Communicating uncertainty in the NSS publication

1. As a producer of official statistics, the Office for Students (OfS) is committed to effectively communicating our statistics so that users can have confidence in their use and interpretation of them. This means we aim to use meaningful and effective ways to report the potential statistical uncertainty within the NSS results that we publish.

2. This document explains what we mean by statistical uncertainty and provides more information about how we calculate uncertainty in the NSS results and present it in the NSS data dashboard. Some of this document – particularly the later sections – are intended for readers who are familiar with statistic concepts and notation.

3. In our approach to communicating uncertainty in the NSS publication, we have drawn on research and consultation carried out to inform the student outcome measures and TEF indicators, also published by the OfS.[1] The approach to uncertainty described below mirrors the approach taken in these other publications and described in their related documentation.

4. The uncertainty in the NSS statistics depends on two things: the number of responses that contribute to the statistic, and to a lesser extent, whether it is an extreme value or a middling value. Statistics based on smaller populations normally have greater uncertainty; the more responses we have to a question, the more confident we can be that our statistic is not affected by the sort of random variation that causes statistical uncertainty (as described in paragraph 6 below). We give a full account of how we calculate the uncertainty measures used in the NSS publication in paragraphs 24 to 28.

What is statistical uncertainty?

5. The statistics calculated from the NSS responses, such as the positivity measure, are factual representations of how students responded to the survey. Considered as such, it would be appropriate to rely solely on the values we calculate from the survey responses. For example (setting aside the possibility of processing errors), we can be sure in a particular case that, of the students who responded to the survey, 71 per cent gave a positive response.

6. However, this factual measure will be of less value to those who are interested in understanding the underlying positivity of students' academic experience, and the potential for its improvement. Thinking about statistics in these terms means that we instead want to think about them as representing the underlying academic experience in relation to a whole

[1] See, in particular, 'Description of student outcomes and experience measures used in OfS regulation', paragraphs 99 to 103, Annex C and Annex D available at Description and definition of student outcome and experience measures - Office for Students.

population of students who could have attended that provider and responded to the NSS, or may do so in the future. This whole population is known as a **superpopulation.**

7. It is not possible to say exactly what the academic experience looks like for the superpopulation, because students who could have attended the provider in question and responded to the survey but did not do so, and students who may attend the provider in future, cannot be known to us.

8. The group of students which did attend and respond to the NSS are just one set of students from this superpopulation, and the statistics calculated from data about this group are used to infer what we would expect in the superpopulation. However, this group is – in various respects – a random realisation of the whole population who could have done so. For example, perhaps one of the students who answered the survey was feeling particularly positive because it was their birthday. If it happened to be raining on the day that students chose to complete the survey, how differently would student experiences be reported compared with the responses that would have been made if it happened to be sunny instead?

9. This randomness could give rise to a slight difference in the observed NSS responses which could lead to slightly different positivity measures being calculated, even though the underlying academic experience remained the same. This potential for random variation in the values we calculate and interpret as the NSS results, is known as statistical uncertainty.

Why is statistical uncertainty important?

10. Statistical uncertainty is unavoidable in the calculation of any statistic that is unable to identify and refer to its superpopulation: it cannot be rectified through adjustments to the underlying data or the calculations we are performing. This means there will always be a question as to how exact any calculated NSS result is as an estimate for the superpopulation.

11. This question of exactness (or of statistical uncertainty) is important when NSS results are being interpreted to understand and make improvements to the academic experience. We need to understand the extent to which the NSS statistics are affected by uncertainty, because this should inform the way we use the statistics. For example, considerations of uncertainty may tell us that although two numbers in the NSS publication are different, it is very likely that this difference is due to random variation, rather than a true difference in the academic experience. When this is the case, we would be wise to avoid acting on this apparent difference.

Statistical uncertainty, not measurement error

12. Statistical uncertainty should not be confused with measurement error (sometimes known as observational error) or other ways in which survey statistics can become unreliable or inaccurate, such as non-response bias.

