; and b) provide background information on all eligible set members.

Finally, the test we propose to evaluate non-response weight performance could be used by other longitudinal studies to assess whether IP-NR weights deal adequately with differential retention/attrition from one wave to another. Another context where this test could be applied is the evaluation of inverse probability weights designed to account for the loss of eligible set members due to failure to consent or inability to match records when linking survey data with administrative records.

References

Adams-Prassl, A., Boneva, T., Golin, M. & Rauh, C. (2020) Inequality in the impact of the coronavirus shock: Evidence from real time surveys. *Journal of Public Economics*, 189: 104245. DOI: 10.1016/j.jpubeco.2020.104245.

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2020) lassopack: Model selection and prediction with regularized regression in Stata. *The Stata Journal*, 20*:* 176-235.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723.

Baker, R., Blumberg, S., Brick, M.J., Couper, M.P., Courtright, M., Dennis, M., Dillman, M., Frankel, M.R., Garland, P., Groves, R.M., Kennedy, C., Krosnick, J., Lee, S., Lavrakas, P.J., Link, M., Piekarski, L., Rao, K., Rivers, D., Thomas, R.K. Zahs, D. (2010) *AAPOR report on online panels*. American Association for Public Opinion Research.

Becketti, S., Gould, W., Lillard, L., & Welch, F. (1998). The Panel Study of Income Dynamics after fourteen years: An evaluation. *Journal of Labour Economics*, 6:472-492.

Belot, M., Choi, S.,Tripodi, E., van den BroekpAlternburg, E., Jamison, J.C., & Papageorge, N.W. (2021) Unequal Consequences of COVID-19: Representative Evidence from Six Countries. *Review of Economics of the Household,* 19 769-783.

Benzeval, M., Bollinger, C. R., Burton, J., Crossley, T.F. & Lynn, P. (2020) *The representativeness of Understanding Society*. Understanding Society Working Paper Series 2020–08, Institute for Social and Economic Research.

Blom, A.G., Cornesse, C., Friedel, S., Krieger, U., Fikel, M., Rettig, T., Wenz, A., Juhl, S., Lehrer, R., Möhring, K., Naumann, E., & Reifenscheid, M. (2020) High-frequency and high-quality survey data collection: The Mannheim Corona Study. *Survey Research Methods* 14: 171-178. DOI:10.18148/srm/2020.v14i2.7735

Brown, M., Goodman, A., Peters, A., Ploubidis, G.B., Sanchez, A., Silverwood, R. & Smith, K. (2020) *COVID-19 Survey in Five National Longitudinal Studies: Waves 1 and 2 User Guide (Version 2)*. London: UCL Centre for Longitudinal Studies and MRC Unit for Lifelong Health and Ageing.

Burnham, K. P. & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York, NY: Springer.

Burton, J., Lynn, P. & Benzeval, M. (2020) How Understanding Society: The UK Household Longitudinal Study adapted to the COVID-19 pandemic. *Survey Research Methods,* 14: 235–239. DOI: 10.18148/SRM/2020.V14I2.7746.

Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*. John Wiley & Sons.

Chen, J. & Chen, Z. (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95: 759–771. DOI: 10.1093/biomet/asn034

Couper, M. P., Kapteyn, A., Schonlau, M. & Winter, J. (2007) Noncoverage and nonresponse in an Internet survey. *Social Science Research*, 36: 131–148. DOI: 10.1016/j.ssresearch.2005.10.002.

Crossley, T. F., Fisher, P. & Low, H. (2021) The Heterogeneous and Regressive Consequences of COVID-19: Evidence from High Quality Panel Data. *Journal of Public Economics*, 193: 104344.

Deaton, A. (1997). *The analysis of household surveys: a microeconometric approach to development policy*. World Bank Publications.

Denver, J. & Valliant, R. (2018) *Survey Weights: A Step-by-Step Guide to Calculation*. Stata Press.

DuMouchel, W. H., & Duncan, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, *78*(383), 535-543.

