National Student Survey 2022 pilot

Assessing new questions and response options

Reference OfS 2022.42

Enquiries to Nick Bailey: nss@officeforstudents.org.uk

Publication date 28 July 2022

Contents

Background

1. In 2021 we began a review of the National Student Survey (NSS) questionnaire. The overall aim of the review was to ensure that the questions and response options remain relevant and fit for purpose, and that they incorporate best practice for Official Statistics and large social surveys.
2. After discussions with interested parties and sessions of cognitive testing, we developed two alternative questionnaires. Like the established questionnaire, the first uses an agree/disagree format. Respondents are asked to express their attitude towards various assertions (for example, 'Overall, the quality of my course has been good') by selecting a response option such as 'Mostly agree'. This alternative questionnaire incorporated possible improvements to the wording of some questions, as well as some additional questions aiming to cover new areas of interest. It was not intended as a complete questionnaire, including only those question groups in the established questionnaire that we are interested in changing.
3. The second questionnaire uses a different format. Rather than asking respondents to state their agreement or disagreement with an assertion, it poses direct questions (for example, 'Overall, how would you rate the quality of your course?'). The responses to these are tailored to the question, rather than being identical throughout. We developed this direct question format to reflect a near consensus in survey literature, and among the survey practitioners we spoke to, that direct questions often lead to higher quality responses than agree/disagree scales.
4. The two questionnaires were piloted alongside the established survey in January and February 2022. The responses were analysed, with a view to answering the following questions:
    a. Do the modified and new questions work well for our audience?
    b. Do the new questionnaires work well as a whole? For example, are there signs that any questions are redundant, or so divergent that they do not belong in the survey?
    c. Do the new questions work across both modes of response (online and phone)?
    d. Do the proposed response options for the direct questions work well?
    e. Can we see any evidence in practice of the theoretical benefits of the direct question format?
5. Our overall aim was to establish whether either of the questionnaires was good enough to introduce, as it stands, in 2023; and if not, to identify what further research and development is needed.

Note on terminology

6. Throughout, we refer to the current NSS as the 'established questionnaire' (EQ). Of the two alternative questionnaires, we refer to the agree/disagree version as 'pilot one', and to the direct question version as 'pilot two'.
7. Sometimes we need to refer to questions in one of the questionnaires. We use the following format: EQ1 (Question 1 of the established questionnaire); P1Q1 (Question 1 of the pilot one questionnaire); P2Q1 (Question 1 of the pilot two questionnaire).

Key findings

8. The two pilots successfully collected a large number of responses from a wide spread of students.
9. Both questionnaires functioned well but evidence is mixed on whether one performed better than the other. Neither questionnaire appears ready for the field at this point. Further research such as cognitive testing and further piloting could help to refine the questions.
10. There are interesting signs that the direct question format reduces the mode of response effect and may decrease acquiescence amongst respondents.
11. We found removing the midpoint response option – currently 'neither agree nor disagree' – to be viable. However, it has the effect of increasing the proportion of positive responses, which may make the data less effective in identifying highly performing courses and providers.
12. The responses suggest that the questionnaires have a single underlying concept. All the questions are related, including the new questions, which arguably extend beyond the territory of the established questionnaire. However, the proposed sub-groups of questions are not always easy to detect within the patterns of responses.
13. Some of our research questions were difficult to answer because of the number of changes tested, and the design of the pilot, with the new questions asked directly after the established survey. This should be considered when planning future pilots.

Additional data

14. Much of the data referred to in this report can be viewed in more detail using the interactive dashboard available alongside this publication at www.officeforstudents.org.uk/publications/national-student-survey-2022-pilot/.

Method

15. The pilot responses were collected on behalf of the OfS by the NSS survey contractor, Ipsos Mori. All responses come from students who were eligible to participate in the established NSS. Starting in early January 2022, we invited students who had completed the established survey to participate in the pilots. If they accepted this invitation, we assigned them at random to one of the two pilot questionnaires. Students participated in the pilot immediately after completing the established survey.

16. Our aim was to collect 10,000 online responses and 1,000 phone responses for each pilot. Sampling for each pilot stopped in mid-February, once both targets had been achieved. We largely relied on the size of the sample to ensure that we collected responses from a wide range of students. For some smaller groups, such as distance learning students, we implemented quotas. These were monitored by the survey company but were ultimately achieved naturally, without prioritisation.

17. The sample was designed to collect a large number of responses quickly across a wide range of student groups. This is what we needed to test whether the pilot questionnaires worked, and to get results within a tight timescale. The sample was not designed to be 'representative', or to allow us to draw conclusions about the student experience in general. Any such conclusions should be drawn with caution, principally because the pilot data is very unevenly distributed across higher education providers, with some providers excluded altogether or only represented by a small handful of students.

18. The pilot responses, together with related paradata (for example, about mode and time of response), were linked by OfS analysts to information about student and course characteristics held by the OfS and sourced from the Higher Education Statistics Agency and the Education and Skills Funding Agency.

Evaluation of the pilots

Number of responses

19. Table 1 shows the total responses to the pilot questionnaires, split by mode of response.

Table 1: Responses to the pilot

	Online	Phone
Pilot one	11,558	1,018
Pilot two	12,100	1,000

20. We did not include all these responses in our analysis. For example, some students only answered a small handful of questions, and some students gave the same response to every question. Some students gave such discordant answers to the two questionnaires that we judged we could not take their responses at face value. In such cases, we can conclude something about the students' engagement with the pilot – perhaps they did not find it very interesting. But their responses tell us little about the new questions we want to test. To avoid

distracting from more relevant findings about the questions themselves, we have removed these students from the sample we used in analysis (the 'analytical sample').

21. In Annex B, we describe in full how the analytical sample was created. Table 2 shows the number of responses in the analytical sample.

Table 2: Responses in the analytical sample

	Online	Phone
Pilot one	10,202	973
Pilot two	11,529	992

22. The tables show that 1,356 respondents were removed from pilot one. Only 571 respondents were removed from pilot two. This is because more respondents showed signs of disengagement in pilot one, despite the shorter questionnaire. For example, 877 online respondents gave the same answer to every question in pilot one, compared with 465 online respondents in pilot two (despite pilot two consisting of more questions). This could be interpreted as a sign than the pilot two questionnaire is more engaging, but we caution against this. Students may have found pilot two more engaging simply because it was more novel, and not because it is better designed.

Responses by sub-group

23. We reviewed the online samples to check that they contained responses from a wide range of students. We paid particular attention to the following characteristics: subject of study, ethnicity, whether a student is studying by distance learning, and whether a student has a declared disability. This is because we know from earlier research and testing that these characteristics may affect how someone interprets questions about their student experience. We also considered the volume of responses from providers across the four different nations involved in the survey (England, Northern Ireland, Scotland, and Wales). This is because we wanted to be confident that any new questions work equally well across the national boundaries.

24. We found that for each group of interest, we had at least 150 online responses to each questionnaire (and in some cases, many more). The exceptions to this are Northern Irish providers for pilot one (146 responses) and a small number of subject areas. This was as predicted by the sample design and means that we have a reasonable chance of detecting differences in how subgroups respond. Annex B contains a breakdown of the analytical sample by student subgroup.

25. As noted in paragraph 10, the sample was not designed to include responses from students at all providers. 382 providers have students in the NSS population. Around 300 providers were included in each pilot, and there has been no attempt to survey all their students or to achieve a 50 per cent response rate. In the pilot one sample, for instance, 150 providers have responses from fewer than ten students.

Reliability of responses

26. As well as examining whether we had collected enough responses, we considered their quality. As discussed in paragraph 20, we established that a minority of students appeared not to have

engaged with the pilot despite agreeing to take part. These students have been removed from our analytical sample. But of the remaining students, how reliable are their responses? In asking this question, we were mindful of the possibility that the design of the pilot – in particular, the fact that the pilot questions were asked directly after the established questions – could have distorted how students respond. This could happen because of a 'priming effect': students' interpretation of the pilot questions was affected by their exposure to the main NSS questions. Or it could happen because of a mood change related to the pilot: for example, students could have become frustrated by the length of the process and therefore responded more negatively.

27. To address this question, we considered the pairs of identical questions which were included in both the established survey and pilot one. These questions are:
    - **EQ4/P1Q4:** My course has challenged me to achieve my best work.
    - **EQ5/P1Q7:** My course has provided me with opportunities to explore ideas or concepts in depth
    - **EQ6/P1Q18:** My course has provided me with opportunities to bring information and ideas together from different topics
    - **EQ7/P1Q9:** My course has provided me with opportunities to apply what I have learnt
    - **EQ8/P1Q11:** The criteria used in marking have been clear in advance
    - **EQ9/P1Q12:** Marking and assessment has been fair
    - **EQ10/P1Q13:** Feedback on my work has been timely.
28. We would in general expect students to give the same answer each time they encounter these questions. To the extent that this is not the case, there is a risk that something other than the questions themselves are influencing how students respond. This would mean that we need to be cautious in drawing conclusions based on the pilot data.
29. We tested first whether students' responses tend to become more positive, or more negative, the second time they encounter the question. We found that for five of the six questions this is not the case: while students sometimes respond differently to each occurrence of the question, there is no marked shift in either a positive or a negative direction. The exception is the E5/P1Q7 pair: the question about exploring ideas and concepts in depth. In this case, students tend to give a more negative response.
30. We then considered the consistency of responses. To what extent do students give the same response on the two occasions that they encounter these questions? We found that the percentage of students giving the same response on both occasions ranges by question from 59.6 per cent (for the EQ5/P1Q7 pair) to 71.8 per cent (for the EQ10/P1Q113 pair). Of the students who give a different response the second time they encounter the questions, most give a similar response: for example, responding 'Mostly agree' rather than 'Definitely agree'. A smaller proportion of respondents give a very different response. For example, looking at the question 'My course has challenged me to achieve my best work', 3,925 students answered 'Mostly agree' the first time they answered the question. Of these, 97 responded either 'Mostly disagree' or 'Definitely disagree' when they encountered the question for the second time.
31. It is not surprising that a small proportion of students give a different response when they re-encounter a question. For example, consider a student whose true feelings about a statement is on the threshold between 'Definitely agree' and 'Mostly agree'. They initially waver between

the two response options, before choosing ‘Definitely agree’ slightly arbitrarily. When encountering the question for a second time, they might well – equally arbitrarily – select ‘Mostly agree’ instead. This sort of case will arise whenever experience and opinions that are on a continuum are summarised using an interval scale.

32. However, we consider the proportion of students giving a different answer to be unexpectedly high. It may indicate some error in the responses, in the sense that students are sometimes giving answers that do not reflect their attitudes and experiences. This could happen because students’ concentration is waning during the pilot, and they are either misinterpreting questions or else answering them randomly. It could also happen because some questions are, in general, difficult to interpret and so students are prone to misunderstand them, regardless of their participation in the pilot. The data we hold does not allow us to further understand the phenomenon.