13. Measurement error occurs when there are inaccuracies either in the underlying data on which we are performing our calculations (for example, a student is erroneously reported as responding positively rather than negatively), or within the calculations that we are performing (for example, a formula that should include a ‘greater than or equals to’ condition mistakenly includes a ‘strictly greater than’ condition instead).

14. While neither example of measurement error can be entirely ruled out, we aim to identify and reduce the potential for measurement error and non-response bias. We are confident that the statistics we have calculated are an accurate factual representation of students' responses to the NSS. Our approach to reducing other forms of error is described in the NSS quality report.[2]

How we communicate statistical uncertainty in the NSS publication

15. In the NSS data dashboard, we show the value of the positivity measure and the difference from benchmark, and use 'shaded bars' to communicate the statistical uncertainty associated with each of those values.

16. These shaded bars aim to represent the continuous spread (or distribution) of statistical uncertainty around the different values that we have calculated. As such, they indicate the changing likelihood that the underlying academic experience is represented by different values, with the darkest shading representing the range in which there is the greatest likelihood that the true experience is represented. Much like the 'bell curve' of the normal distribution, as the shading lightens in both directions it represents a lower likelihood that the true experience is represented at that point. Wider shaded bars mean we need to consider the potential for the true experience to be represented by a wider range of values around the point estimate that has been observed.

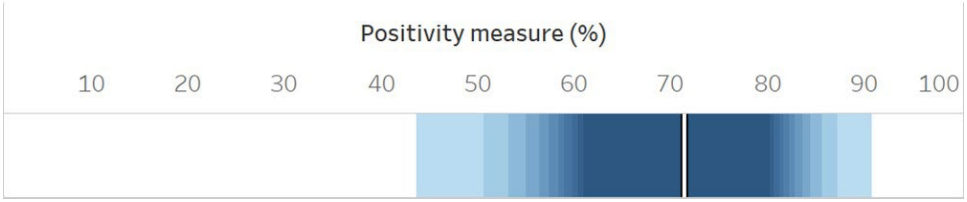
17. The shaded bars can, alternatively, be thought of as representing a series of discrete confidence intervals around the measures we have calculated, where each confidence interval in the series corresponds to a different confidence (or significance) level. The confidence level represents the likelihood that the confidence interval contains the true value in the superpopulation. In other words, on average, 95 per cent of confidence intervals computed at the 95 per cent confidence level would contain the true value in the superpopulation. Similarly, 90 per cent of confidence intervals computed at the 90 per cent confidence level would contain the true value, and likewise for other confidence levels.

18. We illustrate the distribution of statistical uncertainty up to a maximum of a 99.7 per cent confidence interval: the entire shaded bar therefore represents the 99.7 per cent confidence interval. This means it is extremely likely that the true value is covered by this bar.

19. Figure 1 below illustrates how we apply this approach to the positivity measure in the NSS publication. In the example shown in Figure 1, the positivity measure is 71.3 per cent. However, the true positivity measure for the superpopulation may be different from this. The darker shading indicates that the true positivity measure is most likely to fall between 61 per cent and 80 per cent. But we also acknowledge some likelihood that the true positivity measure is higher or lower than this, as indicated by the outer extremes of the shaded bar. If we wished to make a judgment with 99.7 confidence about the true positivity measure in this case, we could only say that it lies between 44 per cent and 90 per cent.