Durrant, G.B, Maslovskaya, O. & Smith, P.W.F (2017) Using Prior Wave Information and Paradata: Can They Help to Predict Response Outcomes and Call Sequence Length in a Longitudinal Study? *Journal of Official Statistics*, 33: 801-833. DOI: 10.1515/jos-2017-0037

European Parliament, Directorate-General for Communication, *Public Opinion Monitoring Unit (2021a): European Parliament COVID-19 Survey – Round 1*. Kantar Belgium. GESIS Data Archive, Cologne. ZA7736 Data file Version 1.0.0, DOI: 10.4232/1.13708

European Parliament, Directorate-General for Communication, *Public Opinion Monitoring Unit (2021b): European Parliament COVID-19 Survey – Round 2*. Kantar Belgium. GESIS Data Archive, Cologne. ZA7737 Data file Version 1.0.0, DOI: 10.4232/1.13709

European Parliament, Directorate-General for Communication, *Public Opinion Monitoring Unit (2021c): European Parliament COVID-19 Survey – Round 3*. Kantar Belgium. GESIS Data Archive, Cologne. ZA7738 Data file Version 1.0.0, DOI: 10.4232/1.13710

Fitzgerald, J., Gottshalk, P. & Moffit, R. (1998) An analysis of sample attrition in panel data. *Journal of Human Resources*, 33(2), 251-299.

FORS (2020) MOSAiCH COVID-19 survey W1+W2+W3: technical information. Available from: https://forscenter.ch/wp-content/uploads/2021/07/mosaich-covid-19-w123-beta_technical-information_with-merging.pdf

Geisser, S. (1975) The predictive sample reuse method with applications. *Journal of the American Statistical Association* 70: 320–328. DOI: 10.2307/2285815.

Groves, R. M. (2006) Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70, 646-675.

Groves, R. M. & Heeringa, S. (2006) Responsive design for household surveys: tools for actively controlling survey errors and costs. *J. Roy. Stat. Soc. Ser. A.*, 169, 439-457.

Groves, R. M., Dillman, D. A., Eltinge, J. L., & Little, R. J. (eds.) (2001) *Survey Nonresponse*. Wiley Series in Survey Methodology.

Harrell, F.E. Jr. (2001) *Regression modelling strategies: with applications to linear models, logistic regression and survival analysis*. New York: Springer.

Institute for Social and Economic Research (2021) *Understanding Society COVID-19 User Guide Version 8.0*. University of Essex.

Institute for Social and Economic Research (2022) *Understanding Society: Waves 1-12, 2009-2021 and Harmonised BHPS: Waves 1-18, 1991-2009, User Guide, 14 November 2022*, Colchester: University of Essex.

Jäckle, A., Burton, J. & Couper, M. (2019) *Event-trigged Data Collection*. Understanding Society Working Paper Series 2019-16, Institute for Social and Economic Research.

Kish, L. (1965) *Survey Sampling*. Wiley: New York.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*, Wiley: New York.

Little, R. J. A. & Vartivarian, S. (2005) Does weighting for nonresponse increase the variance of survey means? *Survey Methods.* 31: 161-168.

Luiten, A., Hox, J. & de Leeuw, E. (2020) Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys. *Journal of Official Statistics*, 36: 469–487.

Lynn, P. (2009) *Sample Design for Understanding Society.* Understanding Society Working Paper Series, 2009–01, Institute for Social and Economic Research.

Lynn, P. & Kaminska, O. (2010) *Weighting Strategy for Understanding Society.* Understanding Society Working Paper Series 2010–05, Institute for Social and Economic Research.

Lynn, P., Burton, J., Kaminska, O., Knies, G., & Nandi, A. (2012) *An initial look at non-response and attrition in Understanding Society.* ISER, University of Essex.

Lynn, P., Nandi, A., Parutis, V. & Platt, L. (2017) Design and implementation of a high quality probability sample of immigrants and ethnic minorities: lessons learnt, *Understanding Society Working Paper 2017-11*, Colchester: University of Essex

Moffit, R., Fitzgerald, J. & Gottschalk, P. (1999) Sample Attrition in Panel Data: The Role of Selection on Observables. *Annales d'Économie et de Statistique*, 9: 129–152. DOI: 10.2307/20076194.

Nicoletti, C. & Peracchi, F. (2005) Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *J. Roy. Stat. Soc. Ser. A,* 168: 763–781. DOI: 10.1111/j.1467-985X.2005.00369.x.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. & Lindblad, M. (2010) Reduction of nonresponse bias in surveys through case prioritization. *Survey Research Methods,* 4: 21-29.

Rothbaum, J. & Bee, A. (2021) *Coronavirus infects surveys, too: Survey nonresponse bias and the Coronavirus pandemic.* US Census Bureau working paper number SEHSD WP2020-10

Schaurer, I. & Weiß, B. (2020) Investigating selection bias of online surveys on coronavirus-related behavioral outcomes. *Survey Research Methods*, 14: 103–108. DOI: 10.18148/srm/2020.v14i2.7751.

Schonlau, M., Soest, A. van, Kapteyn, A. & Couper, M. (2009) Selection Bias in Web Surveys and the Use of Propensity Scores: *Sociological Methods & Research*, 37: 291-318. DOI: 10.1177/0049124108327128.