33. This issue means that we need to be careful when we compare non-identical questions across the established survey and the pilot. We need to be careful not to mistake the sort of variation described in paragraph 28– which occurs even for identical questions – for genuine differences in how students react to a question.

Do the questions work?

34. This section examines whether the pilot questions work, when considered individually. We focus on the newly introduced questions, but also note any issues relating to established questions.

Pilot one

Students responding ‘I cannot answer this question…’

35. For both pilot questionnaires, students could respond to any question by selecting the option ‘I cannot answer this question – I do not understand it, or I do not know the answer’. For most questions in pilot one, fewer than one per cent of students selected this response option, which is a reassuring result: we want to ask questions that almost every student can answer.

36. Within pilot one, four questions appear to be less well understood, or harder to answer. These are listed below, with the percentage of students using the ‘I cannot answer this question…’ option given in brackets. Most of these questions deal with entirely new themes, while P1Q18 (the students’ union question) is a new way of asking about an established theme.

    - **P1Q5:** There is an appropriate balance of breadth and depth in my course (8.2 per cent)
    - **P1Q18:** Overall, I am content with the students’ union (association or guild) at my institution (3.2 per cent)
    - **P1Q20:** My institution provides a free environment for the expression of ideas, opinions, and beliefs (2.8 per cent)
    - **P1Q23:** My institutions’ services to support my mental wellbeing were available when I needed them (2.3 per cent).

37. We examined whether students with particular characteristics were more likely to respond by selecting the option ‘I cannot answer this question…’ We found that:

a. Students with a declared disability were particularly likely to respond in this way to P1Q5 ('There is an appropriate balance of breadth and depth in my course'). 10.2 per cent of students with a declared disability selected this response option.

b. Distance learning students were particularly likely to choose this option when responding to P1Q18 ('Overall I am content with the students' union…') (7.7 per cent), P1Q20 ('My institution provides a free environment for the expression of ideas, opinions and beliefs') (6.6 per cent) and P1Q23 ('My institution's services to support my mental wellbeing were available when I needed them') (6.4 per cent).

Students responding 'Not applicable'

38. In pilot one, students could respond to any question by selecting the option 'Not applicable'. Since the questionnaire is not routed – all respondents are asked all questions – we aim to design questions that are applicable to most students. We found that for almost all pilot one questions, a very small proportion of students selected the 'This does not apply to me' response option. Two questions have lower levels of applicability:

- **P1Q18:** Overall, I am content with the students' union (association or guild) at my institution (6.2 per cent)
- **P1Q23:** My institutions' services to support my mental wellbeing were available when I needed them (17.5 per cent).

39. Looking at sub-groups of students, we can see that the finding in paragraph 37 is particularly pronounced for students studying through distance learning. 15.2 per cent of distance learning students used the 'Not applicable' option in response to P1Q18. 30.1 per cent of distance learning students used this response option to answer P1Q23. Students with a declared disability were **less** likely than other students to find P1Q23 inapplicable (10.7 per cent). We did not find other marked differences in use of 'Not applicable' by student subgroup.

Students exiting the questionnaire

40. In the established NSS and in the pilots, students cannot skip a question. They can only avoid responding to a question by exiting the questionnaire or hanging up the phone. We examined whether the number of students participating in the pilot decreased sharply at any point in the questionnaire. This would be an indicator that the question was particularly problematic. For pilot one, we found that almost all students who started the questionnaire reached the end. No question was associated with a marked drop in participation.

Range of responses

41. We examined the range of response options used by those participating in pilot one. We found that for every question, each response option was used by at least 1 per cent of respondents, and that no response option was used by more than 50 per cent of respondents. For most questions, the positive response options of 'Definitely agree' and 'Mostly agree' were used by around 80 per cent of respondents, and the middle option 'Neither agree nor disagree' used by around 10 per cent of respondents. The question about the students' union is atypical, with a lower proportion of positive options. In no case was the distribution of response options so surprising as to suggest that the question was flawed or had been comprehensively misunderstood.

Comparison with the established survey

42. When a pilot one question was based on an established survey question, we compared how individual students responded to the pair of questions, to see how changes in wording had affected their responses. To acknowledge the variation in responses that occurs even when the question itself has not changed (see paragraphs 25 to 32), we considered a change to be significant only when the net shift from positive to negative, or vice versa, was greater than five percentage points

43. For example, the first pair of questions is 'Staff are good at explaining things' versus 'Teaching staff are good at explaining course content'. We found that when individual students confronted the second question, 20 per cent of students gave a more positive response than they gave to the established question, and 13.5 gave a more negative response – resulting in a net difference of 6.5 percentage points.

44. Even with this restriction, there remain complexities. One question – 'My course has provided me with opportunities to explore ideas and concepts in depth' – was repeated word-for-word in both established survey and pilot one, yet individual students tended to give a more negative response when they encountered the question for the second time. We judge that this effect is too extreme to be due to random variation in responses. It may be that earlier questions in the pilot questionnaire have affected the considerations students bring to bear when they answer the question. We have included this pair of identical questions in Table 3.

Table 3: Changes in students' responses (for similar questions) between established survey and pilot one

Question pair	Finding
EQ1: Staff are good at explaining things. **P1Q1:** Teaching staff are good at explaining the course content.	Students tended to respond **more positively** to the pilot question, perhaps because of the narrower focus.
EQ2: Staff have made the subject interesting. **P1Q2:** Staff have made the subject engaging.	No significant change.
EQ3: The course is intellectually stimulating. **P1Q3:** My course is intellectually stimulating.	No significant change.
EQ5: My course has provided me with opportunities to explore ideas and concepts in depth. **P1Q7:** My course has provided me with opportunities to explore ideas and concepts in depth.	The two questions are identical, but students tended to respond **more negatively** during the pilot. This may be because answering the preceding questions, P1Q5 and P1Q6, which are new, has shifted how students interpret the question.
EQ11: I have received helpful comments on my work. **P1Q14:** Feedback has enabled me to improve my work.	No significant change.

Question pair	Finding
EQ18: The IT resources and facilities provided have supported my learning well. **P1Q16:** It has been easy to access learning resources (digital and physical) provided by my institution when I needed to.	Students tended to respond **more positively** to the pilot question.
EQ19: The library resources (e.g. books, online services and learning spaces) have supported my learning well. **P1Q16:** It has been easy to access learning resources (digital and physical) provided by my institution when I needed to.	Students tend to respond **more negatively** to the pilot question.
EQ18: I have been able to access course-specific resources (e.g. equipment, facilities, software, collections) when I needed to. **P1Q6:** It has been easy to access learning resources (digital and physical) provided by my institution when I needed to.	No significant change.
EQ18: The IT resources and facilities provided have supported my learning well. **P1Q17:** Learning resources (digital and physical) provided by my institution have supported my learning well.	Students tended to respond **more positively** to the pilot question.
EQ19: The library resources (e.g. books, online services and learning spaces) have supported my learning well. **P1Q17:** Learning resources (digital and physical) provided by my institution have supported my learning well.	Students tended to respond **more negatively** to the pilot question.
EQ20: I have been able to access course-specific resources (e.g. equipment, facilities, software, collections) when I needed to. **P1Q17:** Learning resources (digital and physical) provided by my institution have supported my learning well.	No significant change.

Question pair	Finding
EQ26: The students' union (association or guild) effectively represents students' academic interests. **P1Q18:** Overall, I am content with the students' union (association or guild) at my institution.	Students tend to respond **more positively** to the pilot question.
EQ27: Overall, I am satisfied with the quality of the course. **P1Q19:** Overall, the quality of my course has been good.	Students tend to respond **more positively** to the pilot question.

45. These findings do not imply that there are problems with the pilot questions. When a question has changed, it is unsurprising that many students respond differently. However, we should reflect upon whether the changes we see are consistent with the theories informing the questionnaire development. For example, the theory behind the changes to the learning resource questions was that, in the modern world, students cannot effectively distinguish between IT resources, library resources and course-specific resources. The pilot data casts some doubt on this. Students tend to respond more negatively when asked about IT, rather than learning resources in general, and more positively when asked about libraries. Similarly, we should consider whether we expected to see students tending to give a more positive response when asked about quality, rather than about overall satisfaction.

Time taken to respond

46. In the online versions of the pilots, as in the established survey, each question is displayed on a single screen. Students move between screens by clicking a 'next' button. We analysed the time that elapsed between each click of the 'next' button, as a way of understanding how long students spent on each question. We were looking for questions that took a long time to answer, or questions that were skimmed over very quickly. This data was only available for online respondents.

47. This analysis did not raise any concerns about the pilot one questions. For most questions, the median time spent on a question is four or five seconds, with an upper value (95th percentile) of between ten and 15 seconds. Understandably, the students spent longest answering the first question, presumably because the format is new to them. Aside from this, the following four questions had slightly longer response times:

    - **P1Q19:** Overall the quality of my course has been good (median: six seconds, upper value: 19 seconds)
    - **P1Q20:** My institution has made me aware of services to support my mental wellbeing (median: seven seconds, upper value: 21 seconds)
    - **P1Q21:** My institution's services to support my mental wellbeing were available when I needed them (median: five seconds, upper value: 18 seconds)
    - **P1Q22:** My institution provides a free environment for the expression of ideas, opinions and beliefs (median: five seconds, upper value: 18 seconds).

48. These longer times suggest that the questions require more thought to answer, but they are not so extreme as to suggest that these questions do not work.

Individual questions: Summary of findings

49. Most of the new questions included in pilot one worked well, as, unsurprisingly, did the established questions repeated in the pilot. The following areas may require further consideration or research:

    a. Some of the new questions are relatively problematic. These are the questions relating to breadth and depth, the students' union, free expression of ideas, and the availability of mental wellbeing services. Some of these issues are particularly pronounced for distance learning students, and students with a declared disability.

    b. In some cases, rewording a question appears to elicit a different response from students, often in a more positive direction. While this is not of itself a concern, we should consider whether we understand these changes and whether they are consistent with our aims in rewording the question.

Pilot two

Students responding 'I cannot answer this question'

50. For both pilot questionnaires, students could respond to any question by selecting the option 'I cannot answer this question – I do not understand it, or I do not know the answer'. For most questions in pilot two, fewer than 2 per cent of students selected this response option.

51. Within pilot two, the questions that appear to be less well understood, or harder to answer, are listed below, with the percentage of students using the 'I cannot answer this question…' given in brackets. Some of these questions – P2Q8, P2Q9 and P2Q29 – deal with entirely new themes, but the other questions ask about established themes in new ways.

