Methodology document for National Student Survey characteristic analysis

1. This document explains the methodology used to produce the National Student Survey (NSS) characteristic analysis published on 19 February 2020.

Background and scope

2. This data has been developed as an experimental statistic. It explores the extent to which responses to the National Student Survey (NSS) differ across student groups. This data was published alongside the Insight brief 'The National Student Survey: Consistency, controversy and change'.[1]

3. The data includes NSS results from the 2018 and 2019 surveys. Respondents are included from all domiciles and all levels of study. The analysis includes students studying in England, Northern Ireland, Scotland and Wales. For more information about which students are eligible to participate in the survey, see the NSS good practice guide.[2] The analysis is at sector level, so no provider level results are included.

4. The data is presented in three interactive charts. These display the percentage of respondents who agreed (selected 'Definitely agree' or 'Mostly agree') to each question and scale[3] in the NSS. Responses have been separated according to six student and course characteristics: age, sex, ethnicity, disability status, subject classification and mode of study. Alongside the actual percentage of respondents who agreed to each question, benchmark values have been calculated using these characteristics. We have shown when the agreement rate for a characteristic group differs from the benchmark.

5. There is always a degree of uncertainty around this sort of measurement. For example, the NSS survey is a measurement at a point in time – it is possible that some respondents would respond differently on a different day. To account for this, we have indicated when the difference from the benchmark is likely to be significant, rather than due to this sort of random variation.

6. A full description of the characteristics used in this analysis and their values is shown in Table 2. The data is shown separately for UK-domiciled and overseas students. Categories will not be shown where the response rate is lower than 50 per cent or there are fewer than 10

---

[1] See www.officeforstudents.org.uk/publications/the-national-student-survey-consistency-controversy-and-change

[2] Available at www.officeforstudents.org.uk/media/18b0f1e0-e085-405f-b5f6-fbb05c7b0167/nss-good-practice-guide-2020.pdf pp. 7-8.

[3] A scale is a group of questions addressing a common theme e.g. 'Assessment and feedback'.

responses. No categories failed to meet these publication criteria in the first publication of the data on 19 February 2020.

Benchmarking and significance testing

7. We have used benchmarking to take into account factors that are known to influence NSS agreement rates. The benchmark is a weighted sector average. It gives information about the agreement rates that might be expected given the student and course characteristics included in the weighting. Where differences exist between the rate and the benchmark, this suggests that the agreement rates are affected by a factor that is not included in the weighting.

8. The benchmarks are based on the six factors standardly used to analyse NSS data, with five of these used to calculate each individual benchmark. For example, if the characteristic analysed is age, then the benchmark is based on the sex, ethnicity, disability status, subject classification and mode of study of respondents.[4] If the agreement rate for an age group is different from the benchmark, we can infer that this difference is unlikely be due to any of the characteristics included in the benchmark: for instance, we can ignore the fact that older students are more likely to study part-time. The difference may instead be due to the age of the students, or to a further unknown factor.

9. A full explanation of the benchmarking methodology used here and in other higher education statistics can be found on the Higher Education Statistics Agency's (HESA's) website.[5]

10. As part of this analysis, we have evaluated the extent to which a user can be confident that the agreement rate differs from the benchmark. This is equivalent to evaluating the probability that the difference has arisen due to random variation in the responses: a user has 95 per cent confidence if the probability that the difference is due to chance is less than 0.05. We have presented this evaluation as follows:

   a. For the first two charts, 'Results by characteristic' and 'Differences by characteristic', statistical tests have been performed at three fixed confidence levels: 99 per cent, 95 per cent and 90 per cent.[6] This approach was chosen to reduce the 'cliff edge' effect produced by testing against a single, fixed confidence level, and to reflect the fact that different users may be willing to tolerate higher levels of uncertainty.

   b. The third chart, 'Explore the data' extends this idea, allowing users to control the confidence levels displayed. The lower confidence limit can be set in the range 85-90 per

---

[4] If the characteristic being analysed were to be included as a benchmarking factor, there would be no difference between the actual agreement rates and the benchmarks.

[5] See 'Technical notes': https://www.hesa.ac.uk/data-and-analysis/performance-indicators/benchmarks. Please note that the method described for calculating benchmarks and standard deviations has been used for this analysis, but not the method for producing significance markers; our approach to marking significance is described below. Further note that the technical notes refer to provider-level benchmarks, whereas benchmarks have been calculated at sector level only in this analysis.

[6] Because results are shown for each student and course characteristic, each question and scale in the NSS, and for UK-domiciled and overseas students separately, there are a large number of possible variations of each chart (420 in total). It is possible that not all confidence levels will be shown in every single one of these, so the total number of confidence levels shown in the key may vary as the filters on each chart are used. For example, if the '99% (above benchmark)' and '90% (above benchmark)' confidence levels are shown in the key, but the '95% (above benchmark)' level is not, this means that there are no categories displayed in the current version of the chart where the confidence level is greater than or equal to 95 per cent and less than 99 per cent (that the actual value is above the benchmark).

cent, the middle confidence limit can be set in the range 91-98 per cent, and the upper confidence limit remains fixed at 99 per cent.[7]

The Office for Students is currently exploring different ways of showing statistical significance and we welcome feedback on the approaches taken here.