Schouten, B., Cobben, F., Lundquist, P. & Wagner, J. (2016) Does more balanced survey response imply less non-response bias? *J. Roy. Stat. Soc. Ser. A,* 179: 727-748. DOI: 10.1111/rssa.12152

Schwarz, G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics* 6: 461–464.

Solon, G., Haider, S. J. & Wooldridge, J. M. (2015) What Are We Weighting For? *Journal of Human Resources*, 50, 301–316. DOI: 10.3368/jhr.50.2.301.

StataCorp (2017) Stata Lasso Reference Manual Release 17. StataCorp LLC, College Station, Texas.

Steyerberg, E.W., Eijkemans, M.J.C. & Habbema, J.D.F. (2001) Application of shrinkage techniques in Logistic regression analysis: a case study. *Statistica Neerlandica,* 55: 76–88. DOI: 10.1111/1467-9574.00157.

Sugiura, N. (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics—Theory and Methods* 7: 13–26. DOI: 10.1080/03610927808827599.

Tibshirani, R. (1996) Regression and shrinkage via the Lasso, *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

University of Essex, Institute for Social and Economic Research (2021). Understanding Society: COVID-19 Study, 2020-2021. [data collection]. 9th Edition. UK Data Service. SN: 8644, 10.5255/UKDA-SN-8644-9.

University of Essex, ISER., NatCen Social Research and Kantar Public (2019) Understanding Society: Waves 1-9, 2009-2018 and Harmonised BHPS: Waves 1-18, 19912009, 12th Edition. DOI: 10.5255/UKDA-SN-6614-13.

Valliant, R., Dever, J. A., & Kreuter, F. (2013) *Practical tools for designing and weighting survey samples*. New York: Springer.

Wagner, J. R. (2008) *Adaptive Survey Design to Reduce Nonresponse Bias.* PhD diss., University of Michigan, Michigan.

Williams, D. & Brick, M.J. (2018) Trends in US face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, 6: 186–211.

Wooldridge, J. M. (2002). Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Portuguese economic journal*, 1(2):117-139.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of econometrics*, 141(2):1281-1301.

Yang, Y. (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 92: 937–950. DOI: 10.1093/biomet/ 92.4.937.

Yang, Y. (2006) Comparing learning methods for classification. *Statistica Sinica* 16: 635–657.

Zhang, Y., Li, Y. & Tsai, C-L. (2010) Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* 105: 312–323. DOI: 10.1198/jasa.2009.tm08013.

Zou, H., and Hastie, T. (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67: 301–320. DOI: 10. 1111/j.1467-9868.2005.00503.x.

Table 1: COVID-19 Study wave 1 eligible set size, and UKHLS main survey wave 9 information availability and response rates for different components. 'All eligible' is all eligible set members. 'Regular internet users' are those reporting using the internet 1-2 times a week or more. 'Non-regular internet users' are those reporting less frequent internet use, or for whom information is unavailable. 'Issued to Tel. Survey' are the non-regular internet user web survey non-respondents who were issued to the telephone survey (hence, they are also counted in (iii)). 'All' is respondents from all three components (i.e. the web and telephone surveys) combined.

	COVID-19 Study Eligible	Respondents			
	(i)	(ii) Regular internet Users	(iii) Non-regular internet Users	(iv) Issued to Tel. Survey	(v) = (ii) + (iii) + (iv) All
N Eligible	43981	29723	14258	3398	
N Eligible with w9 info	35352	29723	5629	2944	
N Respondents		15514	2247	718	18479
Response rate		0.52	0.16	0.21	0.42
N Respondents with w9 info		15514	747	674	16935
Response rate, with w9 info		0.52	0.13	0.23	0.48

Table 2: Socio-demographic characteristics for the UKHLS COVID-19 Study. 'All eligible' (column (i)) is all UKHLS main survey participants who were eligible for the Study and had main survey wave 9 information. The other columns include respondents only. 'Regular net users' are web survey respondents who reported using the internet 1-2 times a week or more in main survey wave 9. 'Non-regular net users' are web survey respondents reporting less frequent internet use, or for whom information is not available. 'Issued to Tel. Survey' are the Covid-19 Study web non-respondents who subsequently responded to the telephone follow-up survey. 'All' is respondents from all three components (i.e. the web and telephone surveys) combined. 'GCSE' denotes General Certificate of Secondary Education.