- **P2Q1:** Are teaching staff good at explaining course content? (2.3 per cent)
- **P2Q8:** Does your course contain the right balance of depth and breadth? (8.9 per cent)
- **P2Q9:** Does your course contain the right balance of directed and independent study? (2.3 per cent)
- **P2Q17:** Are you able to get good advice about study choices? (3.1 per cent)
- **P2Q23:** Do staff value students' opinions about the course? (2.9 per cent)
- **P2Q24:** Do staff act on students' feedback? (5.6 per cent)
- **P2Q25:** Has the students' union (association or guild) had a positive impact on your experience? (5.2 per cent)
- **P2Q29:** How easy is it to access your university or college's mental wellbeing services? (6.2 per cent)

52. We examined whether students with particular characteristics were more likely to select the option 'I cannot answer this question…' We found that:

    a. Students with a declared disability were particularly likely to respond in this way to P2Q8 ('Does your course contain the right balance of depth and breadth?'). 10.9 per cent of students with a declared disability selected this response option.

b. Students with no known disability were particularly likely to select this option for P2Q29 ('How easy is it to access your university or college's mental wellbeing services?'). 6.8 per cent of respondents without a declared disability responded this way.

c. Distance learning students were particularly likely to choose this option when responding to P2Q23 ('Do staff value students' opinions about the course?') (13.5 per cent), P2Q24 ('Do staff act on students' feedback?') (21.9 per cent), P2Q25 ('Has the students' union (association or guild) had a positive impact on your experience?') (9.7 per cent) and P2Q29 ('How easy is it to access your university or college's mental wellbeing services?') (18.5 per cent).

53. It is interesting that students participating in pilot two are more likely to use the 'I cannot answer this question…' response option. This finding is in tension with the theory that direct questions are easier to understand and respond to than questions using an agree/disagree format. For example, the direct question 'Are teaching staff good at explaining course content?' is no harder to understand than the corresponding statement 'Teaching staff are good at explaining course content'. Yet 2.3 per cent of students said they could not answer the direct question, whereas 0.4 per cent of students said they could not respond when asked to agree or disagree with the corresponding statement. Although these differences in percentage are relatively small, they are statistically significant.

54. There are a number of possible explanations here:

a. Students in pilot two are confronting a new questionnaire format, directly after completing the established questionnaire (which uses an agree/disagree format). It may have taken some students a little time to adjust to the new format. This could explain why the very first question ('Are teaching staff good at explaining course content?') appears to be harder to answer than some of the later questions, despite its apparent simplicity.

b. Feedback we have received from interviewers suggests a disconnect between some of the pilot two questions and their response options. For example, the response options for P2Q1 ('Are teaching staff good at explaining course content?') has response options based on frequency, namely:

i. Very often.

ii. Fairly often.

iii. Not very often.

iv. Rarely or never.

v. This does not apply to me.

These response options were selected to address a problem identified in earlier NSS research: students may struggle to answer a question about their course when their experiences have been mixed.[1] For example, if the teaching staff are usually very good at explaining things but occasionally very bad, how should they react to the statement 'Teaching staff are good at explaining course content'? It may be the

[1] See for example 'UK review of information about higher education: Cognitively tested questions for the National Student Survey', section 4.1.1, available at https://dera.ioe.ac.uk/24581/.

case, however, that in addressing this initial problem we have created a new problem: the response options given are slightly unexpected as responses to the question asked, and this means that some students struggle to respond.

c. In some cases, the higher use of the 'I cannot answer the question…' response option may reflect better engagement with the questions. This takes a little explaining. Some of the new questions piloted are quite hard to answer, both because they are conceptually complicated and because they require judgment. The general theory is that respondents engage better with direct questions, in the sense that they read the question more carefully and give answers that better reflect their opinions and experiences. In contrast, respondents to agree/disagree questionnaires may develop strategies for answering questions regardless of whether they fully understand the question (such as always responding 'Mostly agree'). If all this is true, then the higher usage of 'I cannot understand the question…' may not indicate a problem with the question wording and format. It may be the case that questions that are more engaging are prompting students to realise and declare when they find the question hard to answer.

d. Finally, pilot one included a middle option ('Neither agree nor disagree'), whereas pilot two does not. It may be that some students who did not feel able to answer the question defaulted to the middle option in pilot one, whereas in pilot two the absence of this alternative led them to use the more appropriate 'I cannot answer this question…' response option.

55. The pilot design does not allow us to determine which of these explanations is correct. All may have some elements of truth. This could be explored through further research.

Students responding 'This does not apply to me'

56. Students were able to respond to any question in pilot two by selecting the option 'This does not apply to me'. In addition, for some questions, students were able to declare that the question did not apply to them by using a response option specific to the question. For example, P2Q14 is 'Are you able to contact teaching staff when you need to?' An additional response option for this question is 'I have not needed to contact teaching staff'. These additional options were introduced for questions that are clearly conditional. For example, it is not possible to answer P2Q14 using the four main response options if you have never needed to contact teaching staff.

57. Table 4 shows those questions for which more than 5 per cent of students used the response option 'This does not apply to me', or an additional response option specific to the question.

Table 4: Questions that more than 5 per cent of respondents declared to be not applicable

Question	Proportion responding 'This does not apply to me' (%)	Proportion using question-specific response option (%)	Total proportion declaring that the question does not apply (%)
P2Q7: When working with other students as part of your course, how helpful was this for your learning? Additional response option: 'I did not work with other students as part of my course.'	1.3	4.8	6.1
P2Q17: Are you able to get good advice about study choices? Additional response option: 'I have not needed advice about study choices.'	2.3	6.6	8.9
P2Q26: Has the students' union (association or guild) had a positive impact on your experience? Additional response option: 'My university/college does not have a students' union (association or guild).'	14.8	1.2	16
P2Q29: How easy is it to access your university or college's mental wellbeing services? No variant of 'not applicable', but students could respond 'Prefer to not say.'	16.5	–	16.5

58. Reviewing the response by sub-group, we found that distance learning students were particularly likely to respond to these questions with the 'This does not apply to me' option. For example, 28.3 per cent gave this response for P2Q25 and 29.6 per cent for P2Q29. Students with no known disability were also particularly likely to select this response option for P2Q29 (18.8 per cent). We did not identify any other marked differences in the use of the 'This does not apply to me' option by student subgroup.

59. Pilot two has slightly higher levels of inapplicability than pilot one. This seems to be closely related to the additional response options, and may not be a bad thing. It may be the case that the direct questions used in pilot two, together with the additional response options, prompt students to realise that questions do not apply to them. This is preferable to the alternative, which would be to choose a response option that does not reflect their experience in order to proceed with the questionnaire.

60. While the higher levels of inapplicability may not reveal a problem with pilot two, the questions listed should be reviewed to see whether their applicability can be approved, and to consider the possibility of routing.

Students exiting the questionnaire

61. In the established NSS and in the pilots, students cannot skip a question. They can only avoid responding to a question by exiting the questionnaire or hanging up the phone. We examined whether the number of students participating in the pilot decreased sharply at any point in the questionnaire. This would be an indicator that the question was particularly problematic. For pilot two, we found that almost all students who started the questionnaire reached the end. No question was associated with a marked drop in participation.

Range of responses

62. We examined the range of response options used by those participating in pilot two. We found that for every question, each response option was used by at least 1 per cent of respondents, and that no response option was used by more than 65 per cent of respondents. On average, the two most positive response options for each question were selected by 84.2 per cent of respondents. The question about the students' union is atypical, with a lower proportion of positive options.

63. For the following questions, responses were particularly clustered at the positive end of the scale:

    a. **P2Q4:** 'Do you feel challenged by your course?' 52.2 per cent of respondents gave the most positive response, and 34.9 per cent gave the second most positive response.

    b. **P2Q13:** 'Were you given the marking criteria in advance?' 63.6 per cent of respondents gave the most positive response, and 24.2 per cent gave the second most positive response.

    c. **P2Q15:** 'Are you able to contact teaching staff when you need to?' 52.9 per cent of respondents gave the most positive response, and 37.5 per cent gave the second most positive response.

    d. **P2Q20:** 'Have you been able to access the learning resources (either digital or physical) that you need?' 54.9 per cent of respondents gave the most positive response, and 36.8 per cent gave the second most positive response.

64. None of these distributions are so radical as to suggest that the question is not worth asking, or that it has been radically misunderstood. However, we should consider whether changes in the question wording, or in the response options, might yield a more nuanced understanding of the student experience.

Comparison with the established survey

65. When a pilot two question was based on an established survey question, we compared how individual students responded to the pair of questions. In almost all cases, students responded more positively to the pilot two question. To a large extent, this can be attributed to the removal of the middle response option. Students who gave a neutral 'Neither agree nor disagree' response to a pilot one question tended to give a positive response option to the corresponding pilot two question.

66. It is hard to disentangle this general increase in positivity from the impact of changes in wording. We have identified question pairs in which net positive change is greater than 15 percentage points, or there is a net negative change. These are listed in Table 5. In these

cases, it is likely that something has affected students' responses other than the removal of the middle option.

67. For the purpose of comparing responses between the two questionnaires, we mapped response options to each other depending on their position in the response scale. For example, we considered the response option 'Definitely agree' to be as positive as the response option 'Very well', given that both are the most positive option available. This, of course, is a simplistic assumption. In some cases, the change in students' responses will reflect the fact that the response options do not map directly to each other, rather than any change in question wording.

Table 5: Changes in students' responses (for similar questions) between established survey and pilot two

Question pair	Finding
EQ1: Staff are good at explaining things. **P2Q1:** Are teaching staff good at explaining course content?	Students tended to respond **more positively** to the pilot question. Given similar findings for pilot one, this is likely to be partly due to the change in wording.
EQ4: My course has challenged me to achieve my best work. **P2Q4:** Do you feel challenged by your course?	Students tended to respond **more positively** to the pilot question. This is not surprising: one can feel challenged, without being challenged to achieve one's best work. It is possible that the challenge reported in response to P2Q4 is not a wholly positive experience.
EQ8: The criteria used in marking have been clear in advance. **P2Q13:** Were you given the marking criteria in advance?	Students tended to respond **more positively** to the pilot question. This may suggest a separation between receiving the criteria in advance and regarding it as clear. But it may also reflect the use of frequency response options for pilot two.
EQ9: Marking and assessment has been fair. **P2Q11:** Has marking and assessment been fair?	Students tended to respond **more positively** to the pilot question. Apart from the effect of removing the middle option, this shift largely relates to students who responded 'Mostly agree' to EQ9 and who selected the most positive response option – 'Very often' when answering P2Q11.
EQ18: The IT resources and facilities provided have supported my learning well. **P2Q20:** Have you been able to access the learning resources (either digital or physical) that you need?	Students tended to respond **more positively** to the pilot question. Apart from the effect of removing the middle option, the shift in positivity is mainly due to students who responded 'Mostly agree' to EQ18 and who selected the most positive response option – 'Very often' when answering P2Q20.
EQ18: The IT resources and facilities provided have supported my learning well. **P2Q21:** How well have the physical and/or digital resources supported your learning?	Students tended to respond **more positively** to the pilot question. As above, a large proportion of students who responded 'Mostly agree' to EQ18 selected the most positive option – 'Very well' – when answering P2Q1.