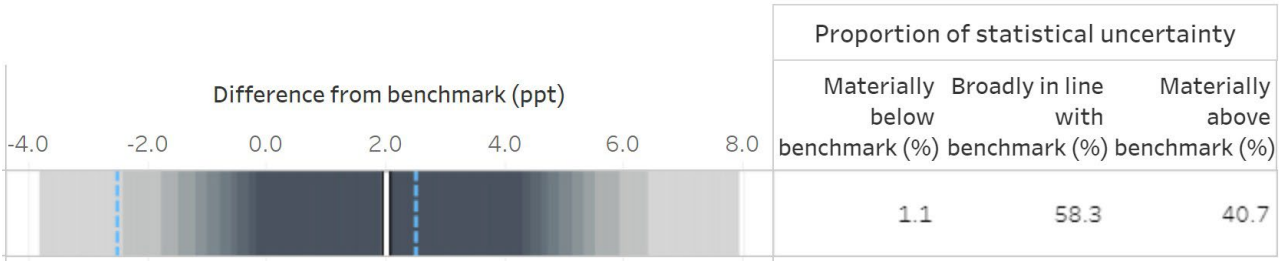
[2] Available at NSS data: quality report - Office for Students

Figure 1: Shaded bars around the positivity measure



20. We take a similar approach to reporting the statistical uncertainty in the differences from benchmark. This is shown in Figure 2. In this example, the actual difference between the positivity measure and the benchmark is 2.0 per cent. The shaded grey bar shows the uncertainty around this difference.

Figure 2: Shaded bars around the difference from benchmark

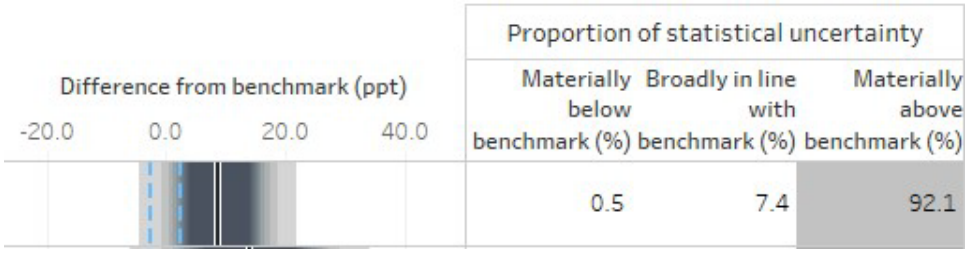


21. To support consistency and transparency of interpretation about the shaded bars around the difference from benchmark, and the statistical uncertainty they represent, Figure 2 shows that we also include summary figures in a table to the right of the shaded bars. These summary figures describe the proportion of the distribution of statistical uncertainty, represented by the shaded bar, that falls materially above or below the provider's benchmark value, and the proportion broadly in line with benchmark. The blue lines at 2.5 per cent and -2.5 per cent, as shown in Figure 2, indicate the point at which a difference from benchmark may be considered to be material.[3] In Figure 2, we can see that 1.1 per cent of the uncertainty distribution is materially below benchmark; 58.3 per cent is broadly in line with benchmark; and 40.7 per cent is materially above benchmark.

22. It is intended that the summary figures are used together with the shaded bars to aid interpretation of users' statistical confidence. The summary figures are highlighted where they show that at least 75 per cent of the distribution falls above or below those values, but users can use the shaded bars to make other interpretations of a provider's performance. As the proportion of the distribution within a category increase, the shading becomes darker. For example, in Figure 3 below, the final column is shaded, indicating that most of the uncertainty distribution is above benchmark. The shading is fairly dark, reflecting that proportion of the distribution that is materially above benchmark is approaching 100 per cent.

[3] The term 'materially' and the definitions of materially above and below benchmark for the purposes of interpreting NSS results are not intended to be statistical concepts and do not have particular statistical meanings. The guiding lines are intended to aid consistent interpretation of the difference from benchmark. We have used the same guiding lines as are used in TEF assessments, which improves consistency for participating providers.

Figure 3: shading of the 'Materially above benchmark' column



23. The shading of the summary figures can be used to easily identify cases in which the true positivity measure for the superpopulation is very likely to be higher, or lower, than the benchmark. We acknowledge, however, that for some measures the uncertainty distribution is so wide (due to small populations) that it is hard to achieve this threshold of 75 per cent, for any category. The shading of the columns should therefore not be taken as a sole measure of performance but should be used in conjunction with the other measures of uncertainty.