	All Eligible	Respondents			
	(i)	(ii) Regular net Users	(iii) Non-regular net Users	(iv) Issued to Tel. Survey	(v) = (ii) + (iii) + (iv) All
Sex: Male	0.45	0.42	0.43	0.38	0.42
Age: 20-29	0.15	0.11	0.06	0.03	0.10
Age: 30-39	0.13	0.14	0.02	0.04	0.13
Age: 40-49	0.17	0.18	0.04	0.05	0.17
Age: 50-59	0.19	0.22	0.15	0.10	0.21
Age: 60-69	0.16	0.20	0.26	0.18	0.20
Age: 70-79	0.13	0.13	0.32	0.29	0.15
Age: 80-89	0.06	0.02	0.13	0.25	0.04
Age: 90+	0.01	0.00	0.02	0.07	0.00
Qualifications: Degree	0.39	0.51	0.22	0.18	0.48
Qualifications: A-level	0.22	0.21	0.18	0.14	0.21
Qualifications: GCSE or lower	0.39	0.28	0.60	0.67	0.31
Family type: Couple, kid(s)	0.25	0.27	0.10	0.05	0.25
Family type: Couple, no kid(s)	0.38	0.43	0.61	0.23	0.43
Family type: Single, kid(s)	0.04	0.03	0.01	0.03	0.03
Family type: Single, no kid(s)	0.34	0.27	0.28	0.69	0.28
Ethnic minorities	0.19	0.12	0.12	0.15	0.12
Country: England	0.79	0.82	0.77	0.75	0.81
Country: Wales	0.06	0.06	0.07	0.08	0.06
Country: Scotland	0.08	0.09	0.09	0.09	0.09
Country Northern Ireland	0.07	0.04	0.07	0.08	0.04
Tenure: Owned	0.35	0.37	0.64	0.50	0.39
Tenure: Mortgage	0.38	0.44	0.18	0.11	0.41
Tenure: Rented	0.11	0.10	0.05	0.10	0.10
Tenure: Social Housing	0.16	0.09	0.13	0.29	0.10
Household net income (£/month)	3633.65	3842.49	4038.23	1882.50	3773.62
Long-standing illness: Yes	0.35	0.33	0.49	0.55	0.34

Table 3: Top 20 Lasso response propensity model predictor average marginal effect sizes for the UKHLS COVID-19 Study web respondent only dataset (both regular internet users and non-regular internet users). The average marginal effect is the average change across eligible set members in the estimated response probability as the predictor changes from 0 to 1, when other predictor values for eligible set members are at their observed values. All predictors reported in the table are coded as indicator variables, with in each case the reference category being those sample members not identified in the predictor name: for example, for the first predictor listed, females. Predictors are also grouped by subject matter.

Predictor	Marginal Effect	p-value
Sex: Male	-0.07	<0.001
Ethnicity: Irish	-0.09	<0.001
Region: Northern Ireland	-0.07	<0.05
Age band: 16-29	-0.10	<0.001
Age band: 30-39	-0.06	<0.001
Age band: 80+	0.11	0.990
Qualifications: General Certificate of Secondary Education or lower	-0.07	<0.001
Occupation: Professional	0.13	0.968
Occupation: Administrative and secretarial	0.13	0.980
Occupation: Associate professional and technical	0.06	0.992
Standardised income decile: 6	0.06	<0.001
Standardised income decile: 5	0.06	<0.001
Reported income from savings and investment: Yes	0.08	<0.001
Household type: 3 or more adults, no kids, incl. at least one couple	-0.06	<0.001
Mode at main survey wave 9: Web	0.23	<0.001
Email known at start of Covid-19 Study	0.28	<0.001
Internet use: Every day	0.07	<0.001
Internet use: Never use / no access	-0.08	<0.001
Voted in last election: Inapplicable	0.08	<0.001
Voted in last election: Yes	0.06	<0.001

Table 4: Statistical tests of UKHLS COVID-19 Study weight performance. For each survey dataset, we test if differences between UKHLS main survey wave 9 weighted survey variable prevalence estimates ('Wave 9: wt. est.) and estimates computed given Covid-19 Study calibration and Inverse Propensity (IP-NR) weights (respectively, 'C. wt. diff.' and 'IP-NR wt. diff.') equal zero. We consider variables in the IP-NR weighting model and variables in neither weighting model. `Core benefits' include Income Support, Job Seekers Allowance and Universal Credit. * equals $P<0.05$, ** equals $P<0.01$, *** equals $P<0.001$.