Question pair	Finding
EQ20: I have been able to access course-specific resources (e.g. equipment, facilities, software, collections) when I needed to. **P2Q20:** Have you been able to access the learning resources (either digital or physical) that you need?	Students tended to respond **more positively** to the pilot question. As above, a large proportion of students who responded 'Mostly agree' to EQ20 selected the most positive option – 'Very often' – when answering P2Q0.
EQ22: I have had the right opportunities to work with other students as part of my course **P2Q7:** When working with other students as part of your course, how helpful was this for your learning?	Students tended to respond **more negatively** to the pilot question. Large proportions of students responded 'Definitely agree' to EQ22 and 'Fairly helpful' to P2Q27, and 'Mostly agree' to EQ22 and 'Not very helpful' to P2Q7. Students who gave a neutral response option to EQ22 shifted in almost equal volumes in a negative and positive direction. It seems likely that there is some distance between the concept of the 'right opportunities', and the opportunities being helpful.
EQ25: It is clear how students' feedback on the course has been acted on. **P2Q24:** Do staff act on students' feedback?	Students tended to respond **more positively** to the pilot question. This positivity shift is mainly explained by the large proportion of students (22.5 per cent) who responded to EQ25 using the middle option. A large majority of these shifted in a positive direction. In addition, a large proportion of students responded 'Mostly agree' to EQ25 and 'To a great extent' to P2Q24.
EQ26: The students' union (association or guild) effectively represents students' academic interests. **P2Q25:** Has the students' union (association or guild) had a positive impact on your experience?	Students tended to respond **more negatively** to the pilot question. 31.5 per cent of respondents responded to EQ26 using the middle option and, unusually, the majority of these gave a more negative response option 'Hardly at all' or 'Not at all' to P2Q25.

Time taken to respond

68. In the online version of the pilots, as in the established survey, each question is displayed on a single screen. Students move between screens by clicking a 'next' button. We analysed the time that elapsed between each click of the 'next' button, as a way of understanding how long students spent on each question. We were looking for questions that took a long time to answer, or questions that were skimmed over very quickly.

69. This analysis did not raise any concerns about the pilot two questions. For most questions, the median time spent on a question is four or five seconds, with an upper value (95th percentile) of between 11 and 16 seconds. Understandably, the students spent longest answering the first question, presumably because the format is new to them. Aside from this, the following four questions had slightly longer response times:

    - **P2Q20:** Have you been able to access the learning resources (either digital or physical) that you need? (Median: eight seconds, upper value: 22 seconds)

- **P2Q29:** How easy is it to access your university or college's mental wellbeing services? (Median: seven seconds, upper value: 21 seconds)
- **P2Q30:** During your studies, have you felt free to express your ideas, opinions, and beliefs? (Median: six seconds, upper value: 20 seconds)
- **P2Q31:** Has your course given you the knowledge and skills you think you will need for your future? (Median: six seconds, upper value: 19 seconds)

70. These longer times suggest that the questions require more thought to answer, but they are not so extreme as to suggest that these questions do not work.

Individual questions: Summary of findings

71. Most of the questions in pilot two worked well, but the following areas may require further consideration or research:
    a. Some questions appear to be harder to answer and require further attention. This is particularly true of the question about the balance of depth and breadth (P2Q8) and the question about accessing mental wellbeing services (P2Q29).
    b. It may be possible to improve the applicability of some questions by rewording. Introducing routing to the questionnaire would be an alternative approach.
    c. Some questions – most strikingly, P2Q13 ('Were you given the marking criteria in advance?') – received very positive responses. While this is not necessarily undesirable, we should consider whether these questions could be refined to measure with more nuance the student experience.
    d. Changes in wording appear to have made a difference to how students respond. To give just one example, P2Q7 ('When working with other students as part of your course, was it helpful?') invokes a much more negative response than the established EQ22 ('I have had the right opportunities to work with other students as part of my course'). In these cases, we should consider carefully whether the new question captures the concepts of interest.

Do the subscales work as expected?

72. In the established questionnaire, eight subscales are presented to respondents. They are represented by a heading in the online survey and introduced on the phone by the interviewers as the start of the new section. Agreement rates known as 'scale scores' are calculated for each subscale and are often used as a helpful way of summarising the survey responses by theme.
73. For the pilot questionnaires, the question items were also grouped into sections presented as headings to the online respondents. We have analysed the structure of the pilot surveys to see whether the sections in these new surveys correspond with our expectations and can be used to construct subscales as in the established questionnaire. This analysis does not include the phone respondents, as these responses may have different patterns of response.
74. Our approach relies on four main stages:
    - reviewing item response distributions
    - examining inter-item correlations

- identifying the structure of underlying factors in the responses
- reviewing underlying factors for internal consistency.

We describe each of these stages in turn for both the pilot questions.

Pilot one

Reviewing item response distributions

75. Pilot one was not intended as a complete questionnaire: it included only question groups that we were considering changing. Therefore, when considering the relations between the pilot one questions, we have included questions asked as part of the established survey which would combine with the pilot two questions to create a complete questionnaire. The question groups that we have added from the established questionnaire are academic support, organisation and management, learning community and student voice.

76. Our review of the item response distributions for pilot one shows that, consistent with previous iterations of the NSS, there is a clustering of response options at the positive end of the scale. For all questions, the most common response options were either 'Mostly agree' or 'Definitely agree'. See paragraph 41 for discussion of the distribution of response options.

77. All questions had an agreement rate lower than 90 per cent, indicating that although there is a clustering at the positive end of the scale there is enough variation in responses to retain them in the item pool for correlation analysis and principal components analysis.

Examining inter-item correlations

78. We found that all items are correlated to some extent.[2] There are very few correlations lower than 0.3, indicating a moderately strong correlation between all items. P1Q18 ('Overall, I am content with the students' union (association or guild) at my institution'), is the item with the lowest average correlation with the other questions.

79. Looking at the correlations within each of the sections of the pilot one questionnaire, we see that these are in many cases higher than correlations with other questions. The correlations are often above 0.5, which indicates that there may be more overlap between the concepts each item measures than is ideal.[3] The correlations within each subscale are not very much higher than the correlations of those questions with items outside the sections, which indicates that the sections in the questionnaire are not entirely conceptually separate from each other. For the purposes of the NSS this is as expected and replicates findings from the established questionnaire. This shows that the questions are covering a domain of experiences that are closely related to each other and to the underlying quality of the student experience. It does, however, mean that in constructing scores for each subscale there is a risk of counting experiences that are very similar twice, and thus overweighting that aspect of the scale score.

80. In particular, the correlations between P1Q16 and P1Q17 ('Learning resources' subsection) and P1Q22 and P1Q23 ('Mental health' subsection) are above 0.7. This indicates that the

[2] We provide the correlation matrices showing the correlations between question pairs as part of the additional data supporting this report, available alongside the report at www.officeforstudents.org.uk/publications/national-student-survey-2022-pilot/.

[3] See, for example, Clark, L, and D Watson (1995), Constructing validity: Basic issues in objective scale development. Psychological Assessment, 7, pp309-319.

questions are substantially overlapping in the concepts they cover; a higher response to one of these questions is highly predictive of a higher response in the other. These questions should be considered carefully when selecting questions to include in the final version of the survey following this pilot, as one of the questions in each pair may be redundant.

Identifying the structure of underlying factors

81. The established questionnaire results show an underlying structure of eight factors, which map closely to the expected subscales shown as sections within in the survey:[4]
    - Teaching on my course
    - Learning opportunities
    - Assessment and feedback
    - Academic support
    - Organisation and management
    - Learning resources
    - Learning community
    - Student voice.
82. When identifying a factor structure to describe the pilot one data, we found it necessary to remove some questions. We removed the 'Mental health' and 'Learning resource' questions because of their high degree of inter-correlation, as described in paragraph 79, and because they created factors with only two items each. We removed the students' union question (P1Q16) and the freedom of expression question (P1Q20) because of their weaker correlations with the other questions. We found that with these questions removed, a factor structure of five underlying factors describes the data well. Removal of the questions from this exercise does not suggest that they should be removed from the questionnaire. But it does imply that it would not be easy to construct formal subscales including these questions.
83. In our principal components analysis model, the initial factor solution shows that all items loaded onto a single underlying factor, with loadings between 0.56 and 0.79. This shows that all the items included in our model are measuring different aspects of the same underlying thing, which we term 'student academic experience'.
84. Table 6 shows the factor loadings for the question items grouped by the subsections they are presented in within the questionnaire.
85. For our purposes we are terming as a 'loading' anything with an absolute regression coefficient above 0.3. Anything below this would be a 'low loading' and anything below 0.1 would be a 'zero loading'. It is possible for factors to load negatively. Where a question item has more than one loading above 0.3, this indicates the item is a complex variable. The number of these included in the survey should be limited.

[4] For the established questionnaire questions, see Table A1 in www.officeforstudents.org.uk/advice-and-guidance/student-information-and-data/national-student-survey-nss/.

Table 6: Factor loadings for question items in pilot one (supplemented with questions from the established questionnaire)

Questionnaire subsection	Question	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Teaching on my course	**P1Q1**	0.51	0.16	0.01	0.35	-0.02
	P1Q2	0.70	0.05	-0.01	0.28	-0.05
	P1Q3	0.86	-0.07	0.04	0.14	-0.12
	P1Q4	0.73	0.10	0.05	0.02	0.02
Learning opportunities	**P1Q5**	0.67	0.07	0.16	-0.04	0.00
	P1Q6	0.44	0.15	0.25	-0.06	0.15
	P1Q7	0.70	0.04	0.12	-0.05	0.11
	P1Q8	0.66	0.04	0.05	-0.06	0.20
	P1Q9	0.62	0.07	-0.06	-0.09	0.32
	P1Q10	0.23	0.00	0.00	-0.09	0.73
Additional questions	**P1Q21**	0.64	0.09	-0.01	-0.01	0.19
Academic support (EQ)	**EQ12**	0.10	0.09	0.06	0.67	0.05
	EQ13	0.23	0.22	0.09	0.44	0.13
	EQ14	0.22	0.10	0.10	0.42	0.21
Organisation and management (EQ)	**EQ15**	0.18	0.07	0.52	0.32	-0.03
	EQ16	0.02	-0.02	0.94	-0.15	-0.01
	EQ17	-0.01	0.07	0.65	0.24	0.03
Assessment and feedback	**P1Q11**	-0.08	0.82	0.02	-0.06	0.11
	P1Q12	-0.02	0.85	0.01	0.00	0.00
	P1Q13	-0.06	0.82	0.02	0.07	-0.05
	P1Q14	0.21	0.73	-0.02	0.02	-0.07
	P1Q15	0.39	0.51	0.05	-0.08	-0.02
Student voice (EQ)	**EQ23**	-0.01	0.08	0.10	0.29	0.59
	EQ24	0.05	0.15	0.17	0.40	0.38
	EQ25	-0.02	0.18	0.20	0.28	0.45

Note: Grey shading indicates a loading over 0.3.

