Understanding Society: COVID-19 Study teaching dataset, 2020-2021

USER GUIDE

Version 1, October 2022

Economic and Social Research Council

01/10/22

Version No. 01.00

Table of contents

1. Introduction

How did the pandemic affect peoples' lives? As the UK went into the first lockdown of the COVID-19 pandemic, the team behind the biggest social survey in the UK, Understanding Society (UKHLS), developed a way to capture these experiences. From April 2020, participants from this Study were asked to take part in the Understanding Society COVID-19 survey, henceforth referred to as the COVID-19 survey or the COVID-19 study.

The COVID-19 survey regularly asked people about their situation and experiences. The resulting data gives unique insight into the impact of the pandemic on individuals, families, and communities. The COVID-19 Teaching Dataset contains data from the main COVID-19 survey in a simplified form. It covers topics such as:

- Socio-demographics
- Whether working at home and home-schooling
- COVID symptoms
- Health and well-being
- Social contact and neighbourhood cohesion
- Volunteering

The resource contains two data files.

1. **Cross-sectional:** contains data collected in Wave 4 in July 2020 (with some additional variables from other waves)
2. **Longitudinal:** contains mainly data from Waves 1, 4 and 9 with key variables measured at 3 time points.

This User Guide provides background information about Understanding Society and the COVID-19 survey and details the contents and structure of the COVID-19 Teaching Dataset including details of survey weights and missing data.

What are longitudinal surveys?

Longitudinal surveys or panel surveys are surveys where the same set of individuals or households (or other units of observation) are followed over time and generally asked the same questions to allow a dynamic picture of their lives (in contrast to cross-sectional surveys which provide a snapshot). In these surveys, each round of interview is referred to as a wave or a sweep. These waves or sweeps could be at regular intervals or at varying intervals.

1.1 Understanding Society

Understanding Society: the UK Household Longitudinal Study started in 2009 with a large representative sample of individuals living in households across the UK, with oversamples or boost samples of ethnic minority groups and immigrants. As a longitudinal study, Understanding Society seeks to follow households over a long period of time, regularly interviewing participants to collect information about different aspects of their lives. By interviewing the same individuals repeatedly (at approximately one-year intervals) and asking them the same questions, the study provides information to measure change over time and across the life course. For more details of Understanding Society, see the Understanding Society User Guide (Institute for Social and Economic Research 2020).

1.2 COVID-19 Study

From April 2020 to September 2021, participants from the main Understanding Society sample were asked to participate in a short web-survey. This survey covered topics designed to evaluate the changing impact of the pandemic on the welfare of UK individuals, families, and wider communities. The 20-minute questionnaire included core questions designed to track change, as well as varying content, which was adapted as the coronavirus situation developed.

How often were the interviews?

The first wave of the COVID-19 survey was fielded in April 2020, with monthly waves until July 2020 (Waves 1-4). From September 2020 onwards, the survey was fielded every two months until March 2021 (Waves 6-8). A final 9th wave was fielded in September 2021.

How were the data collected?

While the questionnaire was implemented as a web survey, in the 2nd and 6th waves (May and November 2020) there was an additional telephone survey for households without internet access to reduce coverage error. While the survey was mainly designed to interview adults (16+ year olds as of April 2020), in a few waves paper self-completion surveys were also sent to 10-15 year olds in the households of the adult respondents. Wave 8 (March 2021) also included COVID-19 antibody testing kits.

Who is in the sample?

All adults (16+ year olds as of April 2020) in households who had participated in at least one of the last two waves of the main study Understanding Society, were invited to participate in this survey. From the September 2020 (Wave 5) survey onwards, only sample members who had completed at least one partial interview in any of the first four web surveys were invited to participate. From the November 2020 (Wave 6) survey onwards, those who had only completed the initial survey in April 2020 and none since were no longer invited to participate.

Further information about the Covid-19 survey and the resources to support its use can be found in Section 5.1.