	Wave 9	Reg. internet users		Web		Web and telephone	
Variable	wt. est.	C. wt. diff.	IP-NR wt. diff.	C. wt. diff.	IP-NR wt. diff.	C. wt. diff.	IP-NR wt. diff
In IP-NR model:							
Subjective financial situation (SFS): comfortable or OK	0.71 (0.00)	-0.05***	0.00	-0.04***	0.00	-0.04***	0.00
SFS: just about getting by	0.21 (0.00)	0.03***	0.00	0.03***	0.00	0.03***	-0.00
SFS: finding it quite/very difficult	0.08 (0.00)	0.01***	-0.00	0.01***	-0.01	0.01***	-0.00
Tenure: Owned	0.34 (0.00)	-0.06***	0.01*	-0.06***	0.01	-0.05***	-0.00
Tenure: Mortgage	0.34 (0.00)	-0.07***	-0.02***	-0.06***	-0.01*	-0.06***	0.01
Tenure: Rented	0.13 (0.00)	0.03***	-0.01	0.03***	0.00	0.03***	0.00
Tenure: Social Housing	0.19 (0.00)	0.09***	0.01	0.09***	0.00	0.08***	-0.01
Low skill Occupation	0.36 (0.00)	0.05***	0.00	0.05***	0.00	0.04***	-0.00
Any savings income	0.36 (0.00)	-0.08***	-0.00	-0.08***	-0.01	-0.07***	0.00
Behind with some or all bills	0.06 (0.00)	0.02***	0.00	0.02***	0.00	0.02***	0.00
Neither model:							
Income poverty	0.14 (0.00)	0.03***	0.01	0.03***	0.01	0.02***	-0.00
Receives core benefit	0.05 (0.00)	0.02***	-0.00	0.02***	-0.00	0.02***	-0.01
Visited GP in last year	0.78 (0.00)	-0.01	-0.00	-0.00	-0.00	-0.00	-0.00
Smoker	0.15 (0.00)	0.05***	0.01	0.05***	0.01*	0.05***	0.01
Hospital outpatient in last year	0.46 (0.00)	-0.00	-0.00	-0.00	-0.01	-0.00	-0.01

Table 5: Kish's DEFF estimating variance inflation due to weight use and weight Coefficients of Variation (CVs) for inverse probability weights (IP-NR) and calibration weights for each of the cumulative UKHLS COVID-19 Study (respondent) dataset components. "Raw" are the weights computed as the product of

the main survey weights and the inverse of the estimated response probability from the response probability model. "Trimmed" are these weights after trimming at 25 times the median response probability, which we undertake to reduce the variability of the weights and hence survey estimate precision loss. Calibration weights are not trimmed (see text), so for these only values for raw weights are presented.

		(i) Regular internet users		(ii) All web = (i) + non-regular internet users		(iii) All web + telephone = (ii) + telephone	
		Raw	Trimmed	Raw	Trimmed	Raw	Trimmed
IP-NR	DEFF	12.5	2.6	6.4	2.7	2.4	2.2
	CV	339.0	124.9	231.5	132.2	118.8	111.7
Calibration	DEFF	1.3		1.2		1.2	
	CV	52.2		46.5		39.9	

Table 6: UKHLS COVID-19 Study survey variable inverse propensity (IP-NR) weighted means / prevalences and their standard errors (brackets) for each cumulative dataset (regular internet users, web = regular internet users + non-regular internet users, web and telephone).

Variable	Regular internet users	Web	Web and telephone
Advised to shield by National Health Service	0.07	0.08	0.08
	(0.00)	(0.01)	(0.00)
Reported suffering from asthma	0.15	0.15	0.14
	(0.01)	(0.01)	(0.00)
Reported suffering from arthritis	0.12	0.12	0.14
	(0.00)	(0.00)	(0.00)
Reported suffering from cancer	0.04	0.04	0.05
	(0.00)	(0.00)	(0.00)
In paid work	0.63	0.62	0.58
	(0.01)	(0.01)	(0.01)
Household net earnings (£/month)	1935.50	1907.91	1790.51
	(29.66)	(28.91)	(25.87)
On benefits	0.14	0.14	0.14
	(0.01)	(0.01)	(0.01)
Carer in own or other household	0.47	0.46	0.44
	(0.01)	(0.01)	(0.01)

Online Appendix

1. Lasso variable selection methods: details and use

Lasso procedures (Tibshirani 1996; Steyerberg et al. 2001) are regularised regression methods. As with other regularised regression methods, they minimise the sum of squared deviations between predicted and observed values similar to Ordinary Least Squares (OLS), but in addition impose a regularisation penalty on model complexity (Ahrens et al. 2020). Due to the imposition of this penalty, such methods tend to outperform OLS in terms of out of sample prediction, as reducing model complexity and inducing shrinkage bias decreases prediction error. In doing so, they also address the problem of model overfitting: high in-sample fit, but poor prediction performance on unseen data.