86. As can be seen in Table 6, the factor loadings are similar to the subsections, which is reassuring. There is one factor loading which might be considered anomalous – P1Q10 ('I have had the right opportunities to work with other students as part of my course') – which loads strongly with the questions in the 'Student voice' subsection. We will examine further the structure of each of these factors and explore their internal consistency in paragraphs 87 and 88.

87. There are four complex variables in the Table 6: P1Q1, EQ15, P1Q15, and EQ24. These are questions that load onto more than one factor. These are not surprising, in the sense that the factors they load onto correspond to related subsections (for example, 'teaching on my course' and 'academic support'). There are no questions that load onto more than two factors.

Counterintuitively, EQ24 ('Staff value students' views and opinions about the course') does not load as strongly with the other 'Student voice' questions as with the 'Academic support' questions. However, the differences between the two loadings are small and so it may make semantic sense to include it with the other 'Student voice' questions rather than the 'Academic support' questions when creating subscales.

88. We can also see in Table 6 that most questions have a zero loading on at least one factor. Considering the factors, none of them share more than one complex variable, which shows that they are mostly distinct from each other. Each factor contains zero loadings for most of the items that do not load on it. All of these observations increase our confidence that the factor model represents an underlying structure that is present within the data.

Internal consistency of the underlying factors

89. We reviewed the internal consistency of the factors we identified. We can see from Table 7 that each of these factors has an internal consistency that is close to or above 0.8. This indicates that these factors are consistent with themselves in terms of the responses to the questions included. Looking at our anomalous loading for P1Q10, we can see that removing this item from the student voice factor slightly increases the internal consistency.

90. We have chosen not to include EQ24 in the academic support factor, despite its slightly higher loading on that factor, because we do not find the same pattern of loadings in the established questionnaire from which the student voice questions have been replicated. We prefer instead to include it with the other student voice questions.

Table 7: Cronbach's alpha for each subscale identified in the principal components analysis for pilot one

Factor	Alpha
Factor 1 (Teaching and learning)	0.93
Factor 2 (Assessment and feedback)	0.86
Factor 3 (Organisation and management)	0.79
Factor 4 (Academic support, excluding EQ24)	0.85
Factor 5 (Student voice, including P1Q10)	0.83
Factor 5 (Student voice, with P1Q10 removed)	0.84

Pilot two

Reviewing item response distributions

91. As for pilot one, but to a greater extent, there is a clustering of responses at the positive end of the scale. See paragraph 62 for discussion. Nonetheless, we concluded that there is enough variation in responses to retain all questions in the correlation analysis and principal components analysis. In particular, no more than 63 per cent of respondents chose a single response option for any question.

Examining inter-item correlations

92. Compared with pilot one the inter-item corelations are lower. The strength of the correlations varies between 0.14 and 0.76. There are five items that are on average more weakly correlated

with the other items. The questions are P2Q4 (mean correlation 0.28), P2Q7(mean correlation 0.28), P2Q13 (mean correlation 0.26), P2Q25 (mean corelation 0.29) and P2Q28 (mean correlation 0.28). This suggests that these questions may be more weakly related to the core underlying concept of ‘student academic experience‘.

93. Some of the questions correlate highly with each other, which may suggest that there is some redundancy between these questions. This is true of two of the academic support questions P2Q16 and P2Q17 (correlation=0.66):

    - How well have teaching staff supported your learning?
    - Are you able to get good advice about study choices?

    It also applies to the organisation and management questions P2Q18 and P2Q19 (correlation=0.67):

    - How well organised is the course?
    - Have changes to the course been clearly communicated?

    Finally, it is true of two of the student voice questions P2Q23 and P2Q24 (correlation=0.76):

    - Do staff value students' opinions about the course?
    - Do staff act on students' feedback?

94. This last pair has a correlation over 0.7, and one of these questions should be considered for removal from the survey as they are substantially overlapping in the patterns of responses elicited from the pilot participants.

Identifying the structure of underlying factors

95. As with pilot one, we found it necessary to remove some questions to find a factor structure that describes the data, as follows:

    a. The learning resources questions (P2Q20 and P2Q21), based on the high correlation between these items, which creates a two-item factor.

    b. The mental health questions (P2Q23 and P2Q24), based on very high correlation with each other.

    c. The students’ union question (P2Q25) and the working with other students question (P2Q7). In exploratory attempts, these two questions loaded together, forming another unhelpful two-question scale.

    d. P2Q13 (‘Were you given the marking criteria in advance?’) This question loaded to a single factor. It is independently known to be anomalous based on the very positive responses.

96. Again, removal of the questions from the factor analysis does not imply that they should be removed from the questionnaire. But it does suggest that, in their present form, it would be harder to defend the construction of formal subscales that include them.

97. With these questions removed, the data is described by a four-factor structure, as shown in Table 8. The factors appear to be compound in nature, covering seemingly disparate aspects of student experience.

Table 8: Factor loadings for question items in pilot two

Survey section	Question	Factor 1	Factor 2	Factor 3	Factor 4
Teaching on my course	**P2Q1**	0.53	0.14	0.08	0.17
Teaching on my course	**P2Q2**	0.42	0.25	0.03	0.27
Learning opportunities	**P2Q9**	0.44	0.16	0.07	0.19
Academic support	**P2Q16**	0.67	0.20	0.03	0.01
Academic support	**P2Q17**	0.63	0.23	0.03	-0.01
Organisation and management	**P2Q18**	0.79	-0.16	0.03	0.20
Organisation and management	**P2Q19**	0.81	-0.24	0.09	0.14
Student voice	**P2Q22**	0.68	0.05	0.10	-0.03
Student voice	**P2Q23**	0.83	0.06	0.04	-0.08
Student voice	**P2Q24**	0.84	0.04	0.02	-0.09
Free speech	**P2Q30**	0.44	0.45	0.02	-0.16
Skills for future	**P2Q31**	0.25	0.52	-0.01	0.17
Learning opportunities	**P2Q5**	-0.09	0.77	0.08	0.12
Learning opportunities	**P2Q6**	0.03	0.67	0.09	0.12
Assessment and feedback	**P2Q10**	0.21	0.30	0.31	0.16
Assessment and feedback	**P2Q11**	0.29	0.06	0.52	0.02
Assessment and feedback	**P2Q12**	0.11	-0.02	0.78	0.01
Assessment and feedback	**P2Q13**	-0.13	-0.03	0.90	-0.03
Academic support	**P2Q14**	0.30	0.23	0.36	0.02
Teaching on my course	**P2Q3**	0.20	0.20	0.00	0.62
Teaching on my course	**P2Q4**	-0.07	0.02	0.02	0.89
Learning opportunities	**P2Q8**	0.35	0.21	0.05	0.36

Note: Grey shading indicates a loading over 0.3.

98. The first factor contains questions related to the quality of teaching, academic support, organisation and management, student voice and free expression of ideas. The second factor relates to assessments and feedback, although one of the academic support questions also loads here. The range of learning opportunities offered is the theme of the third factor, and the fourth factor relates to the intellectual challenge of the course.

99. The first factor contains many of the key domains of student experience that would, in the established questionnaire, usually load onto five separate factors. This suggests that changes to the response scales and question wordings have impacted how students respond to the survey. It also suggests that there is a single underlying driver of these aspects of student experience. It may be the case that something about the culture at a provider leads to good learning and teaching, well organised courses and receptiveness to the student voice.

100. The other questions cover sufficiently distinct aspects to load onto identifiably separate factors. It is especially interesting that the teaching questions to do with quality of interactions (P2Q1, P2Q2) load onto a separate factor from the questions related to how challenging the content of the course is (P2Q3, P2Q4).

101. There are three items with significant cross-loadings (where an item loads on more than one factor):

- **P2Q30:** During your studies, have you felt free to express your ideas, opinions and beliefs?
- **P2Q14:** Are you able to contact teaching staff when you need to?
- **P2Q8:** Does your course contain the right balance of depth and breadth?

102. In each case, these load onto the first factor as well as one of the other factors. None of the variables loads onto more than two factors. The loading on the first factor is the weakest of the two loadings in every case.

Internal consistency of the underlying factors

103. Each factor shows a respectably high degree of internal consistency, indicating that the items included in each subscale are measuring aspects of student academic experience that are consistent with each other. An alpha of over 0.9 may indicate some redundancy within the first subscale, and we may want to review these questions to ensure that each question contributes a distinct aspect of the student experience. We should note that alpha tends to increase as more items are added to a subscale, possibly explaining some of the very high value for 'Factor 1'. This is illustrated in Table 9.

Table 9: Cronbach's alpha for each subscale identified in the principal components analysis for pilot two

Factor	Alpha
Factor 1 (Composite factor)	0.92
Factor 2 (Assessment and feedback)	0.79
Factor 3 (Learning opportunities)	0.74
Factor 4 (Intellectual challenge)	0.75

Conclusions on subscales

104. Broadly, the intended sections of the survey load onto factors indicating that they could be combined into subscales usefully, this is true for both the pilot one and pilot two questionnaires.

105. There are questions in both pilot questionnaires which measure very similar aspects of the student experience and as such correlate very highly, indicating that some of the questions may be redundant.

106. In both questionnaires most of the question items correlate with each other at least moderately – this suggests that although we can construct subscales from these items careful consideration should be given to the construction and use of such subscales.

Mode of response

107. In most of our analysis, we focus on the online respondents because of the greater size and variety of the population. In our analysis of phone responses, we concentrated on whether they different significantly from the online responses. This would be cause for concern: in a

mixed mode survey, we would like an individual's responses to be the same regardless of whether they respond online or by phone.

Pilot one

108. We treated a response of 'mostly agree' or 'definitely agree' as positive. In pilot one we found that for all but one of the questions, phone respondents tended to be more positive than online respondents. The one exception to this was P1Q10 ('I have had the right opportunities to work with other students as part of my course').