2. The COVID-19 Study Teaching Dataset

The COVID-19 Teaching Dataset includes data from the *Understanding Society* COVID-19 survey with some additional background information collected about the participants during their main *Understanding Society* annual interviews. The resource includes two files

- a cross-sectional file with data mainly from Wave 4 (July 2020), and
- a longitudinal file with data mainly from Wave 1, 4 and 9 (April & July 2020 and September 2021).

2.1 Topics

Both the files contain data on varied topics relating to socio-demographics and experiences during the pandemic. The topics are summarised in the table below along with information about the source of the data. A full list of variables is available in the appendix with information about the source and a link to the corresponding questionnaire.

Table 1 Topics in the Covid-19 Teaching dataset with information about the source of data

Topics	Cross-sectional	Longitudinal
Background variables (inc. health/shielding)	Main Survey (Wave 10)	Main survey (Wave 10)
Socio-demographic	Wave 4 (July 2020)	Wave 1 (April 2020) Wave 4 (July 2020) Wave 9 (Sept 2021)
Working at home	Wave 4 (July 2020)	Wave 1 (April 2020) Wave 4 (July 2020) Wave 9 (Sept 2021)
COVID symptoms and test	Wave 4 (July 2020)	Wave 1 (April 2020) Wave 4 (July 2020) Wave 9 (Sept 2021)
Health and well-being	Wave 4 (July 2020)	Wave 1 (April 2020) Wave 2 (May 2020) Wave 4 (July 2020) Wave 9 (Sept 2021)
Home schooling	Wave 5 (Sep 2020)	Wave 1 (April 2020) Wave 5 (Sep 2020)
Social contact	Wave 6 (Dec 2020)	Wave 3 (June 2020) Wave 6 (Dec 2020)
Neighbourhood cohesion	Wave 6 (Dec 2020)	Wave 3 (June 2020) Wave 6 (Dec 2020)
Volunteering	Wave 4 (July 2020)	Wave 4 (Jul 2020)

2.2 Structure

Cross-sectional file
The cross-sectional file (**c19_teaching_xw**) includes 53 variables, each measured at one time-point only. The data comes mainly from Wave 4 of the Covid-19 survey in July 2020. There are also background variables from the main Understanding Society study and variables on some key topics taken from neighbouring waves when not asked in Wave 4.

Longitudinal file
The longitudinal file (**c19teaching_lw**) includes 92 variables in total. Many variables are repeated measures with measurements from up to three time points. Data comes mainly from Waves 1, 4 and 9. However, due to the variability in when topics were included in the COVID-19 study, variables on some topics are taken from neighbouring waves. The variable list in the appendix lists the source of the variables as well a link to the relevant questionnaire.

This datafile is a **balanced panel**, which means it only includes the participants that took part in every wave.

The file is in **wide format**. In wide format, each participant has one row of information, and each measurement of the same variable is a different variable. In this dataset, this is operationalised by having the same root name with a wave prefix signifying the wave it was asked/measured in: **ca_**, **cd_**, **ci_**. These prefixes indicate the time point as follows:

- ca_ = variables from Wave 1 (or taken from Waves 2 and 3)
- cd_ = variables from Wave 4 (or taken from Waves 5, 6 and 7)
- ci_ = variables from Wave 9.

For example, the three variables indicating whether someone 'Has had symptoms that could be coronavirus' are called **ca_hadsymp**, **cd_hadsymp**, and **ci_hadsymp**.

Variables without prefixes are stable characteristics that do not change over time. These were asked in the main Understanding Society annual survey and copied across into this dataset. While some of these are characteristics that do not change over time, such as year of birth (**doby_dv**), others are information that was collected pre-Covid and taken to be fixed during the Covid survey, such as the highest educational qualification reported in Wave 10 of the main survey (**hiqual_dv**).

Identifiers: The variable pidp is a unique cross-wave identifier that can be used to identify individuals across the waves. The variable j_hidp is the household identifier from Wave 10 of the main Understanding Society survey and can be used to identify which individual respondents were living together (in the same household) at that time.

2.3 Missing data

In the web survey, respondents are initially only shown the substantive response options. If they click "Next" without selecting a response option, they are shown response options for "Don't know" and "Prefer not to say". In the telephone survey interviewers record if respondents spontaneously say "Don't know" or "Prefer not to say".