How we calculate the measures of statistical uncertainty

24. The confidence intervals used to create the shaded bars around the positivity measure use the Jeffreys interval.[4] We have used the Jeffreys interval because it has been shown to perform well in a wide range of circumstances, including when the denominator is small, or the positivity measure is close to 0 or 100.[5] The Jeffreys interval is calculated using the Jeffreys prior[6] for the binomial proportion, *p*, given *n* trials. Confidence intervals are calculated from the posterior distribution for p which is a Beta distribution with parameters (np + 0.5, $n - np$ + 0.5). In our case, *p* is the observed number of students giving a positive response to the question; and *n* is the number of students responding to the question. As the standard deviation of the binomial distribution decreases as the probability of success approaches 1 (i.e. an observed rate near 100 per cent), this results in a clear asymmetry in some of the bars.

25. The confidence intervals around the difference from benchmark depend on the standard deviation of the difference between the positivity measure and the benchmark, which incorporates the uncertainty in both components. The method for calculating this standard deviation is described by Draper and Gittoes (2004).[7] They describe the relationship between the indicator value and the benchmark and present evidence that the differences are normally distributed.

[4] Jeffreys, Harold (1946). An invariant form for the prior probability in estimation problems. Proc. Royal Society, London. A186453–461. https://royalsocietypublishing.org/doi/10.1098/rspa.1946.0056

[5] Brown et al (2001). Interval estimation for a binomial proportion Statistical Science. Vol. 16, No. 2, pages 101-133. http://dx.doi.org/10.1214/ss/1009213286.

[6] Although the Jeffreys interval has a Bayesian derivation it can also be justified from a frequentist perspective. See Brown et al (2001).

[7] Draper, D and Gittoes, M (2004). Statistical analysis of performance indicators in UK higher education. Journal of the Royal Statistical Society. Series A (Statistics in Society), 167, Part 3, pages 449-474.

26. Each of the shaded bars for the difference from benchmark represents a normal distribution with the distribution mean equal to the observed difference from benchmark and the distribution variance as the standard deviation squared. The distribution formula for the difference is

$$N(\text{Difference}, (\text{Standard deviation})^2)$$

27. Where the observed positivity measure is near 0 per cent or 100 per cent, it is possible for the distribution of the difference from benchmark represented by the shaded bar to imply that the measure value (i.e. if you centred this distribution around the observed indicator value) could extend below 0 per cent or above 100 per cent. In constructing the shaded bars for the difference from benchmark, we have explicitly not adjusted for this, except for cases where the provider's contribution to benchmark (or to the component parts of the benchmark) is 100 per cent because we cannot meaningfully calculate the standard deviation in such cases. Similarly, in rare cases where both the benchmark and the positivity measure are 100 per cent (or both 0 per cent), we are unable to calculate the uncertainty. We have instead tried to mitigate the issue by also presenting the shaded bar for the positivity measure. This is because the shaded bar for the positivity measure does not have this issue due to its derivation. The use of both charts reduces the risk that a user will misinterpret the uncertainty on the difference from benchmark in these cases.

28. The summary figures for the differences from benchmark represent the proportions of the uncertainty distribution which fall materially above and below a provider's benchmark. We regard a difference of at least 2.5 percentage points as material. To determine the proportions, we use the cumulative distribution function (CDF) for the normal distribution. To the left of the materiality boundary (-2.5 percentage points) the proportion is given by the CDF, while to the right of the boundary (2.5 percentage points) the proportion is given by one minus the CDF.

Multiple comparisons

29. In statistics, the issue of 'multiple comparisons' arises when a user considers multiple statistical tests at once. With more tests, there is more opportunity for unlikely events to occur simply due to the influence of random chance. To account for this, when conducting multiple tests, it may be appropriate to make formulaic adjustments to what we consider to be unlikely to have occurred by random chance alone – for example, by extending the confidence intervals around a measure.