Regularised regression methods incorporate tuning parameters that determine the amount and form of regularisation penalty. With Lasso procedures (Tibshirani 1996; Steyerberg et al. 2001), the mean squared error is minimised subject to a penalty on the absolute size of coefficient estimates:

$$\hat{\beta}_{lasso}(\lambda) = \arg\min \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x_i'\beta})^2 + \frac{\lambda}{n}\sum_{j=1}^{p}\psi_j\,|\beta_j|, \qquad (1)$$

where $\hat{\beta}_{lasso}(\lambda)$ are the Lasso estimated coefficients for each predictor in the considered set p given the tuning parameter $\boldsymbol{\lambda}$ that determines the overall penalty level, n is sample size, $\boldsymbol{y_i}$ is the value of the response variable for subject i = 1,..n, $\boldsymbol{x_i'}$ are the values of the predictors for the same subjects, $\boldsymbol{\beta}$ are the OLS estimated coefficients for the predictors, and $\boldsymbol{\psi_j}$are (given $\boldsymbol{\lambda}$) predictor-specific penalty loadings. A $\boldsymbol{\lambda}$ of zero results in the OLS model. Increasing $\boldsymbol{\lambda}$ ultimately results in an empty model, with all coefficients set to zero. It is this setting of some coefficients to zero and removal of predictors from models that enables Lasso to be used as a model selection technique. Note that in this paper we assume that predictors are uncorrelated and hence that Lasso-type penalisation is all that is necessary, enabling us (after standardising predictors so that they have equal variances) to set $\boldsymbol{\psi_j}$ all to unity: for methods suitable when this assumption is relaxed, see Zhou & Hastie (2005) & Ahrens et al. (2020).

Several techniques exist to choose the value of the tuning parameter $\boldsymbol{\lambda}$. The first of these is cross-validation, which explicitly evaluates out of sample prediction performance. The data in question are split into training and validation datasets. The models for different values of $\boldsymbol{\lambda}$ are then estimated and variables selected using the training dataset. Next, they are fitted to the validation dataset, and mean squared prediction errors calculated to quantify performance (Ahrens et al. 2020). For example, with the commonly used *K*-fold cross-validation technique datasets are split into *K* groups of approximately equal size (Geisser 1975). One group is treated as the validation dataset, and the others combined as the training dataset. Then, for each value of $\boldsymbol{\lambda}$, models are identified and their performance quantified multiple times in a process that involves each data point being used for validation once.

The second technique is the use of information criteria. Information criteria are closely related to regularised regression methods, being interpretable as likelihood methods that penalise the number of parameters in models. Again, models for different $\boldsymbol{\lambda}$ are estimated and variables selected, then the best performing is chosen based on information criteria value. The Akaike Information Criterion (Akaike 1974) or the Bayesian Information Criterion (Schwarz 1978) may be used, along with their extensions (for small n / high p relative to n) the corrected AIC (AIC_c: Sugiura 1978) and the Extended BIC (EBIC: Chen & Chen 2008).

When producing the inverse propensity non-response (IP-NR) weights released with the UKHLS COVID-19 Study and in the main text of this paper, we use information criteria techniques to choose

values of λ and identify models for estimating subject response probabilities. Specifically, we utilise the EBIC (in the Stata 16 package 'lassologit': see Ahrens et al. 2020), which is:

$$EBIC_{\xi}(\lambda) = n\,log\big(\hat{\sigma}^2(\lambda)\big) + df(\lambda)\,log(n) + 2\xi df(\lambda)\,log(p), \qquad (2)$$

where $\hat{\sigma}^2(\lambda) = n - 1\ \sum_{i=1}^{n} \varepsilon_i^2$ and ε_i are the residuals. df is the effective degrees of freedom, the penalisation parameter common to all information criteria, and in this case is quantified as the number of coefficients estimated to be non-zero. $\xi[0,1]$ is a second penalisation parameter included in the EBIC to prevent over-selection of variables when *p* is relatively large, and is quantified as:

$$\xi = 1 - log(n)/\big(2\,log(p)\big) \qquad (3)$$

We ultimately utilise EBIC techniques because in simulations Ahrens et al. (2020) show that in the majority of scenarios they perform best out of those mentioned earlier (all of which are available in the 'lassologit' package and its sister package 'lassopack': see Ahrens et al. 2020) in terms of model identification, that is, in terms of lowest rates of false positives (identifying predictors not correlated with the response variable) and false negatives (not identifying actual correlates of the response variable). We note though, that some of the findings replicate earlier work: see, for example, Chen & Chen (2008) for simulations showing that the EBIC performs better than the BIC. Moreover, there are theoretical reasons why such findings might be expected. First, supporting their use for model identification, BIC techniques are the only ones of those tested that are model consistent, that is, will select the 'true' model (if in the potential set) with a probability nearing one as sample size tends to infinity (Yang 2005; Zhang et al. 2010). Second, when model identification is the goal, theory indicates that cross-validation training datasets should be small and validation datasets should be close to *n*, because more data are required to identify the correct model than to reduce bias and variance (Yang 2006). This does not occur with the *K*-fold cross-validation technique included in 'lassopack' (and in most other Lasso software packages: see, for example, StataCorp 2017), with which the training dataset is ~ n/K (see earlier). We note here that given this, intuitively at least the relatively small size of most survey datasets may preclude the use of more appropriate cross-validation techniques for identifying response probability models anyway: a sufficiently large evaluation dataset may lead to too small a training dataset for initial model selection to reliably take place.

As mentioned in the second paragraph of this section, for the above techniques to be used as described predictors must first be standardised so that they have unit variance. Hence, when modelling COVID-19 Study response probabilities we first converted all multi-category predictors and interactions into dummy variables. We also set up models so that the predictors 'Gender', 'Age' and 'Education' and their interactions could not be removed during Lasso procedures, and included in the final selected predictor sets all dummy variables associated with Lasso-selected predictors: this approach reduced biases in weighted estimates compared to main survey wave 9 values (unpublished results). After model identification, we utilised post-Lasso OLS estimation to estimate subject response probabilities for weight calculation. This is because Lasso estimated coefficients are subject to attenuation bias (Ahrens et al. 2020). We fitted probit models including the Lasso-selected predictors, then computed estimated response probabilities using model coefficients and subject characteristics.

References

Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2020) lassopack: Model selection and prediction with regularized regression in Stata. *The Stata Journal*, 20*:* 176-235.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19: 716–723.

Chen, J. & Chen, Z. (2008) Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95: 759–771. DOI: 10.1093/biomet/asn034

Geisser, S. (1975) The predictive sample reuse method with applications. *Journal of the American Statistical Association* 70: 320–328. DOI: 10.2307/2285815.

StataCorp (2017) Stata Lasso Reference Manual Release 17. StataCorp LLC, College Station, Texas.

Steyerberg, E.W., Eijkemans, M.J.C. & Habbema, J.D.F. (2001) Application of shrinkage techniques in Logistic regression analysis: a case study. *Statistica Neerlandica,* 55: 76–88. DOI: 10.1111/1467-9574.00157.

Sugiura, N. (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics—Theory and Methods* 7: 13–26. DOI: 10.1080/03610927808827599.

Tibshirani, R. (1996) Regression and shrinkage via the Lasso, *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

Yang, Y. (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 92: 937–950. DOI: 10.1093/biomet/ 92.4.937.

Yang, Y. (2006) Comparing learning methods for classification. *Statistica Sinica* 16: 635–657.

Zhang, Y., Li, Y. & Tsai, C-L. (2010) Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* 105: 312–323. DOI: 10.1198/jasa.2009.tm08013.

Zou, H., and Hastie, T. (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B 67: 301–320. DOI: 10. 1111/j.1467-9868.2005.00503.x.

3. COVID-19 Study wave 1 web respondent weight DEFFs, CVs and weighted estimate biases at different weight trimming levels

Following creation of the IP-NR weights for the COVID-19 study and for this paper, we employed trimming techniques (replacing values beyond a threshold of *n* times the median with the threshold value) to reduce the impact of extreme weight values on weight variability and hence weighted survey estimate variance inflation (see also section 3.2.2 of the paper main text). Obviously, such trimming may also lead to an increase in non-response biases. In the absence of any published guidance on acceptable levels of weight variability / estimate variance inflation, we settled on a strategy of choosing a trimming level that gave a DEFF (a measure of variance inflation: see section 3.2.2) of <3 while not leading large non-response biases (measured as difference between COVID-19 Study weighted estimates of main survey wave 9 measured subject characteristics and similar estimates computed using main survey wave 9 weights). In practice, this meant using a trimming threshold of 25 times the median value. To show how weight variability / variance inflation and bias vary with trimming levels (and demonstrate the validity of our choice), in the two Tables below we report DEFFs, CVs (the Coefficient of Variation of the weights, a measure of their variability used in the paper main text) and main survey wave 9 variable non-response biases at the different trimming levels computed for wave 1 web respondents. We report biases for the variables considered in Table 5 in the paper, with variables in the IP-NR weight response probability model in the first table, and variables not in the model in the second table. Results for the chosen 'times 25' strategy are italicised.