Missing observations are recorded using negative values such as

- -1 for Don't know,
- -2 for Refusal,
- -8 for Inapplicable
- -9 for Missing.

All of the questions are not relevant for everybody. For example, a question about pay is only relevant for those who have a job. So, those who say they don't have a job are not asked the question about pay. This means that the value of the pay variable would be missing for them. This type of missing is called inapplicable and assigned a value of -8. In rare cases, there may be coding error or some other type of random error resulting in the data team not being sure of the correct value of the variable. In those cases, a value of -9 is assigned.

For many analyses, users will need to set these values as missing.

Stata code for setting negative missing values to system missing

```
mvdecode _all, mv(-9/-1)
```

SPSS code for setting negative missing values to system missing

```
MIS VAL ALL (-9 THRU -1)
```

2.4 Weights

Both the cross-sectional and longitudinal files include survey weights that adjust the sample to represent the UK adult population.

What are weights and why do we need them?

Estimates based on sample survey data could be biased because of two main reasons.

Sample design

Some population sub-groups constitute a very small proportion of the population. If everyone is chosen with the same selection probability then the number of individuals from these groups will be too small to allow robust statistical analysis. So, to address this issue, individuals from these sub-groups are chosen with a higher selection probability than they appear in the population (such samples are called boost samples or over-samples). This however, means that estimate of any characteristic based on such samples will be biased in favour of the over-sampled group if the groups differ on that characteristic. For example, in the Understanding Society sample, some ethnic minority groups are over-sampled. If the self-reported health of these ethnic minority groups are on average lower than that of the white majority group, then any estimate of general health of the UK population based on this sample will be biased downwards.

Non-response

Not everyone selected into a sample agrees to participate in the survey. If those who agree to participate are on average different from those who don't, in terms of the characteristics being estimated, then the estimate based on the sample survey responses will be biased on favour of the respondents. For example, as young people are less likely to respond and young people are more likely to use computers than older age groups, any estimate of computer usage based on the sample will be biased downwards.

What do weights do?

Weights are designed to counter the impact of these biases. Design weights are the inverse of the selection probability and non-response weights are the inverse of the response propensity. So, those who are less likely to be included in the sample get a higher weight.

The cross-sectional file includes a single cross-sectional weight called **betaindin_xw**. This weight adjusts for unequal selection probabilities in the sample design and non-response and ensures that this subsample of individuals can be used to make inferences about the UK adult population. This weight is based on the **cd_betaindin_xw** weight and is sourced from Wave 4 of the Covid-19 survey.

The longitudinal file includes a longitudinal weight**: ci_betaindin_lw**. In addition to correcting for unequal selection probability, this weight adjusts for the fact that not all those invited to participate in the survey, do participate in all waves. Weighted estimates produced using these weights ensure that those estimates can be used to make inferences about the UK adult population (just as **betaindin_xw** does for the cross-sectional dataset).

For details of how these weights are calculated, see the main Understanding Society Covid-19 Study User Guide.

2.5 Clustering and stratification

The samples in Understanding Society are drawn using clustered and stratified designs. The names of the clustering and stratification variables are **psu** and **strata**. These two variables are included in both the cross-sectional and longitudinal datasets.

What is a clustered sample?

Sometimes it is not possible to select a sample directly from the population. For example, if you want to interview school children, you will first need to select a few schools and then get permission from the school authorities and then select a sample of children from the schools that have provided permission. If interviews are conducted in people's homes and

the sample is such that the chosen addresses scattered in very faraway places then it is very costly to send interviewers to all those different places. In such cases, it is cheaper to first select a few areas randomly and then select addresses from those areas. In Understanding Society, for such cost reasons, first postcode sectors are selected and then a few addresses are selected from the chosen postcode sectors. As the postcode sectors are first selected, they are referred to as Primary Sampling Units (PSUs).

What is a stratified sample?