109. The difference between the percentage of phone and online respondents giving a positive response ranged from five to ten percentage points for most of the questions, with two questions having a difference of more than ten percentage points. We used a chi-squared test to help determine whether the association between mode of response and giving a positive response is statistically significant. For pilot one, 20 of the 23 questions were found to have a statistically significant association.

110. For comparison, we carried out the same tests using the same students' responses to the established survey. We found a very similar effect: that is, for almost all questions, students who answered the questions on the phone tend to respond more positively. In other words, this mode of response has not been introduced by pilot one, but is already a feature of the established survey.

111. There is a possibility that the students who respond by phone are different from the students who respond online, in a way that explains their positivity. For example, it might be the case that part-time students are more likely to respond by phone and, independently, more likely to be positive. To understand this, we created a model accounting for other characteristics that could explain how a student responds. This allowed us to obtain the predicted probability of a student responding positively to a question, assuming mode of response not to be a relevant factor. When looking at these predictions compared with the actual results, we can better determine how much of the effect is due to the mode of response.

112. Using these predicted responses slightly reduces the average difference between phone and online respondents. This suggests that other factors to some extent explain this difference. However, even using the predicted responses, it remained the case that phone respondents are more positive than online respondents, including for the question P1Q10 (which we had found to be an exception in the raw analysis). The findings were similar when looking at the main survey for the same group of students.

113. Table 10 shows the pilot one questions for the proportion of positive online respondents, and the proportion of positive phone respondents, differs by more than ten percentage points.

Table 10: Pilot one questions with a mode of response difference greater than ten percentage points

Question	Proportion of positive phone responses	Proportion of positive online responses
P1Q4: My course has challenged me to achieve my best work	88.5	78.5
P1Q12: Marking and assessment has been fair	85.1	72.5
P1Q14: Feedback has enabled me to improve my work	83.7	72.4
P1Q19: Overall, the quality of my course has been good	90.1	80.1

Pilot two

114. We counted the first two responses presented to students – for example ‘very good’ and ‘fairly good’ – as positive responses. We examined the percentage of students who gave a positive response to each individual question in pilot two, for both phone respondents and online respondents. For 24 questions, a higher proportion of phone respondents gave a positive answer. For 6 questions, a higher proportion of online respondents gave a positive answer.

115. Taking the questions where phone respondents had a higher percentage giving positive responses, most of the differences are in the range of three to six percentage points. We used a chi-squared test to determine whether these differences were significant. 25 of the questions saw a significant difference between online and phone responses. Of these questions, 23 had higher agreement rates among phone respondents and two had higher among online respondents.

116. As with pilot one, we created a model accounting for other characteristics that could affect how a student responds. The overall pattern is that, once other factors are accounted for, the difference between phone and online responses decreases slightly, though the effect is smaller than we saw in pilot one. Even taking into account student characteristics, it remains the case that, for some questions, online respondents were more positive than telephone respondents. We also noted that the pilot two population displayed a larger mode of response effect when answering the established survey: this suggests that the reduced mode of response effect is due to the new questionnaire, rather than the characteristics of the students.

117. Given the seemingly reduced mode of response effect for pilot two there are no questions with a difference of greater than 10. Therefore, Table 11 shows questions for which the proportion of positive responses differs, by mode of response, by more than five percentage points.

Table 11: Pilot two questions with a mode of response difference greater than five percentage points

Question	Proportion of positive phone responses	Proportion of positive online responses
P2Q10: Have assessments allowed you to demonstrate what you have learnt?	90.2	84.4
P2Q11: Has marking and assessment been fair?	91.2	84.1
P2Q12: Did you understand the marking criteria used to assess your work?	91.7	86.6
P2Q17: Are you able to get good advice about study choices?	83.7	78.1
P2Q18: Is the course well organised?	87.8	78.7
P2Q24: Do staff act on students' feedback?	83.3	77.5
P2Q26: Overall, how would you rate the quality of your course?	95.3	86.5
P2Q28: Are you aware of services at your university/college to support your mental wellbeing?	76.9	84.2
P2Q29: How easy is it to access your university or college's mental wellbeing services?	78.5	72.7

Conclusion

118. In both pilots, there are some differences between phone and online responses. These are not great enough to suggest that any questions are being misheard when asked by phone, or misread when asked online. For both pilots, as for the established questionnaire, phone respondents tend to be more positive, but there are some signs that this tendency is reduced in pilot two.

Comparison of direct questions and agreement scales

119. Much survey literature suggests that questionnaires using direct questions, with item specific responses, yield better quality data than questionnaires that ask respondents to agree or disagree with a statement. Direct questions are viewed as more likely to measure what we intend to measure and generate reproducible results.[5]

120. The perceived disadvantages of questions that use an agree/disagree format (many of which overlap) are as follows:

    a. Acquiescence bias – the tendency of respondents to agree with a statement regardless of their experience or opinion. This can be either as a shortcut to avoid

[5] See, for example, Saris, Willem, et al (2010) 'Comparing questions with agree/disagree response options to questions with item-specific response options', in Survey Research Methods May 2010, pp61-79.

cognitive burden, or due to a predisposition to respond positively when presented with a statement.

b. Higher cognitive burden in processing agree/disagree format questions, resulting in acquiescence bias as well as poor responses due to fatigue.

c. False correlations between unrelated items due to acquiescence bias and other errors in responses.

d. Mode of response effects, due to acquiescence bias being stronger when questions are asked by phone.

e. Overuse of the middle option or 'not applicable' due to difficulties in processing the questions.

121. We have examined whether, comparing our two pilot questionnaires, pilot two performs better than pilot one in relation to the issues in paragraph 119.

122. We began by looking at the percentage of respondents giving the same response to every question. These would include the extreme 'acquiescers': students who always give the same positive response. For pilot one, this figure was 7.6 per cent among online respondents, and for pilot two 3.8 per cent. As noted previously, however, we cannot discount the fact that the pilot surveys are carried out after the main survey and therefore pilot one's similarity to the current survey might lead to less engagement.

123. We also examined the longest chain of consecutive, identical answers given by respondents. We did this by looking at the fifth percentile, median and 95th percentile of longest blocks of consecutive answers. The two surveys barely differ when looking at the fifth percentile and median values both for online and phone respondents. When it comes to the 95th percentile, pilot two has longer chains of responses, although the differences are only seen in online responses. This is not unexpected as there are a greater number of questions.

124. To examine whether there is a difference in cognitive burden between the two surveys we considered whether students were more likely to respond with 'I cannot answer the question…' There does not appear to be any evidence of this. For both pilot one and pilot two, most questions had between 0.5 and 1.5 per cent of students responding this way. Both had a few questions with a larger proportion, in particular the question about depth and breadth. We can also look at the time taken to respond to the survey online to see if there is a difference. Overall, the average time taken to respond to each question is similar for both pilot surveys. Pilot two takes very slightly longer, but within half a second.

125. It is not easy deciding on what would constitute an overuse of the middle option or 'This does not apply to me.' But we did not find evidence that these options are overused in pilot one compared with pilot two. No middle option was provided in pilot two, making a comparison impossible. Respondents were more likely to respond 'This does not apply to me' in pilot two. However, this may be related to the removal of the middle option in this questionnaire, which some students may have previously used as a proxy.

126. It is difficult to answer with any certainty the question around false correlations. However, we can examine the correlations between questions that have not been put in the same scales. To feel confident that the groups picked out are correct we would expect there to be little correlation between questions not within these groups. Higher correlations between supposedly unrelated questions suggest that other questions could be more closely related and therefore we can be less sure that the groups picked out are correct. Overall, the

correlations between unrelated questions were higher in pilot one, suggesting there is more chance that something more is going on outside the questions we have grouped together.

127. Acquiescence bias tends to be stronger on the phone, and therefore a stronger mode of response effect could indicate a higher acquiescence bias in the survey. As previously noted, pilot one had a similar mode of response effect to the main survey, whereas pilot two looked to have a slightly reduced effect. This picture did not change once other factors were accounted for. This is perhaps the strongest indicator that any acquiescence bias could be reduced in pilot two. Independently, the reduced mode of response effect is a factor that counts in favour of the pilot two questionnaire.

Summative questions

128. This section examines the summative questions trialled in pilot one and pilot two. These questions are designed to prompt students to reflect on their overall feelings about their course, unlike the other questions, which ask about specific elements of their experience such as teaching or assessments. Summative questions are valued by some users of the data. They also serve as a way of checking the relevance of the other questions. For example, if a question correlates very weakly with a summative question about quality, it may suggest that the concept covered by the first question does not bear directly on students' views about quality.

129. The summative questions we trialled in the pilot are:

- **P1Q19:** Overall, the quality of my course has been good.
- **P2Q26:** Overall, how would you rate the quality of your course?
- **P2Q27:** On a scale of 0 – 10, how likely are you to recommend your course to a friend or a colleague?

130. We compared them with the summative question used in the established survey, which is:

- **EQ27:** Overall, I am satisfied with the quality of my course.

P1Q19: Overall, the quality of my course has been good

131. Students responded slightly more positively to this question than to the established question. The agreement rate (the proportion of respondents that selected the 'Definitely agree' or 'Mostly agree' response options) for P1Q19 is 80.1 per cent. This compares with an agreement rate of 77.9 per cent for EQ27 from the analytical subpopulation. The net difference in responses is 6.2 percentage points, which is slightly above the threshold we set for significant difference in paragraph 42. The increase in positivity is largely related to the 3.1 per cent of students who responded 'Neither agree nor disagree' to EQ27 and 'Mostly agree' to P1Q19, and the 8.6 per cent of students who responded 'Mostly agree' to EQ27 and 'Definitely agree' to P1Q19.

132. It would be helpful to explore further, through cognitive testing, whether these two questions prompt students to think about different concepts. It seems at least possible that a student could think that the quality of the course is good, and yet not be satisfied (since they expected something even better – excellence, say). If so, there is a further question about which of the two measures is better as an overall measure of the students' perception of their experience.

133. A good summative question should be correlated with other questions in the questionnaire. We tested this by examining the pairwise correlations between P1Q19 and the other questions in pilot one. The questions are shown in full in Annex A and summarised in Table 12, compared with similar statistics (using the same population) for the established summary question.

Table 12: Summary of pairwise correlations between summative questions and other questions (pilot one and established questionnaire)

Summative question	Mean correlation	Minimum correlation	Maximum correlation
EQ27	0.52	0.37	0.67
P1Q19	0.52	0.35	0.64

134. Like EQ27, P1Q19 is moderately or strongly correlated with almost all the other questions. Also like EQ27, it has the highest mean correlation of all the questions in the questionnaire. These are features you would expect of a good summary question. The weakest correlation is with the students' union question.

P2Q26: Overall, how would you rate the quality of your course?