11. It is expected that users of the data will wish look at multiple data items, and to consider together several or many comparisons with the benchmark. For example, when viewing responses split by subject classification, users may consider the agreement rates per subject as a whole, noting cases in which the rate is significantly better or worse than the benchmark. In this scenario, standard statistical tests can overstate confidence, as the probability of a false positive increases along with the number of data items under consideration[8]. To overcome this, an adjustment is often made to the calculation to control the false discovery rate (Benjamini and Yekutieli, 2001[9]): in this case, a Bonferroni correction has been used.

12. Considering the total number of questions, scales, categories of characteristics, and the number of years of data gives a total number of comparisons of 2,520. We have considered the number of comparisons within each broad domicile group (i.e. UK or overseas) as the total number of comparisons; this approach is cautious but stops short of assuming that users will consider all possible comparisons together. For simplicity we have assumed that all comparisons are independent and chosen to use the same correction for all characteristics. This is a conservative approach, which means that for most users the level of confidence will be understated. The OfS is reviewing its approach to dealing with multiple comparisons, and this approach may be refined in future.

13. Corrected significance levels calculated using the Bonferroni method[10] and their corresponding critical z-scores are shown in Table 3.Table 3

14. Colour coding has been used to indicate the confidence level met by each bar in the charts. Keys are provided to show the confidence level corresponding to each colour.

Quality profile

15. With any statistical output, there are factors that influence its reliability and accuracy. We have assessed some of the factors impacting this analysis, which are explained in this section. We welcome feedback on how the reliability and accuracy of this statistic can be improved.

16. This analysis is produced using NSS data. This data is linked to the Higher Education Statistics Agency (HESA) Student and Alternative Provider (AP) Student records, and the Individualised Learner Record (collected by the Education and Skills Funding Agency); all three are established data collections with their own quality assurance mechanisms. We therefore have a high level of confidence in the source data and the subsequent processing. Although errors

---

[7] When the values of the confidence limits are changed, the key will update automatically to show the new values, and the colours of the bars will update accordingly.

[8] For example, if 20 comparisons are made at a 95 per cent confidence level, the probability of at least one comparison resulting in a false positive is $1 - 0.95^{20} \approx 64.2$ per cent.

[9] Benjamini, Yoav; Yekutieli, Daniel. "The control of the false discovery rate in multiple testing under dependency." Ann. Statist. 29 (2001), no. 4, 1165--1188. doi:10.1214/aos/1013699998

[10] The formula for corrected confidence levels using this method is given by *1 – ((1 – C) / 2N))* where *C* is the confidence level expressed as a decimal (e.g. 0.95), and *N* is the total number of comparisons (i.e. 2,520).

sometimes occur in these data sources, we judge it unlikely that an undiscovered error in a provider's data would materially affect this analysis.

17. The benchmarking method used in this analysis is an established method used by HESA and the Office for Students in a range of contexts, including the calculation of provider benchmarks for the routine NSS data release. We have chosen to minimise risk by using the same method, and only varying the factors we use in the benchmarks.

18. Our approach to understanding and displaying significant differences is new for this analysis, and different from the approach used elsewhere in the NSS results. We have chosen not to include materiality as a factor when assessing significance. We have also applied significance tests at multiple levels; we believe this is appropriate in this context because it allows users to decide the sort of differences they are interested in, rather than assuming that a particular sort of difference is relevant and interesting to all users. We have also used a different level of Bonferroni correction to take account of the large number of data items that can be considered in one viewing. We have been particularly careful to ensure the accuracy of these aspects of the analysis, due to their relative novelty.

19. We have identified the following limitations with our approach:

- The benchmarks used have been identified in prior reviews as being likely to influence responses to the NSS. There may be other factors that influence responses, and which may better explain differences from the benchmark in some cases. For example, if students with certain characteristics are clustered at certain providers, and these providers offer a distinctive experience, then our analysis may show that this group of students have responses that differ from the sector as a whole. Yet, in this case, it would be a mistake to infer that the students' characteristics explain or cause their different responses, as this difference may be better explained by the provider they attend.

- The NSS has a high response rate. Nonetheless, the fact that some students do not respond introduces the possibility of non-response bias. This would occur if the non-respondents are different from the respondents in ways that systematically affect student or course characteristics.[11] In this analysis we have guarded against non-response bias by suppressing results for characteristic groups with response rates below 50 per cent (if any exist), and checking that response rates do not differ greatly across characteristic groups. The Office for Students is presently exploring other ways to identify and correct for non-response bias.