If you want to make sure that the sample includes people of different types to allow inter-group comparisons, then one of the ways to do that is select a stratified sample. Here the entire population is divided into groups based on different characteristics (and these groups are referred to as strata) and then a few addresses are selected from every strata. In case of Understanding Society, the UK population was divided into groups based on region, economic level of the areas and population density or urbanicity.

Standard errors and survey design

Standard errors of estimates show the precision of these estimates, and so are as important as producing the estimates. When standard statistical softwares (Stata, SPSS, R and SAS) compute standard errors, they assume that the survey data is based on a simple random sample which is not clustered or stratified. But if in reality the sample is clustered or stratified then the computed standard errors will be biased. To correctly compute the standard errors, you will have to tell the software the names of the clustering and stratification variables and use special syntax.

3. Data access

The Understanding Society COVID-19 Study Teaching Dataset is available from the UK Data Service under Study Number 9019: https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=9019.

The dataset is classified as End User Licence (EUL) or safeguarded. Full details of the access requirements and the application process can be found on the UK Data Service website: https://ukdataservice.ac.uk/find-data/access-conditions/

4. Citation and acknowledgements

The bibliographic reference for this study is as follows:

University of Essex, Institute for Social and Economic Research, University of Manchester, Cathie Marsh Institute for Social Research (CMIST), UK Data Service. (2022). *Understanding Society: COVID-19 Study Teaching Dataset, 2020-2021* [data collection]. UK Data Service. SN: 9019, DOI: 10.5255/UKDA-SN-9019-1

All works which use or refer to these materials should acknowledge these sources by means of bibliographic citation. To ensure that such source attributions are captured for bibliographic indexes, citations must appear in footnotes or in the reference section of publications.

4.1 Citing this User Guide

When citing this User Guide, you can use the citation of this version quoted below. Note that where an online version is available on the Understanding Society website it is always the most up to date.

Institute for Social and Economic Research and Cathie Marsh Institute for Social Research. (2022), *Understanding Society: COVID-19 Study teaching dataset, 2020-2021, User Guide, Version 1.0, October 2022*, Colchester: University of Essex and University of Manchester: UK Data Service.

4.2 Acknowledgements

We would like to acknowledge Professor Sin Yi Cheung, Professor of Sociology at the University of Cardiff for proposing the creation of, and contributing to, the COVID-19 teaching dataset. The dataset is a collaboration between the UK Data Service and Understanding Society with acknowledgments going to the authors Dr. Jennifer Buckley, Cathie March Institute, University of Manchester, UK Data Service and Dr. Piotr Marzec and Dr. Alita Nandi, Understanding Society, Institute for Social and Economic Research, University of Essex.

5. User Support

5.1 COVID-19 Study documentation

The COVID-19 Topic page contains an overview of the Study with key links to documentation such as the user guide, questionnaires, briefing notes, dashboard tool and how to access the data from the UK Data Service:

The long-term content plan gives details on which topics were asked in which month.

The COVID-19 User Guide contains information on the content of each file in the dataset, sample and questionnaire design, data collection and data processing and information about using the dataset for research.

The COVID-19 dashboard tool can be used by students to build charts to show trends over time and compare different population groups.

The Briefing notes give key findings on subjects ranging from working at home to health and caring. The briefing notes include graphs, frequency charts and explanations of their content. They also include notes on different types of methodology and reasons for use.

5.2 COVID-19 Study

Students can go on to explore the whole COVID-19 Study. In Wave 8 the Study contained antibody testing, consent to NHS and Registry data linkage, a second respondent incentive experiment, and an experiment with the position and order of the consent questions. The serology file includes the results of these antibody tests. For information on the methodology and participant correspondences see the COVID-19 User Guide.

5.3 Training

The Help and Support section of the Understanding Society website provides links to the FAQs, online training courses and upcoming in-person training workshops. Training videos and webinars are available on our YouTube channel.

5.4 User Support Forum

Questions about the data can be posted on the Understanding Society User Support Forum. Questions asked by other data users are also visible and searchable.

Questions about accessing the data can be sent to the UK Data Service help desk help@ukdataservice.ac.uk.

Access the data

www.ukdataservice.ac.uk

help@ukdataservice.ac.uk

+44 (0) 1206 872143