

COVID-19 Schools Infection Survey, 2021 to 2022: methods and further information

Information on the methods used to collect the data, process it, and calculate the statistics produced from the COVID-19 Schools Infection Survey (SIS).

Contact:
Andrea Lacey
schools.infection.survey@ons.
gov.uk
+44 1633 651663

Release date:
28 February 2022

Next release:
To be announced

Table of contents

1. [COVID-19 Schools Infection Survey](#)
2. [The study sample](#)
3. [Data we collected](#)
4. [Timing of the study](#)
5. [Weighting](#)
6. [Seroconversion](#)
7. [Linking survey data and biological samples](#)
8. [Test sensitivity and specificity](#)
9. [Uncertainty in the data](#)

1 . COVID-19 Schools Infection Survey

The COVID-19 Schools Infection Survey (SIS) is jointly led by the Office for National Statistics (ONS), London School of Hygiene and Tropical Medicine (LSHTM) and UK Health Security Agency (UKHSA).

SIS aims to investigate the presence of antibodies to SARS coronavirus 2 (SARS-CoV-2) among pupils in sampled primary and secondary schools in England.

SIS was initially set up during the 2020 to 2021 academic year to detect both current infection and antibodies in pupils and staff. Its purpose was to better understand the transmission of coronavirus (COVID-19) in school settings. SIS that ran during the 2021 to 2022 academic year aims to monitor levels of antibodies in pupils only. We refer to the 2020 to 2021 academic year study as SIS1 and the 2021 to 2022 academic year study as SIS2.

Repeated surveys are being carried out together with antibody samples in a cohort of pupils.

This methodology guide provides information on the methods used to collect the data for, process, and calculate the statistics produced from SIS. We will continue to expand and develop methods as the study progresses.

This methodology guide can be read alongside:

- the [COVID-19 School Infection Survey](#), which gives headline statistics
- the [study guide](#), which explains to participants what taking part in the study entails

2 . The study sample

The COVID-19 Schools Infection Survey (SIS) has a stratified, two-stage sample, designed to support the survey produce estimates of antibody prevalence at national and regional level. The target precision for regional estimates was to have confidence intervals of plus or minus five percentage points (95% confidence level). When calculating required sample sizes, we assumed the most statistically conservative scenario with an estimated prevalence of antibodies of 50%, which would result in the widest confidence intervals.

When a school is sampled, all pupils within participating schools are invited to take part. We assumed participation rates of pupils taking part in the study would be roughly equivalent to the 2020 to 2021 academic year study (SIS1) (25% in primary schools and 15% in secondary schools). We also assumed clustering effects of whole school sampling could be accounted for using design effects estimated from data collected in SIS1 and the COVID-19 Infection Survey. We assumed a design effect (DEFF) of 2.3 in the variance calculation. There are also potentially clustering effects at the local authority (LA) level because schools are selected within a limited number of LAs in each region.

With these assumptions and using standard approaches for sample size calculations (normal approximation), we set out to achieve a sample of 884 participants in primary schools and 884 participants in secondary schools in each of the nine English regions. In total, 16,000 pupils were therefore to be recruited into the study to ensure that we could estimate antibody prevalence for primary and secondary schools separately. Using assumed response rates and average school sizes for both primary and secondary schools, this equates to 13 primary schools and seven secondary schools participating in each region (180 schools in total).

The first stage of sampling was the selection of local authorities. In the second stage, samples of schools (primary and secondary) were drawn from selected local authorities. Our sample frame is restricted to English state schools listed on the school census table for 2020 to 2021.

Schools excluded from the sampling frame were:

- special schools, independent schools, pupil referral units and further education colleges
- primary schools that do not have any pupils in school Years Reception to 11 (such as infant schools); however, these were included in the school counts of the first stage

Sampling of upper-tier local authority areas in England

We first explicitly stratified our sampling frame into the nine regions of England. In each region, we drew five LAs by systematic sampling with probability proportional to size after implicit stratification by the urban-rural split. Size was determined by the total number of schools (primary and secondary) in each LA.

We decided that local authorities selected in the stage one sample should be broadly representative of the urban-rural split in each region. As such, we implicitly stratified the sample of local authorities by using the percentage of primary schools located in urban areas within the LA as a measure. We did this using the [2011 urban-rural definitions](#).

We used primary schools because they are more likely to better represent a local demographic than secondary schools. This is because there are generally fewer secondary schools within a local area, and they are generally located in more central areas.