135. If we construct a positivity measure for this question – the proportion of students who responded with 'Very good' or 'Fairly good' – we find that it is markedly more positive than the agreement rates for EQ27 or P1Q19. The positivity measure is 86.5 per cent, compared with 77.9 per cent for the established summary question. Similarly, the net increase in positivity is 14.7. This is mainly due to the removal of the middle option: 8.7 per cent of students answered 'Neither agree nor disagree' to question 27, and over three-quarters of these gave a positive response to P2Q26. In addition, 11.8 per cent of students responded 'Mostly agree' to EQ27, but gave the most positive option 'Very good' to P2Q26.

136. It is clear, then, that P2Q26 elicits more positive responses from students, but it is hard to disentangle what this should be attributed to. As well as the removal of the middle option, it may be the case that some students are confident that the course is very good, but less willing to say that they 'Definitely agree' that it is good, perhaps because of the certainty this implies. It may also be the case that introducing the concept of 'rating' changes the thoughts prompted by the question.

137. P2Q26 is moderately or strongly correlated with the other pilot two questions, as we would hope for a summary question. Mean, minimum and maximum values are shown in Table 13, alongside comparable values for the established survey.

Table 13: Summary of pairwise correlations between summative questions and other questions (pilot two and established questionnaire)

Summative question	Mean correlation	Minimum correlation	Maximum correlation
EQ27	0.52	0.37	0.67
P2Q26	0.49	0.30	0.64

138. The correlations are slightly weaker than for the established questionnaire or the pilot one summary question. However, pairwise correlations in pilot two tend to be lower for all questions, so this point does not straightforwardly imply that P2Q26 is less adequate as a summary question than P1Q19. The questions with the lowest correlations with P2Q26 are as follows:

- **P2Q28:** Are you aware of services at your university/college to support your mental wellbeing? (0.30)
- **P2Q13:** Were you given the marking criteria in advance? (0.30)
- **P2Q7:** When working with other students as part of your course, was this helpful for your learning? (0.31)
- **P2Q25:** Has the students' union (association or guild) had a positive impact on your experience? (0.32)
- **P2Q29:** How easy is it to access your university or college's mental wellbeing services? (0.37)
- **P2Q4:** Do you feel challenged by your course? (0.37).

139. It is noticeable that all these questions are either new, or substantial reworkings of existing questions. The lower correlations, as well as reflecting on P2Q26, should also prompt us to check whether these new questions are correctly formulated.

P2Q27: On a scale of 0 – 10 how likely are you to recommend your course to a friend or a colleague?

140. The second summative question trialled in pilot two is widely used as a test of customer loyalty. It was added to pilot two as a way of exploring alternative ways to understand the student experience.

141. This question is generally used to form a net promoter score (NPS). To calculate the NPS, respondents are grouped into 'promoters' (those who selected '9' or '10'), 'passives' (those that selected '7' or '8') and 'detractors' (those who selected response options from '0' to '6', inclusive). The score is produced by subtracting the percentage of detractors from the percentage of promoters. For example, if 40 per cent of respondents were 'promoters' and 15 per cent were 'detractors', the NPS would be 25.

142. For pilot two, the NPS is 16.1. Given that the possible range is from -100 to 100, this could be viewed as a positive but not exemplary score.

143. By examining how individual students responded first to the other summative questions, and then to P2Q27, we see that the NPS is measuring something very different from the overall satisfaction agreement rate. Many students who count as 'passives' or even 'detractors' in the NPS calculation gave a positive response to the overall satisfaction questions, as shown in Table 14.

Table 14: Agreement rates for other summative questions by response to P2Q27

P2Q27 responses	EQ27 agreement rate	P2Q26 positivity measure
Detractors (answered 0 to 6)	28.1 per cent	48.2 per cent
Passives (answered 7 to 8)	89.1 per cent	97.4 per cent
Promoters (answered 9 to 10)	98.9 per cent	99.4 per cent

144. We also found that there was much more variation in the NPS between student sub-groups, in comparison with the other summary measures. For example, students aged 25 and over had a NPS of 32.6 (16.5 higher than the NPS for the total analytical sample), whereas for the other summative questions this variation was minimal.

145. It may be the case that P2Q27 is allowing us to identify more granularity in the student experience. For example, it may be the case that students who 'Mostly agree' with the overall satisfaction statement range from those who are barely satisfied to those who like almost everything about their course. However, it might also be that students' responses to P2Q27 are influenced by factors unrelated to the quality of the course, such as the interests of their friends, or their willingness to provide recommendations to anyone. Further research and testing would be required to be confident that, in the context of higher education, this question is gathering useful information about the student experience.

Conclusion

146. We found that two of the alternative summative questions we trialled are working well and would serve as possible replacements for the established overall satisfaction question. We have less confidence in the question 'How likely are you to recommend your course to a friend or colleague?' It raises an intriguing possibility of understanding the student experience with more granularity. But it is not yet clear that categorising responders as promoters or detractors yields a useful measure of the student experience.

Conclusion

147. We found that both questionnaires performed reasonably well. We are confident that neither questionnaire contains any questions that are entirely incomprehensible or highly ambiguous. The length and format of the questionnaires seems to maintain reasonable levels of engagement.

148. Since pilot one was closely based on the established questionnaire, which has itself been refined over several cycles of testing and development, it is not surprising that it performs well. This finding is more impressive for pilot two, which explored a new format and therefore introduced more scope for error. On the other hand, pilot two did not display many of the benefits that are often associated with direct question surveys. This may be because the theoretical benefits of the format were counterbalanced by some immaturity in the questionnaire, or it may have been a restriction in the design of this pilot.

149. Neither questionnaire is ready for the field as it stands. The wording of some questions could be improved. Responses do not correlate entirely as expected, which should prompt us to consider again the underlying concepts behind the questionnaire. Questions remain about whether pilot two has the right number of response options, and whether these are worded correctly.

150. Given the theoretical benefits of the direct question format, a useful next stage would be to revisit the pilot two questionnaire, with a view to improving both the question wording and the response options. Further testing in the field would be required to establish whether this new questionnaire is indeed an improvement, and whether it displays more clearly the benefits we might expect from the format.

Annex A: Question wording for the established NSS and pilots

1. The established NSS uses a five-point agree/disagree response scale for each of the 27 questions, written as statements. The response options here are ‘Definitely agree’, ‘Mostly agree’, ‘Neither agree nor disagree’, ‘Mostly disagree’ and ‘Definitely disagree’. The question wordings for the established NSS are shown in Table A1.

Table A1: Question wording for the established NSS

Question number	Question text
EQ1	Staff are good at explaining things.
EQ2	Staff have made the subject interesting.
EQ3	The course is intellectually stimulating.
EQ4	My course has challenged me to achieve my best work.
EQ5	My course has provided me with opportunities to explore ideas or concepts in depth.
EQ6	My course has provided me with opportunities to bring information and ideas together from different topics.
EQ7	My course has provided me with opportunities to apply what I have learnt.
EQ8	The criteria used in marking have been clear in advance.
EQ9	Marking and assessment has been fair.
EQ10	Feedback on my work has been timely.
EQ11	I have received helpful comments on my work.
EQ12	I have been able to contact staff when I needed to.
EQ13	I have received sufficient advice and guidance in relation to my course.
EQ14	Good advice was available when I needed to make study choices on my course.
EQ15	The course is well organised and is running smoothly.
EQ16	The timetable works efficiently for me.
EQ17	Any changes in the course or teaching have been communicated effectively.
EQ18	The IT resources and facilities provided have supported my learning well.
EQ19	The library resources (e.g. books, online services and learning spaces) have supported my learning well.
EQ20	I have been able to access course-specific resources (e.g. equipment, facilities, software, collections) when I needed to.
EQ21	I feel part of a community of staff and students.
EQ22	I have had the right opportunities to work with other students as part of my course.
EQ23	I have had the right opportunities to provide feedback on my course.
EQ24	Staff value students’ views and opinions about the course.
EQ25	It is clear how students’ feedback on the course has been acted on.

Question number	Question text
EQ26	The students' union (association or guild) effectively represents students' academic interests.
EQ27	Overall, I am satisfied with the quality of the course.

2. Pilot one followed the structure of the established NSS, using the same five-point agree/disagree response scale. A number of questions were unchanged from the established survey, so to minimise repetition only 23 questions were piloted here. The wording for the questions in pilot one is shown in Table A2.

Table A2: Question wording for pilot one

Question number	Question text
P1Q1	Teaching staff are good at explaining the course content.
P1Q2	Staff have made the subject engaging.
P1Q3	My course is intellectually stimulating.
P1Q4	My course has challenged me to achieve my best work.
P1Q5	There is an appropriate balance of breadth and depth in the content of my course.
P1Q6	The balance of directed and independent study on my course supports my learning well.
P1Q7	My course has provided me with opportunities to explore ideas or concepts in depth.
P1Q8	My course has provided me with opportunities to bring information and ideas together from different topics.
P1Q9	My course has provided me with opportunities to apply what I have learnt.
P1Q10	I have had the right opportunities to work with other students as part of my course.
P1Q11	The criteria used in marking have been clear in advance.
P1Q12	Marking and assessment has been fair.
P1Q13	Feedback on my work has been timely.
P1Q14	Feedback has enabled me to improve my work.
P1Q15	Assessments have allowed me to demonstrate what I have learned on my course.
P1Q16	It has been easy to access learning resources (digital and physical) provided by my institution when I needed to.
P1Q17	Learning resources (digital and physical) provided by my institution have supported my learning well.
P1Q18	Overall, I am content with the students' union (association or guild) at my institution.
P1Q19	Overall, the quality of my course has been good.

Question number	Question text
P1Q20	My institution provides a free environment for the expression of ideas, opinions and beliefs.
P1Q21	My course has given me the knowledge and skills I think I will need for the future
P1Q22	My institution has made me aware of services to support my mental wellbeing.
P1Q23	My institution's services to support my mental wellbeing were available when I needed them. If you have not accessed support services, please select 'Not applicable'.

3. Pilot two used a different approach, asking direct questions and offering item-specific response options. These response options followed a general order in a four-point scale. These scales varied to link to the question so we have shown the type of response for reference. For example, if the response type was 'often' then the following options were given: 'Very often', 'Fairly often', 'Not very often' and 'Rarely or never'. The only exception to this was P2Q27, which asked respondents to give a response from 0 to 10. The question wordings and response types for pilot two are shown in Table A3.