- The significance tests are dependent on the population size: when there are many students in a characteristic group, it is easier to be confident that a difference from the benchmark is significant. Correspondingly, we are less likely to find significant differences for characteristic groups that include fewer students. Users should be aware that a lack of a significance marker means that the data does not allow us to conclude with confidence that the difference is real. It does not imply that we can be confident that there is no difference.

---

[11] For example, if male students are more likely to respond than female students, the survey responses will be skewed towards males.

- This analysis looks at students grouped by a single characteristic: age, sex, subject, mode, ethnicity or disability status. We acknowledge that this ignores the interplay between these characteristics. For example, it is possible that the responses of part-time female students differ greatly from the responses of male part-time students; this is not explored here. The OfS is exploring ways of analysing and displaying multiple characteristics. More information is available in our report 'Associations between characteristics of students'.[12]

- Data about the characteristics of overseas students is not always available. This is particularly true of ethnicity, where the data is missing or unknown for around a quarter of respondents to the survey. This means that particular caution is needed in interpreting the results for overseas students, as differences may be hidden or created by the lack of data.

Raw data

20. The raw data that underpins the analysis is provided alongside the analysis itself. Both are available on the same webpage.[13]

21. The raw data is available in .xlsx format.

22. The fields included in the raw data are described in Table 1.

Table 1: Fields included in the raw data

Field	Description	Calculated as
Q	NSS question or scale.	-
Type	The student or course characteristic corresponding to the row e.g. Sex.	-
Description	The value of the student or course characteristic e.g. Male, Female.	-
Domicile	The value of domicile corresponding to the row i.e. UK or Overseas.	-
N_agree	The number of respondents who selected 'Definitely agree' or 'Mostly agree' to the question or scale.	-
N_response	The total number of respondents to the question or scale.	-
Actual_prop	The proportion of respondents who selected 'Definitely agree' or 'Mostly agree'.	N_agree divided by N_response
Benchmark	The benchmark value for the value of the characteristic corresponding to the row.	See footnote[14]
Stdev	The standard deviation of the *Actual_prop* field.	See footnote[15]
Difference	The difference between the percentage of respondents who agreed and the benchmark.	Actual_prop – Benchmark

---

[12] Available at www.officeforstudents.org.uk/publications/associations-between-characteristics-of-students
[13] See www.officeforstudents.org.uk/advice-and-guidance/student-information-and-data/national-student-survey-nss/sector-analysis
[14] See 'Technical notes': https://www.hesa.ac.uk/data-and-analysis/performance-indicators/benchmarks
[15] See 'Standard deviations' section: https://www.hesa.ac.uk/data-and-analysis/performance-indicators/benchmarks

Field	Description	Calculated as
Z	The number of standard deviations by which the percentage of respondents who agreed and the benchmark differ.	Difference divided by Stdev
Year	The year of the NSS survey.	-
Sig_level	The highest of three significance tests met by the difference between the agreement rate and the benchmark: 90 per cent, 95 per cent or 99 per cent.	-

Table 2: Definitions of student characteristics and characteristics of higher education provision

Characteristic	Categories[16]	Definition
Age	Under 21 21 – 24 25 and above	Age of student on the commencement date of their course
Sex	Male Female Other	Indicates the sex of the student
Ethnicity	Asian Black White Mixed Other	Indicates the ethnicity of the student
Disability status	Declared disability No known disability	Indicates whether the student has reported as disabled
Subject classification	Medicine and dentistry Subjects allied to medicine Biological and sport sciences Psychology Veterinary sciences Agriculture, food and related studies Physical sciences Mathematical sciences Engineering and technology Computing Architecture, building and planning Social sciences Law Business and management Language and area studies Historical, philosophical and religious studies Education and teaching	Subject studied according to the Common Aggregation Hierarchy (CAH1)[17]

---

[16] 'Unknown' categories are not shown in this analysis.
[17] https://www.hesa.ac.uk/innovation/hecos

Characteristic	Categories[16]	Definition
	Combined and general studies Media, journalism and communications Design, and creative and performing arts Geography, earth and environmental studies	
Mode of study	Full-time Part-time	Indicates the mode of study of the student

Table 3*: Confidence levels corrected for multiple comparisons and critical z-score values*

Confidence level	Corrected confidence level for individual comparisons	Critical value (two-tailed) corresponding to corrected confidence level[18]
0.85	0.999970238	4.014690665
0.86	0.999972222	4.030934373
0.87	0.999974206	4.048316748
0.88	0.999976190	4.067015643
0.89	0.999978175	4.087254637
0.90	0.999980159	4.109320282
0.91	0.999982143	4.133588387
0.92	0.999984127	4.160565668
0.93	0.999986111	4.190959065
0.94	0.999988095	4.225798252
0.95	0.999990079	4.266669568
0.96	0.999992063	4.316211141
0.97	0.999994048	4.379324740
0.98	0.999996032	4.466889547
0.99	0.999998016	4.613038168

---

[18] Critical values were calculated using the inverse of the standard normal cumulative distribution function.