Table 2.1: Statistical tests of UK Household Longitudinal Survey (UKHLS) COVID-19 Study wave 1 web survey weight performance at different trimming levels. We report results for the untrimmed weights ('Untr.') and for weights given each trimming level ('*8' the median value, '*10' the median value, and so on). In the top two rows, we report DEFFs and CVs for each set of weights. In the following rows, we test if differences between UKHLS main survey wave 9 weighted survey variable means / incidences and estimates computed given COVID-19 Study IP-NR weights equal zero (we also report the wave 9 main survey weighted estimate in the column 'W9'.). We consider variables included in the IP_NR weighting model. * equals $P<0.05$, ** equals $P<0.01$, *** equals $P<0.001$.

	W9	Untr.	*8	*10	*12.5	*15	*20	*25*	*30	*35	*40	*60	*65
DEFF		6.36	1.88	2.00	2.14	2.27	2.50	*2.75*	3.01	3.25	3.47	4.29	4.50
CV		231.54	93.79	100.12	106.75	112.55	122.37	*132.17*	141.90	150.12	157.04	181.48	186.98
SFS: comfortable or OK	0.71	0.02	-0.00	-0.00	0.00	0.00	0.00	*0.00*	0.01	0.01	0.01	0.01	0.01
SFS: just about getting by	0.21	0.01	0.01	0.01	0.00	0.00	0.00	*0.00*	0.00	0.00	0.00	0.00	0.01
SFS: finding it quite/very difficult	0.08	-0.02	-0.00	-0.00	-0.01	-0.01	-0.01	*-0.01*	-0.01	-0.01	-0.01	-0.01	-0.02*
Tenure: Owned	0.34	0.01	0.01	0.01	0.01	0.01	0.01	*0.01*	0.01	0.01	0.01	0.01	0.01
Tenure: Mortgage	0.34	0.00	-0.03***	-0.02***	-0.02***	-0.02***	-0.02**	*-0.01**	-0.01*	-0.01	-0.01	-0.00	-0.00
Tenure: Rented	0.13	0.00	-0.00	-0.00	-0.00	-0.00	-0.00	*0.00*	0.00	0.00	0.00	0.00	0.00
Tenure: Social Housing	0.19	-0.02	0.02***	0.02**	0.01*	0.01	0.01	*0.00*	-0.00	-0.00	-0.00	-0.01	-0.01
Low skill occupation	0.36	-0.01	0.01	0.01	0.01	0.00	0.00	*0.00*	-0.00	-0.00	-0.00	-0.00	-0.00
Any savings income	0.36	0.00	-0.01**	-0.01*	-0.01*	-0.01	-0.01	*-0.01*	-0.01	-0.00	-0.00	-0.00	-0.00
Behind with some or all bills	0.06	-0.01	0.00	0.00	0.00	0.00	0.00	*0.00*	0.00	-0.00	-0.00	-0.00	-0.00

Table 2.2: Statistical tests of UK Household Longitudinal Survey (UKHLS) COVID-19 Study wave 1 web survey weight performance at different trimming levels. We report results for the untrimmed weights ('Untr.') and for weights given each trimming level ('*8' the median value, '*10' the median value, and so on). We test if differences between UKHLS main survey wave 9 weighted survey variable means / incidences and estimates computed given COVID-19 Study IP-NR weights equal zero (we also report the wave 9 main survey weighted estimate in the column 'W9'.). We consider variables not included in the IP-NR weighting model. `Core benefits' include Income Support, Job Seekers Allowance and Universal Credit. * equals $P<0.05$, ** equals $P<0.01$, *** equals $P<0.001$.

	W9	Untr.	*8	*10	*12.5	*15	*20	*25*	*30	*35	*40	*60	*65
Income poverty	0.14	-0.01	0.02***	0.01**	0.01**	0.01*	0.01	*0.01*	0.01	0.01	0.01	0.00	0.00
Receives ore benefit	0.05	-0.01	0.00	0.00	0.00	-0.00	-0.00	*-0.00*	-0.00	-0.00	-0.00	-0.00	-0.00
Visited GP in last 12 months	0.78	-0.00	-0.00	-0.00	-0.00	-0.00	-0.00	*-0.00*	-0.00	-0.00	-0.00	-0.00	-0.00
Smoker	0.15	0.01	0.02***	0.02***	0.02***	0.02**	0.01**	*0.01**	0.01*	0.01	0.01	0.01	0.01
Hospital outpatient in last 12 months	0.46	-0.01	-0.00	-0.00	-0.00	-0.00	-0.00	*-0.01*	-0.01	-0.01	-0.01	-0.01	-0.01