30. In the NSS publication, we do not make any formulaic adjustments for multiple comparisons because we do not consider an arbitrary adjustment based on an assumed number of comparisons to be proportionate. In particular, we consider that the number of comparisons that users might make within and across the full set of available NSS data points could vary substantially depending on the use case and is difficult to predict: some users may choose to view a single positivity measure, others may view hundreds of them. Furthermore, while an adjustment based on an arbitrary number of comparisons may reduce the risk that some data users (those who view many statistics) make a false discovery due to statistical variation, it would simultaneously increase the risk that good statistical evidence is overlooked. Showing artificially wider distributions of the statistical uncertainty associated with each indicator would be a particular issue where users are considering an indicator in isolation or looking across a smaller number of indicators than are accounted for by the arbitrary adjustment.

31. The use of shaded bars around the indicators is intended to reduce the risk of false discoveries due to multiple comparisons, because it guards against an overreliance on one single confidence interval around the statistic. Instead, data users are encouraged to consider what level of confidence is appropriate for their purposes.

32. We acknowledge that there are some circumstances in which it may be desirable for users to consider making adjustments for multiple comparisons. We suggest that when lower levels of statistical confidence are being used to help identify outlying data points, or positivity measures that are above or below a benchmark, users should consider adjusting to a higher level of confidence when making their judgements. This is because of the higher risk of false discovery when using lower levels of statistical confidence. In this context, users may wish to be more conservative in their interpretation of statistical uncertainty the more comparisons they are making. Users can heavily mitigate the risk of making a false discovery by adjusting to use higher levels of statistical confidence. However, in doing so, they should note the consequence of an increased risk that sound statistical evidence may be overlooked.

33. We provide further examples of multiple comparison scenarios, and the way they should be approached, in guidance provided to accompany the OfS's student outcome and experience measures.[8] These examples are also relevant to the NSS publication.

How our approach has changed since the 2022 publication

34. Our approach to communicating uncertainty has changed substantially since 2022. The key changes are:

    a. In 2022, we showed a single confidence interval around the summary statistic, which was then the agreement rate. Our current approach uses a shaded bar to present multiple confidence intervals around the summary statistic, which is now the positivity measure. We regard this approach as an improvement as it recognises that there is no single answer to the question 'What is the likely range of the positivity measure for the superpopulation?'. Instead, we acknowledge that the uncertainty around the positivity is better viewed as a matter of degree: we can say with greater confidence that the true measure falls within a wider range, and with less confidence that it falls within a narrower range.

    b. In 2022, we presented a flag which showed whether the summary statistic for a population differed materially from the benchmark in either a positive or a negative direction. We took the threshold for materiality to be 3 standard deviations. A disadvantage of this approach was a 'cliff-edge' effect: very small changes in the benchmark or the positivity measure (for example, due to data amendments) could add or remove this flag, even though the change in the evidence was very slight. Our current approach instead presents a range of confidence intervals around the difference from benchmark, allowing data users to determine with varying degrees of confidence whether the positivity measure is different from benchmark. We have selected 2.5 percentage points as the threshold for material difference.

[8] See 'Description of student outcome and experience measures used in OfS regulation', Annex D, paragraphs 7-23, available at Description and definition of student outcome and experience measures - Office for Students.

c. In 2022, we made various adjustments for multiple comparisons. In effect, these widened the confidence intervals around the statistics in order to reduce the risk that a data viewer looking at many items encountered a random effect that was reported as significant. A disadvantage of this approach is that it made the same adjustment for everyone, regardless of how they used the NSS statistics. This means that in some cases there was a risk of obscuring real evidence. A second disadvantage is that the approach complicated the calculation of the confidence intervals, making them harder to reproduce. Our current approach does not make an adjustment for multiple comparisons and instead provides data users with the information they need to address risks presented by multiple comparisons. See paragraphs 29-33 for further discussion of the issue of multiple comparisons.

Any queries, please contact the NSS team, email NSS@officeforstudents.org.uk.