To select the local authorities within each region, we first listed all the local authorities ordered by the percentage of primary schools in urban areas. We then took a systematic sample on the local authority list, considering the size of each LA (number of schools). We therefore approximated probability proportional to size selection. We selected our five local authorities per region. This first-stage sample size was a balance between operational efficiency and having sufficient local authorities to gain enough information about the diversity of regional demographics.

LA substitution for practical purposes

Although we sought a random sample, we also wanted to retain a proportion of schools from the previous SIS1 by re-inviting willing schools to take part in this year's study. This would enable longitudinal analysis of repeat antibody samples collected from a cohort of children spanning both studies over an 18 month period.

Using existing relationships with schools also has operational benefits. To allow for this during the 2021 to 2022 academic year study (SIS2), we purposely substituted some of the randomly selected local authorities with local authorities that were already sampled in SIS1. For most regions, we achieved this by swapping the randomly selected local authority with the SIS1 local authority that had the most similar characteristics. We based that measure of similarity on the percentage of urban primary schools and total number of schools.

In addition, there were a small number of LAs that took part in SIS1 that had to be replaced in the LA selection process. This was mainly because of requests from these LAs to no longer involve their schools in the study. In these instances, we followed a similar substitution method to that mentioned above.

For the North West and North East regions (regions with high SIS1 participation), we aimed to retain a high number of schools. This was so a viable cohort could be established across both the SIS1 and SIS2 studies for longitudinal analysis. We therefore sampled five local authorities in these regions but only retained three. This was so we could re-recruit enough schools, ensuring the retained LAs were broadly representative. To do this, we replaced the three most similar sampled local authorities with our three SIS1 local authorities. We then flagged the two remaining randomly sampled local authorities as reserve LAs. If we did not achieve the required number of schools from our SIS1 LAs, we then had the option of sampling new SIS2 schools from the reserve LAs.

Sampling of schools

The aim of the study was to recruit enough schools to allow for region-level estimates as described previously. In the second stage of sampling, schools were selected from within the chosen local authorities. We used systematic sampling, with the list ordered by several indicators that implicitly stratified the sample. These were:

- percentage of pupils with free school meals (high was over 25%, medium was 12% to 25%, and low was under 12%)
- local authority
- whether the school has a sixth form (for secondary schools)

We did not order by urban-rural classification because this had been considered when sampling local authorities. However, it was considered when substituting in similar SIS1 schools. The sample frame for each region was taken and filtered to only include schools from the sampled local authorities, and sampling did not take place separately within each cluster (local authority), but across the whole region. This made the number of schools selected in each LA a random variable. We sampled 13 primary schools and seven secondary schools into the study. We used systematic random sampling with a random start point to identify the sampled schools.

Practical modifications

Several practical modifications were made to the sample selection procedure because of practical constraints on the selection of schools.

For regions where there were only a couple of SIS1 schools to be retained in SIS2, we "swapped in" the SIS1 schools for the SIS2 sampled schools that were the most similar in terms of characteristics. This was a process like the one outlined previously for swapping in SIS1 local authorities.

For regions where there were more SIS1 schools that could be retained than needed, we maintained a reserve list of SIS1 schools. If a SIS1 school declined to take part, we first replaced it with a similar SIS1 school from the reserve list. If a substitute was not available, we selected the originally sampled school from the SIS2 sampling frame that was originally replaced by the SIS1 school. If a SIS2 school declined to take part, we selected the next school in the interval.

There were some SIS1 schools who declined future involvement in the study. Rather than remove them from the sampling frame, we included them to allow for correct estimation of the sampling intervals. However, selection moving to the next school on the list if selected was not allowed.

For the North West and North East regions, where we already had full regional samples comprised of SIS1 schools, we ensured they were representative at a regional level with regards to the sampling characteristics above. We still drew our random sample for those regions (although spread across three local authorities rather than five) and replace all drawn schools with SIS1 schools.

Sampling of individual pupils

Within the selected schools, the headteacher was first invited to register the school. Once a headteacher had registered a school, all pupils were invited to participate in the study. This involved either the parents consenting on behalf of the child (for children in school Years Reception to 11) or the pupils consenting themselves.

3 . Data we collected

In each school that agreed to participate, head teachers were asked to register and complete a short questionnaire. Head teachers were also provided with information about the survey to forward to parents or guardians of pupils aged under 16 years, and pupils aged 16 years or over. After completing a short registration questionnaire and consent form, participants (or their parent or guardian if the pupil was under age 16 years) were enrolled into both antibody testing and questionnaire completion.