Table A3: Question wording and response type for pilot two

Question number	Question text	Response type
P2Q1	Are teaching staff good at explaining course content?	Often
P2Q2	Do teaching staff make the subject engaging?	Often
P2Q3	Is the course intellectually stimulating?	Often
P2Q4	Do you feel challenged by your course?	Often
P2Q5	Have you had the chance to apply the theories and concepts that you have learnt?	Extent
P2Q6	Have you had the chance to bring together information and ideas from different topics?	Extent
P2Q7	When working with other students as part of your course, how helpful was this for your learning?	Helpful
P2Q8	Does your course contain the right balance of depth and breadth?	Extent
P2Q9	Does your course contain the right balance of directed and independent study?	Extent
P2Q10	Have assessments allowed you to demonstrate what you have learnt?	Often
P2Q11	Has marking and assessment been fair?	Often
P2Q12	Did you understand the marking criteria used to assess your work?	Extent
P2Q13	Were you given the marking criteria in advance?	Often
P2Q14	Has feedback helped you improve your work?	Extent
P2Q15	Are you able to contact teaching staff when you need to?	Extent
P2Q16	How well have teaching staff supported your learning?	Well

Question number	Question text	Response type
P2Q17	Are you able to get good advice about study choices?	Often
P2Q18	How well organised is the course?	Well organised
P2Q19	Have changes to the course been clearly communicated?	Often
P2Q20	Have you been able to access the learning resources (either digital or physical) that you need?	Often
P2Q21	How well have the physical and/or digital resources supported your learning?	Well
P2Q22	Do you get the right opportunities to give feedback on your course?	Extent
P2Q23	Do staff value students' opinions about the course?	Extent
P2Q24	Do staff act on students' feedback?	Extent
P2Q25	Has the students' union (association or guild) had a positive impact on your experience?	Extent
P2Q26	Overall, how would you rate the quality of your course?	Good
P2Q27	On a scale of 0 – 10 how likely are you to recommend your course to a friend or a colleague?	0–10
P2Q28	Are you aware of services at your university/college to support your mental wellbeing?	Extent
P2Q29	How easy it is to access your university or college's mental wellbeing services?	Easy
P2Q30	During your studies, have you felt free to express your ideas, opinions and beliefs?	Extent
P2Q31	Has your course given you the knowledge and skills you think you will need for your future?	Extent

Annex B: Derivation of analytical sample

1. The pilots were each designed to collect 10,000 online and 1,000 phone responses. Once these response targets had been achieved, we analysed the data for quality and applied some restrictions in order to reduce bias in the samples. We decided on these criteria by balancing the need for a certain number of responses and the aim to have a high-quality sample. Through this process we applied three exclusion criteria to the populations:
    - students who completed fewer than four pilot questions
    - students who gave the same response to every question in the pilot
    - students who gave a different response to matching pairs of questions in the established survey and pilot.
2. We identified that students responding to the pilot may be disengaged or lose interest during completion. If students completed very few questions in the pilot, we cannot draw much value from their response. We determined that an appropriate cut-off for this would be four questions as this is the size of the first sub-scale, on the quality of teaching. We have therefore removed any respondents who completed fewer than four pilot questions. By 'completed', we mean that they gave an answer other than the 'not applicable' or 'I do not understand the question' responses. This restriction removed 206 responses from pilot one and 87 from pilot two.
3. A further measure of engagement was to identify students who gave the same response to all questions in the pilot. This criterion aimed to remove students who acquiesced throughout the pilot and therefore showed a low-level of engagement with the content of the questions. By 'the same response', we mean that the student picked the same position on the four- or five-point scale for all questions. For this criterion we excluded P2Q27 from pilot two, as this used an 11-point scale. This restriction removed 877 responses from pilot one and 465 from pilot two.
4. The final criterion applied to the population looked at paired questions. Some of the questions in pilot one were unchanged from the established questionnaire, so for each participant we had two responses to the same question. This was not possible for pilot two as all questions were changed. Pilot one repeated eight questions from the established questionnaire but we focused on three of these, Q4/P1Q4, Q5/P1Q7 and Q6/P1Q8, as they appear early in the survey and are therefore subject to minimal impact from changes to ordering or introduction of new questions. The criterion applied was to remove students who gave a different response to two or three of these pairs of questions between the established and pilot surveys. This restriction removed 256 responses from pilot one and none from pilot two.
5. The three exclusion criteria restricted our pilot populations to two analytical samples. We assessed these samples to give ourselves confidence that we still had enough responses to draw conclusions, for the surveys as a whole and for sub-groups. Table B1 shows the final breakdown of our analytical sample by variables of interest to the analysis. It should be noted that we have not drawn conclusions on the phone responses of sub-groups as the number of responses is too small.

Table B1: Breakdown of analytical sample for pilot one and pilot two

Variable	Split	Pilot one		Pilot two	
		Online	Phone	Online	Phone
Country of provider	England	8,909	917	10,053	916
	Northern Ireland	146	5	183	3
	Scotland	785	48	865	68
	Wales	362	3	428	5
Sex	Female	7,209	536	8,091	512
	Male	2,971	437	3,414	478
Disability	Declared	2,452	236	2,597	234
	No known	7,750	737	8,932	758
Ethnicity	Asian	1,093	131	1,349	162
	Black	508	64	615	60
	Mixed	369	53	418	37
	Other	154	22	187	17
	Unknown	311	22	391	27
	White	7,767	681	8,569	689
Age group	Under 21	6,916	476	7,641	501
	21 to 24	997	125	1,177	134
	25 or older	2,289	372	2,711	357
Domicile	UK	8,902	840	9,962	848
	EU	680	37	812	41
	Rest of the world	619	96	754	103
Distance learner	No	9,456	580	10,674	637
	Yes	746	393	855	355
Mode	Full-time	9,228	556	10,430	623
	Part-time	974	417	1,099	369
Subject of study	Agriculture, food and related studies	142	3	115	6
	Architecture, building and planning	189	15	216	11
	Biological and sport sciences	505	36	569	44
	Business and management	1,133	110	1,301	135
	Combined and general studies	139	76	160	43
	Computing	495	94	599	79
	Design, and creative and performing arts	1,102	60	1,200	63
	Education and teaching	429	20	440	29
	Engineering and technology	470	69	541	72

Variable	Split	Pilot one		Pilot two	
		Online	Phone	Online	Phone
	Geography, earth and environmental studies	226	20	239	14
	Historical, philosophical and religious studies	407	35	434	48
	Language and area studies	415	50	465	47
	Law	384	47	459	43
	Mathematical sciences	213	23	237	28
	Media, journalism and communications	183	11	231	8
	Medicine and dentistry	180	16	156	22
	Physical sciences	321	27	384	36
	Psychology	671	86	763	88
	Social sciences	993	92	1,182	84
	Subjects allied to medicine	1,547	77	1,774	89
	Veterinary sciences	59	6	63	6

Annex C: Methodology for principal components analysis

1. Our approach to assessing the validity, reliability and internal consistency of the subscales within the NSS pilots draws on previous work carried out by Richardson in the initial development of the subscales used in the NSS.[6] It is also informed by Clarke and Watson and Thurstone for methods to assess the structure of the data.[7]
2. Using similar methods to Richardson, we have performed principal components analysis on the question items to identify whether there is an underlying factor structure that reflects the expected subscales. We have used the direct oblimin rotation strategy. This rotation allows the underlying factors identified to be correlated with each other. This is important because our inspection of the inter-item correlations shows us that the question items are correlated, and so we would expect the underlying factors to be correlated too. We excluded the summative questions in both pilots from the principal components analysis, because these questions are not designed to be part of any subscales.
3. The principal components analysis gives us 'factor loadings', which show the regression coefficients for each question on each factor – another way to say this is 'How well does the underlying factor predict the answers respondents gave to the question?' A high loading indicates that a question item is closely related to an underlying factor and a low loading shows that the question item is not closely related to the underlying factor. We first considered the unrotated factor loadings. In principal components analysis this demonstrates whether or not the items can all be said to load onto a single underlying factor. We then considered the rotated factor loadings.
4. We consider an item to load onto a factor when its rotated factor loading is above an absolute value of 0.3. We consider an item to have a 'zero' loading where its absolute rotated loading value is below 0.1.
5. In generating our factor model we have run the model iteratively. After each model was produced, we examined the resulting loadings and considered both the loadings and the inter-item correlations. Any problematic items were then removed from the model before the next solution is created.
6. We have excluded items from our analysis where there are fewer than three items loading onto a factor. This means that these items are not included in the factor solution and the number of factors extracted is reduced by one. Where a factor represents a single item (i.e there is only one question with a loading of over 0.3 on a given factor), we first reduced the number of factors by one while retaining the item in the model, to see whether the item can be forced to

[6] Richardson, JTE, 'National Student Survey: Interim assessment of the 2005 questionnaire', September 2005, available at https://webarchive.nationalarchives.gov.uk/20081203012349/http://www.hefce.ac.uk/pubs/rdreports/2005/rd20_05/.

[7] Clark, LA, and D Watson (1995), 'Constructing validity: Basic issues in objective scale development', Psychological Assessment Vol 7 No 3, pp309-319; Thurstone, LL (1947), 'Multiple factor analysis: A development and expansion of vectors of the mind', Chicago: University of Chicago.

load onto another factor. If the item still loads on a separate factor it is then removed from the factor model.

7. Following this approach, we have developed factor models with the pilot data that retain the most detail while removing any extraneous factors. These factors were then examined for internal consistency by calculating Cronbach's alpha for each subscale.

Annex D: Mode of response model

1. As discussed in paragraphs 107 to 118 of the report, we created a model to consider other variables that could have influenced how a student responded to the survey. We did this by creating logistic regression models. The characteristics we tested to see if they had an effect were age on entry, sex, domicile, ethnicity, mode of study, whether a student is distance learning, and whether a student has reported a disability. We modelled the likelihood that a student would give a positive response to a question. We put each of the potential factors into the model to attempt to find the best model. We did this for each of the different questions.
2. Once an optimised model was found, we used this to give us a probability that a positive response would be given by each student individually based on their characteristics. We then found the mean of these predicted probabilities, split into online and phone responses. Given that we did not include any potential mode effect in the model, the result of this step was to find the predicted agreement percentages, assuming that mode of response had no effect on how a student responded to a question.
3. The next step is to find the difference between phone respondents' and online respondents' predicted agreement rates. If there was no mode of response effect, we would expect this difference to be the same as the difference between the actual agreement rates. If this is not the case, we can be surer that any difference has been partially caused by the mode of response.
4. As an example of this, P1Q1 is 'I have had the right opportunities to work with other students as part of my course.' Among online students the percentage of students giving positive responses was 75 and among phone respondents 73.6. This would suggest that the mode of response effect was not large, and slightly in favour in online responses. We found that the optimised model for this question contained ethnicity, sex, distance learning and disability. Once these factors are considered, we would expect the percentage of online respondents giving positive responses to be 75.5, and for phone students the equivalent figure to be 68.2. This suggests that the mode of response is causing a difference in agreement rates of around 5.7 percentage points towards phone responses.