A study team visited each school to collect the biological samples for testing from the pupils who had enrolled in the study. Tests for pupils involved an oral fluid sample to test for SARS-CoV-2 (COVID-19) antibodies against the virus.

For each subsequent round of testing, participants receive advance notification of the date of the sample collection day. New participants were also allowed to join the study at any time.

Alongside each testing round, pupils, parents (of those aged under 16 years) and headteachers are asked to complete a questionnaire. The questionnaire collects demographic information and asks questions about policy-relevant topics, which vary between survey rounds.

4 . Timing of the study

Recruitment of schools to the study began on 4 October 2021. The first round of testing took place between 10 November and 10 December 2021. The second round of testing took place between 10 January and 3 February 2022.

Only those who had enrolled in the study and were in school on the day of testing were tested. This means those with coronavirus (COVID-19) symptoms and those self-isolating would not be present in the school building to be tested on the assigned test day.

The third round is scheduled to take place from 7 March to 25 March 2022.

5 . Weighting

Weighting is applied to the questionnaire and antibody data collected to make the data representative of the wider, target population with regards to specific characteristics. The weighting takes account of the design of the sample and reflects the response patterns and the total numbers of pupils in schools selected.

Response rates to the study may differ between various subgroups of the eligible population, and if this response propensity is correlated with the study outcomes, antibody positivity rates and other estimates will be biased if they are computed from the unweighted, observed data; for example if vaccine uptake (thus antibody positivity is higher in White compared with Ethnic Minorities, and there is higher participation rates in White group, unweighted data from the survey will overestimate the true population antibody positivity). It is important to note that weighting can only be carried out to adjust for differential response in variables that are known for both responders and non-responders. There may be other unobserved biases that affect an individual's likelihood of taking part, which cannot be controlled for by the weights calculated. In addition, some of the observed characteristics used for weighting may be based on small sample numbers, so care may need to be taken.

Separate sets of weights have been computed for each participant group (pupil and parent) for each questionnaire, and separate weights for antibody testing.

We computed initial design weights to reflect the survey sample design and be equal to the reciprocal of the selection probability. Schools that participated in the School Infection Survey (SIS) in the 2020 to 2021 academic year (SIS2) and were invited to take part this year, received different treatment regarding their assigned design weights.

To account for differences in response rates between schools, design weights were multiplied by an adjustment factor. This factor was given by the reciprocal of the school-level response rate, which equals the ratio of the number of participating pupils to the total count of pupils at the school. In this way, the weights of pupils from schools with low participation rates were increased.

To achieve regional level representation regarding selected control variables, we applied a final calibration step to the non-response adjusted design weights. This calibration ensures that the weights for specific subgroups of the sample sum to known population totals. Specifically, the weights were calibrated with respect to:

- school years (grouped as school Years 0 to 2, 3 to 4, and 5 to 6 for primary schools, and school Years 7 to 8, 9 to 11, and 12 to 13 for secondary schools)
- sex (two groups: male and female)
- ethnicity (two groups: non-minority white British and minority)
- free school meals (two groups: pupil has free school meals and pupil does not have free school meals).

Note that in round one, because of time constraints, non-response adjustments and calibration to free school meals were only applied to the antibody data and not the questionnaires.

Calibration group totals were obtained from the 2020 to 2021 school census tables. For the parent survey in secondary schools, the target population is pupils in school Years 7 to 11, and all calibration totals were computed with respect to this group.

The effect of weighting on the survey estimates is moderate. For example, for the 18 regional estimates of seroprevalence in primary and secondary schools in round one, the difference between unweighted and weighted results was:

- less than 1.0 percentage point (pp) in three cases
- between 1.0 and 3.5 pp in nine cases
- between 6.1 and 7.2 pp in three cases

Small differences between unweighted and weighted estimates may indicate that sampling rates within the various calibration groups are similar, but other reasons or chance effects may also be relevant.

6 . Seroconversion

Seroconversion is the development of antibodies to the virus as the result of either infection, vaccination, or both. During an infection, the virus enters the body, and the immune system begins to produce antibodies in response to the viral proteins (or "antigens") present. During vaccination, the components of the vaccine results in production of the spike antigen of SARS-CoV-2, to which the immune system produces an antibody response. Although the presence of immunoglobulins in oral fluids are in concentrations of at least 1 per 1,000th of that found in blood, the reactivity of salivary immunoglobulins mirrors that of serum. Therefore, oral fluids are a non-invasive alternative sample to blood, particularly in children.

Initial data shows that SARS-CoV-2 antibody levels decline over time; this will lead to a point where tests can no longer detect them. The length of time antibodies remain at detectable levels in the body is not fully known.

To account for the different follow-up times between the rounds, the seroconversion rate has been expressed per 1,000 person-weeks. This calculation takes the number of participants who seroconverted between two testing rounds and divides this by the sum of the weeks that each seroconverted participant had between the two rounds of testing.

The antibody tests used in this study are able to determine whether a participant has been infected with SARS-CoV-2 or has been vaccinated. One antibody test looks for antibodies against the viral nucleoprotein (NP) produced by the immune system upon infection. Another antibody test looks for antibodies against surface antigen (S1) produced by the immune system upon infection or vaccination.

7 . Linking survey data and biological samples

Information collected from each participant who agreed to take part is anonymised. An individual serial number or identifier (ParticipantID) is used. This allows for the differentiation of data collected between each pupil. Each ParticipantID is linked to their school by the school's unique reference number.

The biological samples are given a barcode and this barcode is also recorded against the ParticipantID by the study team. This allows the test results to be matched to the correct individual. Personal identifiers (for example, name) are not used to link the data.

8 . Test sensitivity and specificity

The coronavirus (SARS-CoV-2) antibody estimates provided in the COVID-19 School Infection Survey bulletin are the percentage of the school-based population testing positive for current SARS-CoV-2 antibodies on the day of testing. The proportion testing positive for SARS-CoV-2 antibodies should not be interpreted as being the prevalence rate. To calculate prevalence rates, we would need an accurate understanding of the oral fluid test's sensitivity (proportion of true-positive detected by the assay) and specificity (proportion of true-negative detected by the assay).

Results calculated as the proportion of individuals with a positive result can be adjusted to account for the tests specificity and sensitivity using:

$$p = \frac{(q + \textit{specificity})}{(\textit{sensitivity} + \textit{specificity} - 1)}$$

where p is the adjusted proportion positive and q is the observed proportion positive.

Test sensitivity

Test sensitivity measures how often the test correctly identifies those who have antibodies against the virus, so a test with high sensitivity will not have many false-negative results.

Our study involves participants self-collection under the supervision of a study worker. It is possible that some participants may take the sample incorrectly, which could lead to false-negative results. Note that we are measuring antibodies in oral fluids; a person may have antibodies in their blood, but the level in oral fluid is too low to be detected by the assay, leading to a false negative result.

Test specificity

Test specificity measures how often the test correctly identifies those who do not have the virus, so a test with high specificity will not have many false-positive results.

Oral fluid samples from students were collected and sent for detection of antibodies against the SARS-CoV-2 Nucleoprotein (NP) and surface antigen (S1). This was done using two separate Immunoglobulin G (IgG) capture-based enzyme immunoassays (EIA). The NP assay has been shown to have [80% sensitivity and 99% specificity](#), and the S1 assay is estimated to have 75% sensitivity and 98% specificity.

9 . Uncertainty in the data

The estimates presented in the Schools Infection Survey statistical bulletin are subject to [uncertainty](#). There are many causes of uncertainty, but the main sources of uncertainty in the analysis and data presented include each of the following.

Uncertainty in the test (false-positives, false-negatives)

These results derive directly from the tests, and no test is perfect: there will be false-positives and false-negatives from the tests (see section on sensitivity and specificity for details).

The data are based on a sample of people, so there is some uncertainty in the estimates

Any estimate based on a sample contains some uncertainty as to whether it reflects the broader population of interest because of its smaller sample size. A [confidence interval](#) gives an indication of the degree of uncertainty of an estimate, showing the precision of a sample estimate. The 95% confidence intervals are calculated so that if we repeated the study many times, 95% of the time the true proportion testing positive in the population would lie between the lower and upper confidence limits. A wider interval indicates more uncertainty in the estimate. Overlapping confidence intervals indicate that there may not be a true difference between two estimates.

Pupils who chose to enrol in the study may be different to those who do not enrol

As well as random sampling error, samples can be affected by non-response or self-selection bias. This can occur when there is a systematic difference between those who take part in the study and those who do not, meaning participants are not representative of the study population. If this difference is also associated with the likelihood of having SARS-CoV-2 antibodies, then the estimates produced from the data collected cannot be generalised to the study population as a whole.